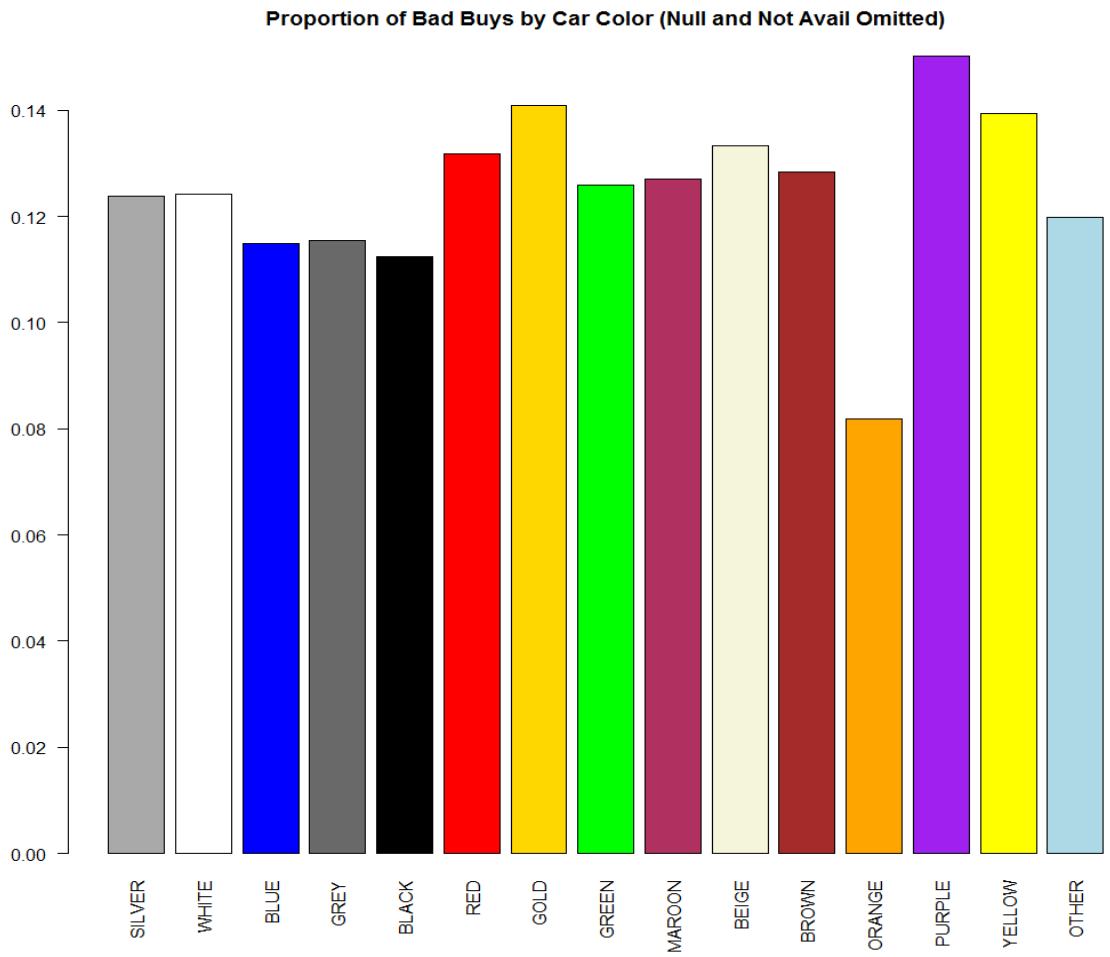
## Are Orange Cars Really not Lemons?

Ben Bullard & John Elder, April 2013
Elder Research, Inc., www.dataminginglab.com

A recent article in The Seattle Times, reported that "an orange used car is least likely to be a lemon." This discovery surfaced in a competition hosted by Kaggle to predict bad buys among used cars using a labeled dataset. Of the 72,983 used cars, 8,976 were bad buys (12.3%). Yet, of the 415 orange cars in the dataset, only 34 were bad (8.2%). The full breakdown of bad buy proportion by car color is shown in Table 1 and Figure 1 below, where the low proportion of bad buys among orange cars stands out prominently.

Table 1: Bad Buys by Color

| Row | Color | Count | Bad Buys | Percent |
|---|---|---|---|---|
| 1 | SILVER | 14875 | 1843 | 12.39% |
| 2 | WHITE | 12123 | 1506 | 12.42% |
| 3 | BLUE | 10347 | 1189 | 11.49% |
| 4 | GREY | 7887 | 911 | 11.55% |
| 5 | BLACK | 7627 | 858 | 11.25% |
| 6 | RED | 6257 | 825 | 13.19% |
| 7 | GOLD | 5231 | 737 | 14.09% |
| 8 | GREEN | 3194 | 402 | 12.59% |
| 9 | MAROON | 2046 | 260 | 12.71% |
| 10 | BEIGE | 1584 | 211 | 13.32% |
| 11 | BROWN | 436 | 56 | 12.84% |
| 12 | ORANGE | 415 | 34 | 8.19% |
| 13 | PURPLE | 373 | 56 | 15.01% |
| 14 | YELLOW | 244 | 34 | 13.93% |
| 15 | OTHER | 242 | 29 | 11.98% |
| 16 | NOT AVAIL | 94 | 24 | 25.53% |
| 17 | NULL | 8 | 1 | 12.50% |
| 18 | TOTAL | 72983 | 8976 | 12.30% |




But how unusual is this low proportion? That is, assuming the true proportion is really equal, what is the likelihood that it could have occurred by chance for a random partition of that size? Such a calculation takes into account the numbers of cars making up both proportions (good and bad Orange vs. good and bad non-Orange[1].) When we apply a 1-sided statistical hypothesis test for equality of proportions between two samples it yields a *p*-value of 0.00675 (see Equation 1). In other words, the hypothesis test reveals that if the underlying reality is that the proportion of bad buys among orange cars is really equal to the proportion of bad buys among all non-orange cars, then the probability that one would observe a sample proportion for orange cars that is so much lower than the sample proportion for non-orange cars (given sample sizes of 415 and 72,466, respectively) is only 0.675%.

```
> prop.test(c(34,8917), c(415,72466), alternative="less")$p.value          Eqn. 1 (in R code)

[1] 0.006754577
```

[1] Note that NULL and NOT AVAIL were removed from the analysis altogether throughout this paper. This was due to small sample size as well as the lack of explanation as to why the color of the car was not reported.

Given such a low *p*-value, it seems likely that orange cars really are better buys. Put another way, since the default or "null" hypothesis (that the proportions are actually equal) is less than 1% likely, there's more than a 99% chance that the alternative hypothesis (that orange cars are really good buys) is true.

**Interpretation**

But why orange? The Seattle Times reports "As for why orange used cars are most likely to be in good shape, the numbers did not hold the answer. One notion was that flashy colors may only attract car fanatics who would be more likely to take care of their vehicles. That didn't pan out, however, since the least well-kept cars turned out to be purple." Brainstorming other explanations, we wondered if orange cars tend to be made by only a few manufacturers, or only represent a few makes or models, or even years of production; i.e., that orange is confounded (mixed up with) another variable actually related to reliability. It's likely not a cause, but a "tag-along" effect. A colleague suggested that orange may be more visible to other drivers and thus those cares are involved in fewer collisions. The opportunity for speculation is endless! Two comments here:

1. To really examine these questions would involve building a data mining model from the full set of Kaggle contest data, which included many other variables.
2. We have learned to not trust the interpretability of a model. That is, that explanations of why some finding might be true are readily invented *after* a finding is made!

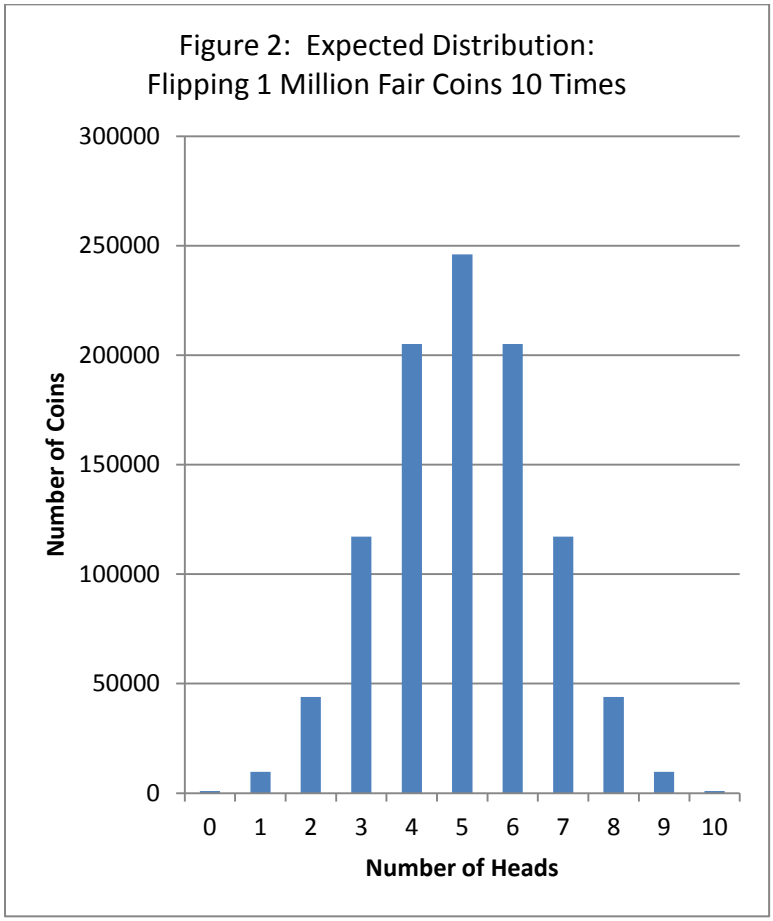
We decided to pause at this point of knowing just the information related to one variable (color) and explore the narrow, but important question of what the true findings should be. What colors are most interesting related to reliability and how confident in those results are we? This paper establishes a framework for approaching problems of this kind and shows how the immediate finding might not be the most interesting, and how the likelihood of finding something that appears interesting only by chance is much greater than traditional statistical tests reveal.

**Where Did Our Hypothesis Come From?**

As we consider the strength of our conclusion about orange cars, the first thing we should note is that that hypothesis was only developed *after* seeing the data. No one surmised it and then went out and collected data to test that idea. Rather, data was collected, and the graph revealed that orange is an outlier, then we applied a hypothesis test to its numbers. The importance of this distinction is probably not obvious, but can be made more clear by a simple illustration. Imagine that I report to you the results of a test I ran to determine if a coin is fair. I flipped the coin 10 times, and discovered that it landed heads every single time. Applying a hypothesis test I find that I can reject the null hypothesis (that the coin is fair) with a p-value of 0.00195[2]. In other words, there is only a 0.195% chance that a fair coin would land heads on all 10 flips, so we could conclude (at a 99.8% confidence level) that this coin must not be fair. It is still possible for the coin to actually be fair, but the hypothesis test tells us how

[2] Probability of 10 consecutive heads or tails with a fair coin is 0.5^10 or 1/1024. Therefore the p-value for a 2-sided test would be 2/1024 or 0.00195.

unlikely that is. Then, if I reported to you that I tested another 999 coins, and found that all 1000 of them landed heads every single time, then you would likely be convinced beyond all doubt that there is something fishy about the whole lot of them. 10,000 heads and 0 tails! All of the coins must surely be biased. However, if I then mentioned in passing that I had also happened to test 999,000 other coins, which resulted in a variety of other proportions of heads and tails, would that change things? These are all separate coins, and not repeated trials of any of the earlier coins I told you about. Every test is independent, so they don't affect each other…. so how could they matter? Intuitively though, you know they could! You would want to know right away whether I identified ahead of time the 1,000 coins which always landed heads or if I first tested all 1,000,000 together and only picked out the 1,000 afterward. You know that if I were to test 1,000,000 coins, I would *expect* some of them to land heads every single time *even if* every single one was fair! In fact, if you do the math[3], you will find that one would expect 977 out of 1,000,000 coins to land heads all ten times on average. Therefore, finding 1,000 coins which landed all heads is not surprising at all, and should not be used as evidence to suggest that those 1,000 coins are biased[4]. The expected distribution of each outcome (i.e., 9 of 10) is shown in Figure 2 and Table 2.




[3] 1,000,000 * 0.5^10 = 976.56

[4] One could, however, *hypothesize* based on this finding that these 1000 coins are biased and run a *subsequent test* on these particular coins to gain evidence either supporting or refuting this hypothesis.

Table 2: Expected Results Flipping 1M Fair Coins 10 Times Each

| Heads | Percent | Count |
|---|---|---|
| 0 | 0.1% | 977 |
| 1 | 1.0% | 9,766 |
| 2 | 4.4% | 43,945 |
| 3 | 11.7% | 117,188 |
| 4 | 20.5% | 205,078 |
| 5 | 24.6% | 246,094 |
| 6 | 20.5% | 205,078 |
| 7 | 11.7% | 117,188 |
| 8 | 4.4% | 43,945 |
| 9 | 1.0% | 9,766 |
| 10 | 0.1% | 977 |

This simple example shows that the significance of the finding that 1,000 coins landed all heads rests entirely on the question of whether or not I had hypothesized *ahead of time* that these particular coins were biased, and the other 999,000 not, or, if I simply tested all 1,000,000 coins indiscriminately and picked out the 1,000 based on the results.

In the same way, for our present investigation involving orange cars, the question of *when* we arrived at the hypothesis that orange cars are good buys is important. Since the hypothesis was arrived at only *after* viewing the data, it follows that had the data been different, the hypothesis itself may have changed. Obviously, if green cars happened to have had a strikingly low proportion, we would have tested the hypothesis that green cars are good buys. Or, if red cars happened to have had a strikingly high proportion, we'd have hypothesized that red cars are bad buys. If we assume that the true proportion of bad buy's among all cars for all colors is actually identical, the probability that one would find a statistically significant difference between red cars and non-red cars is low, and the probability that one would find a statistically significant difference between green cars and non-green cars is low. But the probability that one would find a statistically significant difference between *some* color car and all other colored cars might not be that low! In fact, if the number of colors was great enough, the prospect of finding a statistically significant difference would be almost certain (just as in the case of finding a thousand coins that land heads 10 times in a row if we flip 1 million of them).

What we see is that statistical hypothesis tests only work when the hypothesis comes first, and the analysis second. One cannot use the data to inform the hypothesis and then test that hypothesis on the same data. That leads to overfit and over-confidence in your results, which leads to the model underperforming (or failing entirely) on new data, where it is most needed.

**The Danger of Vast Search**

And yet, how do we know what to hypothesize? Isn't the great strength of data mining that the computer can try out all sorts of things are report back which one might work? Yes, we can and often should use data to drive and develop our hypotheses, but we must then test those hypotheses on *unseen* data. And to get an idea of the significance of a finding without such unseen data we have to ask a broader question than how likely is this *exact* finding to have occurred by chance. We have to ask: "How likely is it that *any* finding that *this interesting* could occur by chance?"

Data Mining has the power and peril of what we call the "vast search effect": If you search hard enough over enough variables, we are sure to find *something* "interesting", whether that finding is real or the effect of random chance. Hypothesis tests are supposed to tell us how likely it is that our finding could have happened by chance, but they fail to do so accurately when the hypothesis itself is contingent on the very same data against which it is tested.

Does this mean that orange cars aren't really good buys after all? No, they still could be. But what this means is that the *p*-value on which that conclusion is based is misleading. We must either take into account the fact that we both developed and tested our hypothesis using the same data, or find new data on which to test our hypothesis in order to calculate a more accurate probability. Three possible approaches to doing this are described below.

**Solution 1: Partitioning**

As mentioned previously, there is nothing wrong with using data to develop hypotheses. A glance at Figure 1 reveals that the proportion of bad buys among orange cars is lower than that of other colors; and unexpected, data-driven hypotheses like this often lead to novel and beneficial discoveries. But, to really *test* this hypothesis, we should now go out and collect data on a whole new group of used cars and see how well it holds up. But that is easier said than done! It takes a lot of work to survey thousands of used car buyers to see if the car panned out or not (and over what time frame, etc.). For this reason, data miners partition the data to mimic repeated experiments. The idea is very simple. After receiving the dataset (in this case 72,983 records) and *before* analyzing it, split it into a training partition and a testing (or evaluation) partition. If the cases are not independent, this split must be done carefully, according to time say, but here may be done randomly. The goal is to make the testing data *simulate future data*. Often one uses say 70% of the data for training, and the remaining 30% for testing. The analyst uses the training dataset to build a model or explore the data to come to some hypothesis, and then employs the testing dataset to ~~to~~ test that model or hypothesis on unseen data. The test step simulates the use of the model in the real world. (Note that if this test is done too often; that is, if there are many iterations of training, then it can become partly "known" to the user/model and lose some of its power to simulate future reality. Also, the training data might – at random or not – be too different from the training data, and thus we recommend cross-validation or bootstrapping to do an even better job of testing.)

Unfortunately, for our problem, reserving data is not something we can go back and do now that we have failed to do it in the first place! We will show a simple example of how one might have applied this approach to the problem at hand, but want to underscore that it is too late to actually use it.

Imagine that we have just received the 72,983 cases of used cars labeled with color and quality. Ideally, we might want to partition according to time, but because we do not have that information, we will randomly partition the dataset into 60% training and 40% testing (Table 3). Having done so, our proportion chart would look like Figure 3:

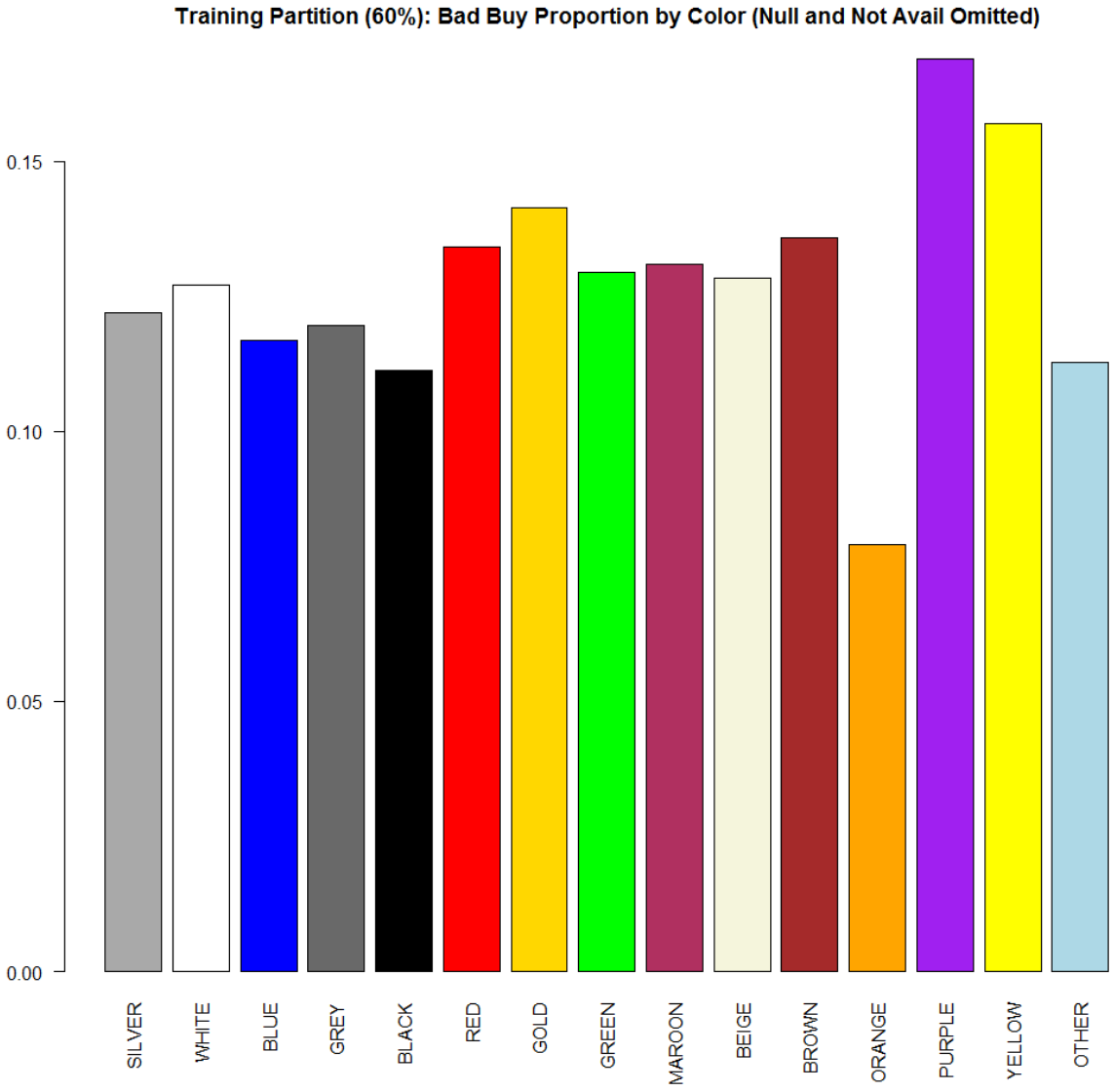



Table 3: A 60% Random Sample

| Row | Color | Count | Bad Buys | Percent |
|---|---|---|---|---|
| 1 | SILVER | 8,858 | 1,080 | 12.19% |

| | | | | |
|---|---|---|---|---|
| 2 | WHITE | 7,233 | 919 | 12.71% |
| 3 | BLUE | 6,210 | 726 | 11.69% |
| 4 | GREY | 4,709 | 563 | 11.96% |
| 5 | BLACK | 4,593 | 511 | 11.13% |
| 6 | RED | 3,749 | 503 | 13.42% |
| 7 | GOLD | 3,136 | 443 | 14.13% |
| 8 | GREEN | 1,893 | 245 | 12.94% |
| 9 | MAROON | 1,223 | 160 | 13.08% |
| 10 | BEIGE | 935 | 120 | 12.83% |
| 11 | BROWN | 280 | 38 | 13.57% |
| 12 | ORANGE | 253 | 20 | 7.91% |
| 13 | PURPLE | 231 | 39 | 16.88% |
| 14 | YELLOW | 153 | 24 | 15.69% |
| 15 | OTHER | 142 | 16 | 11.27% |
| 18 | TOTAL | 43,598 | 5,407 | 12.40% |

Now based on this training dataset (and pretending we have never seen the full dataset!), we would still hypothesize that the orange car proportion is interestingly low. It is conceivable that we could have chosen a partition that would not have led to this hypothesis (or in which it would be hard to tell if we would have come to such a hypothesis), but in this case orange remains an obvious outlier. Now, let's assume that we wish to test the hypothesis that orange cars have a lower proportion of bad buys than non-orange cars. Our testing dataset (Table 4) looks like Figure 4:

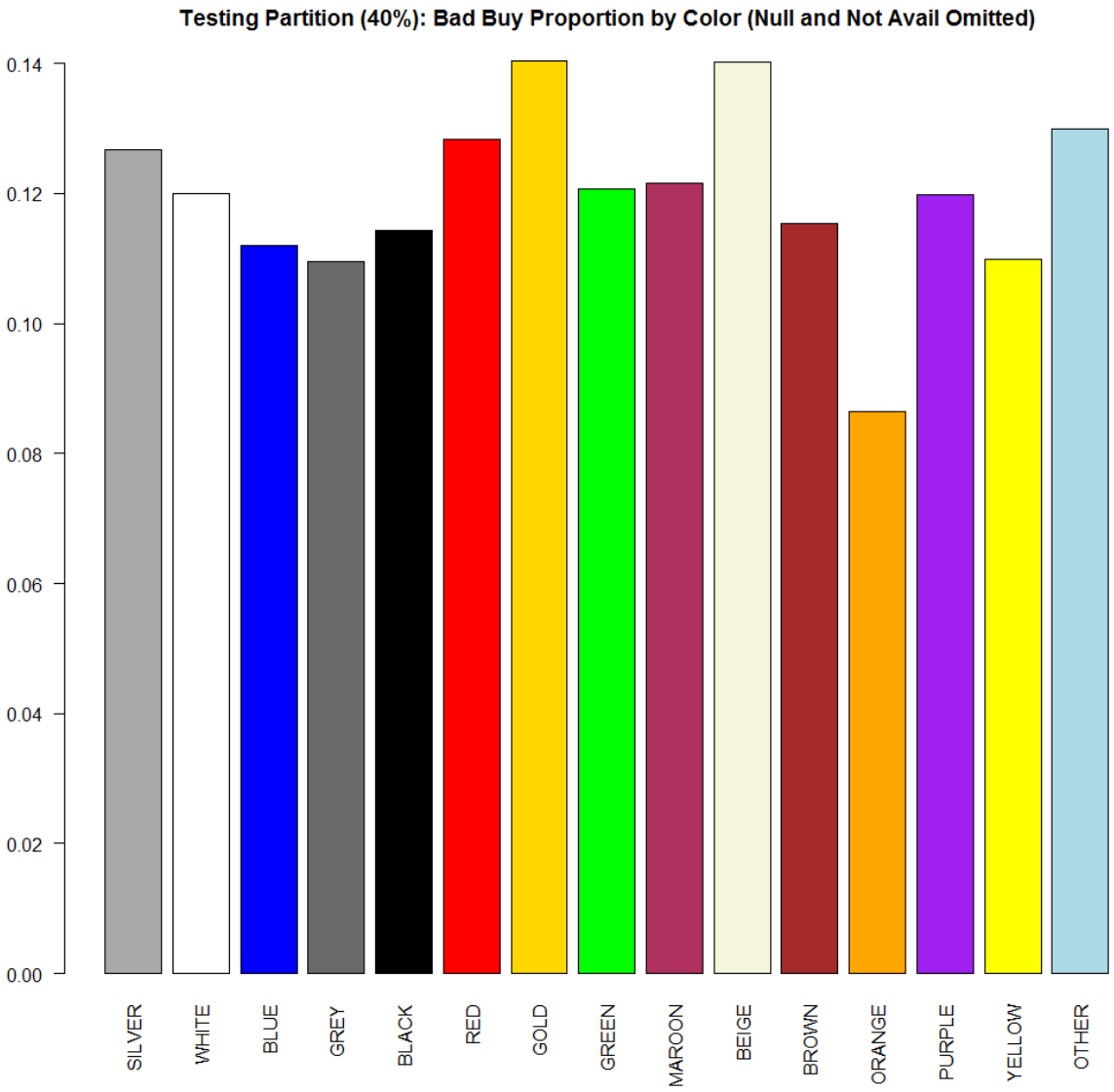



Table 4: Test Dataset (Remaining 40% sample)

| Row | Color | Count | Bad Buys | Percent |
|---|---|---|---|---|
| 1 | SILVER | 6,017 | 763 | 12.68% |
| 2 | WHITE | 4,890 | 587 | 12.00% |
| 3 | BLUE | 4,137 | 463 | 11.19% |
| 4 | GREY | 3,178 | 348 | 10.95% |
| 5 | BLACK | 3,034 | 347 | 11.44% |
| 6 | RED | 2,508 | 322 | 12.84% |
| 7 | GOLD | 2,095 | 294 | 14.03% |
| 8 | GREEN | 1,301 | 157 | 12.07% |
| 9 | MAROON | 823 | 100 | 12.15% |

| | | | | |
|---|---|---|---|---|
| 10 | BEIGE | 649 | 91 | 14.02% |
| 11 | BROWN | 156 | 18 | 11.54% |
| 12 | ORANGE | 162 | 14 | 8.64% |
| 13 | PURPLE | 142 | 17 | 11.97% |
| 14 | YELLOW | 91 | 10 | 10.99% |
| 15 | OTHER | 100 | 13 | 13.00% |
| 18 | TOTAL | 29,283 | 3,544 | 12.10% |

Applying a 1-sided hypothesis test for equality of proportions between the sample of orange cars and non-orange cars in the testing partition yields a *p*-value of 0.109 (Equation 2).

```
> prop.test(c(14, 3530), c(162, 29121), alternative="less")$p.value          (Eqn 2 in R code)

[1] 0.1087065
```

This *p*-value indicates that the proportion of bad buys among orange cars is low, but not low enough to be conclusive at the typical levels of significance. (In medical journals for example, a significance of below 5% is required to publish.) In other words, we would hesitate to conclude that orange cars have a true proportion of bad buys lower than non-orange cars.

However, remember that this outlines a process that *could* have been applied to this problem, and not a solution to the actual question. The *p*-value for the test dataset is much higher than we previously saw primarily because the dataset is smaller, so if we put a difference percentage of cars in the testing sample we'd come up with a different p-value. Partitioning tends to reduce significance, since it's harder for a random finding to show up on both data sets, so it makes a step in the right direction of reducing the vast search effect. Yet even it may not protect us if we are not careful. We often find ourselves making a hypothesis using the training partition, evaluating that hypothesis on the test partition, and then returning to training to revise that hypothesis or make new ones. By alternating between training and testing, we've created an information "leak" from the future (testing) to the present (training), and are increasing the chances of fooling ourselves. For this reason data miners often split the data into 3 groups (training, validation, and testing), to allow themselves to employ some back-and-forth between the first two, but saving the test dataset for a single, final evaluation.

**Solution 2: Mathematical Inference**

We've established that the best way to determine if orange cars are really better buys is to gather brand new data and test our hypothesis on that data. However, when that is impractical, we noted that we could have used partitioning to both develop and test our hypothesis using the existing dataset. But since it is too late for that, is there anything else we can do? Is there some way to *account* for the fact that we both developed and tested our hypothesis on the same dataset? The following two approaches are attempts to do that. Each has~~ve~~ limitations, but both are useful.

The key to both of these solutions is in redefining our question. Previously, we ran a hypothesis test which answered, "How likely is it that the proportion of bad buys among orange cars would be so low by chance alone?" But, this leads to a misleading result because orange was self-selected based on its own "interestingness". A better question would be, "How likely is it that the proportion of bad buys among *some*-colored cars would be so unusual by chance alone?" This question is better because it compares the most interesting observed result (orange), not with what we would expect at random from orange, but with what we would expect at random from the *most interesting color* (whatever it may be). In this way, it assumes that we could have selected any one of the 15 different colors[5], and accounts (at least in part) for how our hypothesis itself could have been different if the data were different.

Now you may have noticed that we used the term "most interesting", rather than "lowest proportion". The reason for this is two-fold. First, we should recognize that a color-group having an especially *high* proportion of bad buys might also be interesting. We might want to know what color car to buy as well as what color to avoid, so we will consider both possibilities. Second, we use the term "interesting" because a low proportion, as shown in the Figures, does not take into account sample size. If, for example, there were 3 neon cars in the dataset, and none were deemed bad buys, then its sample proportion would be 0%! But obviously that would not convince us that neon cars are better buys than other colors, as our intuition would tell us we don't have enough data. For this reason, a better measure of "most interesting" is lowest *p*-value. *P*-values take sample size into account, providing a measure of how *unusually* high or low a proportion is. Therefore, what we want to determine, is the probability that the lowest *p*-value, for any color, is at least as low as the observed *p*-value for orange, under the null hypothesis that it is truly all random.

Our first cut at this is through algebra. Orange's *p*-value of 0.00675, means that 0.675% of the time, a group the size of orange would have a proportion that low, by chance alone[6]. Additionally, it implies that 1.35% of the time (twice that often), orange would have a proportion that is that *extreme* (low or high), by chance alone. Furthermore, because sample size is accounted for, it implies that 0.675% of the time, red would have a proportion that is that *unusually* low; and that 1.35% of the time, red would have a proportion that is that unusually extreme, by chance alone. Now given that this holds for every color, and that there are 15 colors total, we can estimate the probability that *no* color would have a result as interesting as orange did by calculating the probability of "not orange" *and* "not blue" *and* "not gold" *and* "not green" and so on. In probability, "and" means multiply, and "not orange" means "1 – orange". Therefore, we can calculate the probability of "not any color" as follows:

Number of colors = 15

2-sided (as extreme as): P = (1 – 0.0135)^15 = 0.816

1-sided (as low as): P = (1 – 0.00675)^15 = 0.903

[5]

[6] Assuming that the true underlying proportion of bad buys among orange and non-orange cars is equal

This means that 81.6% of the time, no color would have a result as extreme as the result we actually observed in orange, or conversely, that 18.4% of the time, some color would. Now this suggests that our result for orange is somewhat unusual, but not that unusual, and certainly much less unusual than our original *p*-value suggested. Additionally, even if we apply a 1-sided test, and only consider unusually low (not high) proportions, our calculations suggest that we should still expect to find a proportion at least as unusually low as that of orange 9.7% of the time. Therefore, a *p*-value of 0.097 or 0.184 would be a much better indicator of true significance than 0.00675.

**Solution 3: Simulation**

Another way to answer the question "How likely is it that the proportion of bad buys among *some*-colored cars would be that unusual by chance alone?" is to apply a technique called "target shuffling" (invented or rediscovered, most likely, by one of us). This technique is a form of simulation in which we essentially repeat our experiment many times to simulate the results one might expect at random. We call it target shuffling because the technique involves randomly "shuffling[7]" the target (dependent) variable, while leaving the rest of the dataset in place. This is illustrated for a small sample of data in Figure 5.

Figure 5: Example of Target Shuffling

| Input | Target |
|---|---|
| **Color** | **Bad Buy** |
| BLACK | TRUE |
| BEIGE | FALSE |
| BLACK | FALSE |
| MAROON | FALSE |
| GREY | TRUE |
| GREEN | FALSE |
| YELLOW | FALSE |
| BLUE | FALSE |
| BLACK | FALSE |
| GREEN | FALSE |
| GREEN | FALSE |
| PURPLE | FALSE |
| BEIGE | FALSE |
| GOLD | FALSE |
| RED | FALSE |
| YELLOW | FALSE |
| RED | FALSE |
| BROWN | TRUE |

Shuffle

| Input | Target |
|---|---|
| **Color** | **Bad Buy** |
| BLACK | FALSE |
| BEIGE | FALSE |
| BLACK | FALSE |
| MAROON | FALSE |
| GREY | FALSE |
| GREEN | TRUE |
| YELLOW | FALSE |
| BLUE | FALSE |
| BLACK | FALSE |
| GREEN | FALSE |
| GREEN | FALSE |
| PURPLE | TRUE |
| BEIGE | TRUE |
| GOLD | FALSE |
| RED | FALSE |
| YELLOW | FALSE |
| RED | FALSE |
| BROWN | FALSE |

Target shuffling creates a dataset in which we *know* that no real relationship exists between the target variable and any input variable. That is, the null hypothesis holds. It is to this shuffled dataset that we apply our new hypothesis, model or modeling process[8], and then measure the new significance of our

[7] Shuffling denotes random reordering or sampling without replacement

[8] Applying a modeling process would entail re-training the model on the shuffled data and is an effective method of testing for overfit

hypothesis or performance of the model. By repeating this process many times, we are able to create a distribution of "performances" which we *know* to be attributable to random chance alone. Therefore, we are able to compare our results on real data to our "possible" results on random data to get a better sense of just how significant our original results really are.

This technique has great value for at least three reasons. First, with today's computing power it is often much faster and easier to simulate results than to go through the sometimes painstaking effort of accurately calculating them! Second, this technique is typically much more intuitive to a non-statistician, and thereby often leads to results which are viewed as more credible by those who do not understand the underlying statistics. Third, this technique is a great way to confirm and double check a result which has been calculated statistically, since everyone makes mistakes!

In order to apply target shuffling to our problem, we followed this process:

1. Shuffled our vector of bad buys (containing 8,951 Trues and 63,930 Falses)
2. Aggregated to get bad buy count by color
3. Ran an equality of proportions hypothesis test for each color vs. all other colors
4. Determined the minimum *p*-value across all colors
5. Repeated this process 10,000 times

When running a two-sided hypothesis test, and thereby testing for extreme proportions (whether high or low), we found the distribution of Figure 6, in which 1,635 out of the 10,000 trials (16.4%) yielded a minimum *p*-value of less than or equal to our threshold value of 0.0135.

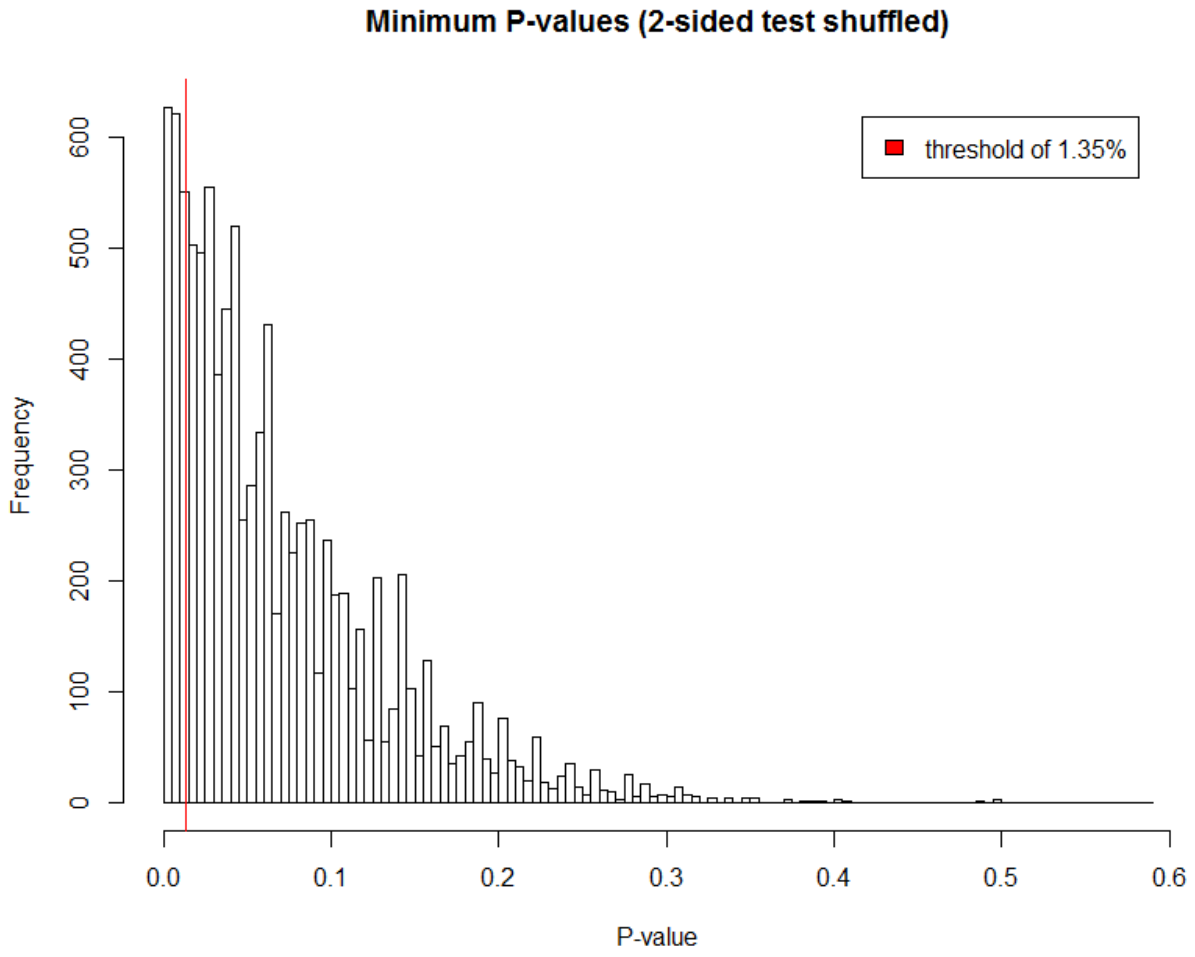



Figure 6

This would indicate that we could expect *some*-colored car to achieve a result as extreme as orange's roughly 16.4% of the time if no difference in proportion between car colors truly exists. As before, we also ran a 1-sided hypothesis test to test for low proportions only, and found that 715 out of 10,000 trials (7.2%) yielded a minimum p-value of less than or equal to our threshold value of 0.00675 (Figure 7).

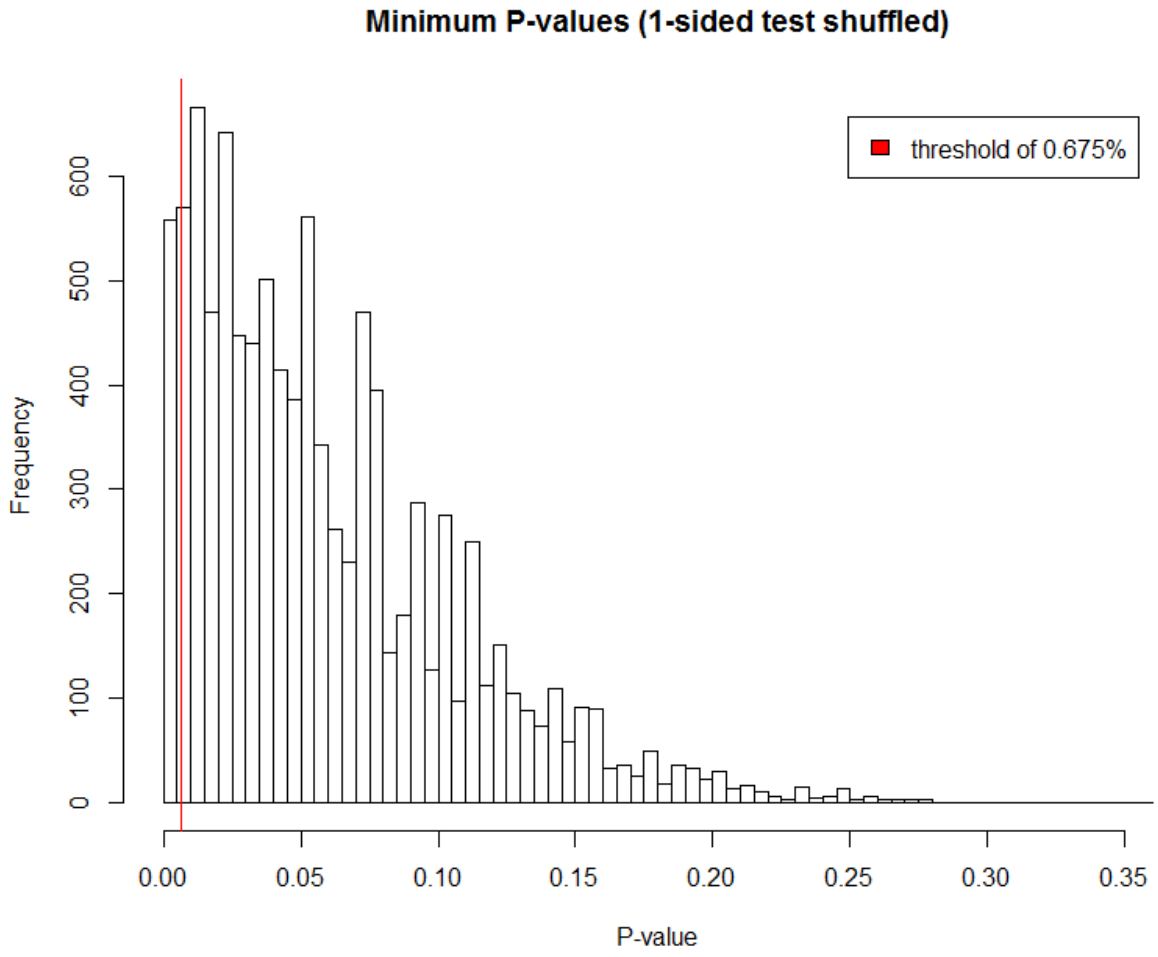



Figure 7

This would indicate that we could expect some-colored car to achieve a result as unusually low as orange's roughly 7.2% of the time if no difference in proportion between car colors exists.

Now, for the more curious reader, there are a few subtleties to observe. First, we have used "shuffling" rather than "re-sampling". Or, in other words, we have used sampling *without* replacement rather than sampling *with* replacement to construct each dummy target vector. Shuffling has the advantage of limiting the sources of variation by keeping the overall bad buy proportion constant. However, it also introduces a slight dependence between the proportion of bad buys for any given color and the proportion of bad buys among all other colors[9]. Therefore, we tried repeating our experiment using sampling with replacement and achieved simulated p-values of 0.1624 and 0.0721, respectively, which nearly match our previous values of 0.1635 and 0.0715. So the type of sampling used is a minor factor.

[9] This is because if the total number of bad buys is t and the number of bad buys for a given color is n, then the number of bad buys for all other colors must be t - n

The second subtlety is that our simulated *p*-values of 0.1635 and 0.0715 are more different from our algebraic p-values of 0.184 and 0.097. A major reason for this becomes clear when we look at the distribution of all *p*-values for all colors over the 10,000 iterations with replacement[10] as in Figure 8:

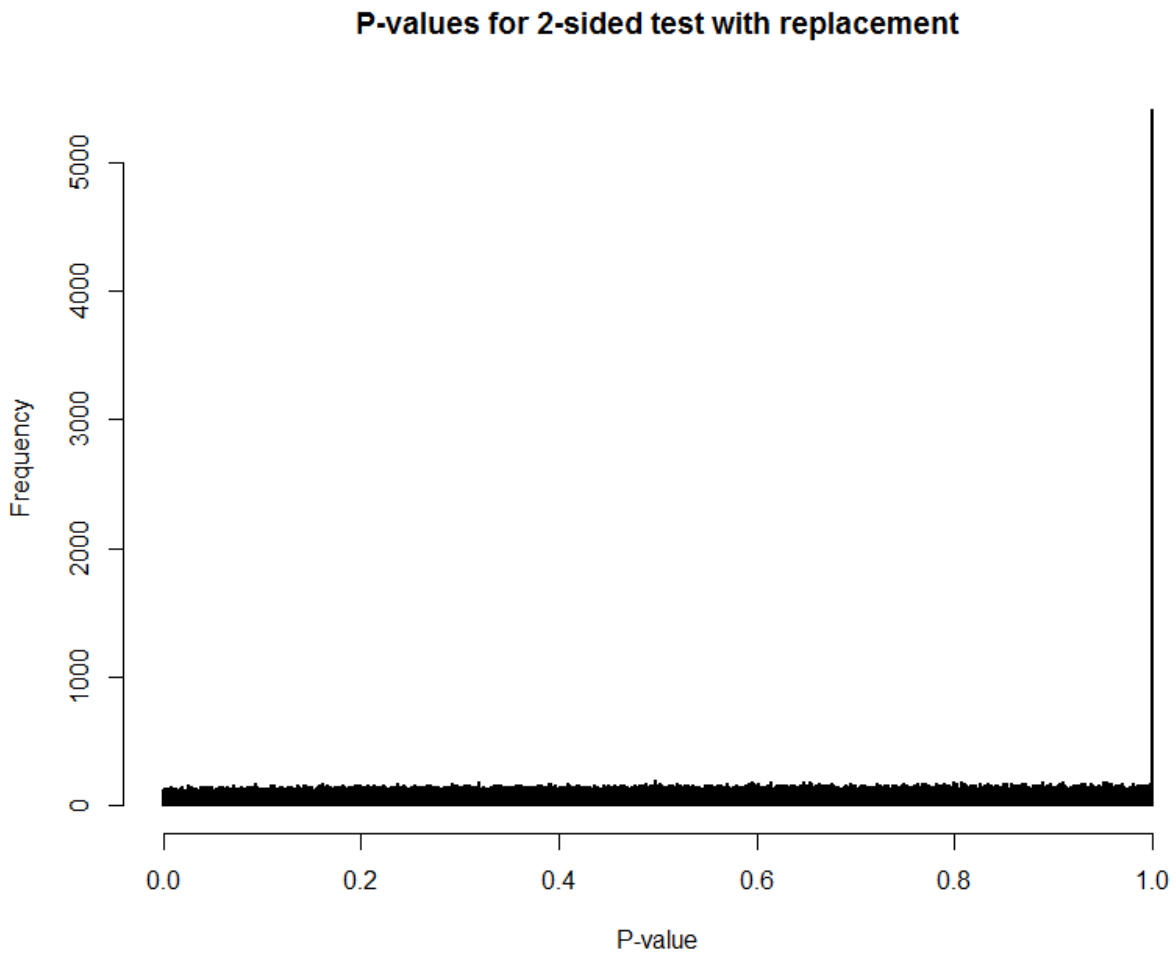



Figure 8

Note that there is a major spike at $p = 1$. The reason for this boils down to the fact that the distribution of sample proportions is non-continuous because we are working with whole numbers. Analogously, if we ran a binomial hypothesis test to test whether a coin is fair with 5 trials, the only possible resulting *p*-values would be 0.0625, 0.375, and 1, as shown in Table 5

[10] The distribution for trials with (rather than without) replacement is shown because the distribution is cleaner and illustrates the point more clearly

**Table 5: P-Values for Coin with 5 Flips**

| Heads | Percent | *P*-value |
|---|---|---|
| 0 | 3.1% | 0.0625 |
| 1 | 15.6% | 0.375 |
| 2 | 31.3% | 1 |
| 3 | 31.3% | 1 |
| 4 | 15.6% | 0.375 |
| 5 | 3.1% | 0.0625 |

Let's consider the *p*-values in the coin experiment of Table 5, 6.25% of the time the outcome is either all or no heads, and the *p*-value for *each* of those cases is 0.0625. Thus, 6.25% of the time the p-value is <= 0.0625. Likewise, 37.5% of the time the *p*-value is less than or equal to 0.375. Yet, it is not true that 50% of the time the *p*-value is less than or equal to 0.5. Rather, that is the case only 37.5% of the time! P-values of 1 are very common. Combining many such distributions together (as happens with multiple bins such as color) produces distributions like Figure 8, in which there is a piling up of values at 1, and in which the proportion of *p*-values less than or equal to a given value are actually not equal to the value itself, as one might theoretically expect. For example, the true proportion of *p*-values that were less than or equal to 0.5 was only 47.8%, rather than 50% as expected; and the true proportion of *p*-values that were less than 0.0135 was only 0.01207[11]. This explains why our simulated p-value is a bit less than the one we calculated, and also serves to illustrate the value of simulation, as effects like this are very hard to anticipate!

**But There is More to Discover…**

Before drawing any final conclusions, note that the hypothesis test for equality of proportions is not limited to comparing only two samples. In fact, it can compare n-samples, and provide the probability that the total variation between them would be as great if all were drawn randomly from populations with matching proportions. Therefore, we can easily test the hypothesis that the proportion of bad buys among all colored-cars is the same.

```
> prop.test(c(1843, 1506, 1189, 911, 858, 825, 737, 402, 260, 211, 56, 34, 56, 34, 29),
+ c(14875, 12123, 10347, 7887, 7627, 6257, 5231, 3194, 2046, 1584, 436, 415, 373, 244, 242),
+ alternative="two.sided")$p.value
[1] 5.149562e-06
```

Surprisingly, this test results in a p-value of 0.00000515! In other words, the test suggests that there is almost no chance that the variation in bad buy proportion by car color is the result of random variation

[11] Not coincidentally, note that 1 – (1 – 0.01207)^15 = 16.7% which is much closer to our simulated p-value of 16.2%

alone. Now to understand why this would be, let us look more closely. Notice the *p*-values we obtain if we apply a two-sided test for equality of proportions between each color and all other colors in the original dataset (Table 6):

Table 6: P-Values for Proportion of Bad across all Colors (Looking for either High or Low)

| Row | Color | Count | Bad Buys | Percent | P-value |
|---|---|---|---|---|---|
| 1 | SILVER | 14875 | 1843 | 12.39% | 0.66220 |
| 2 | WHITE | 12123 | 1506 | 12.42% | 0.61504 |
| 3 | BLUE | 10347 | 1189 | 11.49% | 0.00858 |
| 4 | GREY | 7887 | 911 | 11.55% | 0.03786 |
| 5 | BLACK | 7627 | 858 | 11.25% | 0.00393 |
| 6 | RED | 6257 | 825 | 13.19% | 0.02398 |
| 7 | GOLD | 5231 | 737 | 14.09% | 0.00004 |
| 8 | GREEN | 3194 | 402 | 12.59% | 0.61110 |
| 9 | MAROON | 2046 | 260 | 12.71% | 0.57452 |
| 10 | BEIGE | 1584 | 211 | 13.32% | 0.21678 |
| 11 | BROWN | 436 | 56 | 12.84% | 0.77514 |
| 12 | ORANGE | 415 | 34 | 8.19% | 0.01351 |
| 13 | PURPLE | 373 | 56 | 15.01% | 0.12541 |
| 14 | YELLOW | 244 | 34 | 13.93% | 0.49007 |
| 15 | OTHER | 242 | 29 | 11.98% | 0.96532 |
| 16 | TOTAL | 72881 | 8951 | 12.28% | |

As before, we see that orange has a *p*-value of 0.0135. Yet, surprisingly, this is the 4$^{th}$ lowest (most interesting) *p*-value! In fact, the *p*-value for gold is over 300 times as significant as that of orange! No color has a proportion as different from the mean as orange does, but when sample size is accounted for, we find that the observed proportions for blue, black, and gold are all more *unusually* extreme than that of orange. In fact, the proportion for gold appears to be so unusually high that even with the vast search effect it would appear highly improbable to have occurred by chance.

What then does this mean? Are we to conclude that gold cars are bad buys? Well first, note that statistical significance does not necessarily correspond to practical significance. The observed proportion of bad buys among gold cars is 14.1%, which is only 2% higher than the observed proportion in non-gold cars (12.1%). This might be useful information, but is less useful than it would be to know that the true proportion of bad buys among orange cars is actually ~ 4% lower than that of non-orange cars. Second, we recognize that it is possible that this difference in proportion is attributable to some sampling bias or correlated factor, and not to car color, per se. For example, perhaps bad buy proportion varies with the age of the vehicle, and car color preferences tend to vary with time. Nevertheless, it seems convincing that there is *something* non-random in the relationship between car color (especially gold) and bad buy proportion, and it would likely be worth further investigation to find the reason(s).

**Conclusions**

Note that the truly interesting result (gold) was not identified originally, but orange was, due to the visualization we employed (Figure 1). The visualization was entirely appropriate and accurate, but susceptible to the small-sample effect so it led us astray. Only by testing using p-values, which take into account the sample size, did we learn that there were 3 colors more statistically interesting than the visual outlier color orange. But then we learned to not stop at the p-values, or trust them as indicators of likelihood, since we didn't approach the data with well-formed hypotheses to test, which is what p-values were designed for. Any effort to account for the fact that we have both developed and tested our hypothesis using the same data is imperfect.

Better, is to estimate the probability that *some*-colored cars would have a proportion as unusual as was observed in orange. We tried both mathematical inference and simulation (target shuffling), and checked for unusually low or extreme (low or high) proportions. Our results are in Table 7.

Table 7: Summary of 1-Tailed (as Low as) and 2-Tailed (as Extreme as) Tests by Two Methods

| | **As Low As** | **As Extreme As** |
|---|---|---|
| **Mathematical Inference** | 0.097 | 0.184 |
| **Target Shuffling** | 0.072 | 0.164 |

To assess the true interestingness of orange, we believe that the most realistic probability result is that using target shuffling and extremes (16.4%). It measures the probability of *some* color obtaining a proportion at least as unusually extreme as was observed for orange, if the underlying reality is that there is no relationship with color. We may find this worth testing on new data, or even acting on, but it is by no means as unusual a finding as we first suspected visually, or even after using our first statistical test (0.675%)!

Still, based on the equality of proportions test on all colors, it is highly likely that there is *some* relationship between car color and bad buy proportion (p-value of 0.00000515). Using *p*-values to account for sample sizes, 3 other colors were identified as more interesting (statistically) than orange. Further investigation would be required to reveal whether color is fate or, really, whether color is confounded with a more meaningful variable. Of course, if the relationship holds up out-of-sample, it may be worth acting on whether we are satisfied with an explanation or not!