### “I am curious if you could maybe explain me some more about the statistical parts (i.e., against which significant value are you explaining? 0.05 I presume?) and how you established that stuff.”

Let’s take the example of reviews for bunq containing the term “Apple Pay”. We want to know if these reviews are generally positive or negative. One obvious way is just taking the mean review rating; i.e.

*Mean rating for apple pay = ( 5 x (number of 5 star reviews that mention “Apple pay”) +*

*4 x (number of 4 star reviews that mention “Apple pay”) +*

*… +*

*1 x (number of 1 star reviews that mention “Apple pay”) )*

*/(Total number of reviews that mention “Apple pay”)*

But one problem of this is that it doesn’t take into account the selection bias inherent in reviews, that is to say people only tend to review things if they are very dis/satisfied with a product or service, so the mean rating might not accurately represent the mean population rating (i.e. the hypothetical mean rating if every member of the population who purchased the product was required to rate it).

One way we can resolve this is by asking a slightly different question from the data. That is, we don’t really care so much about the absolute review rating for a given feature, rather the thing which interests us is how the ratings for that feature compares with ratings for the app on the whole. For example, if an app has generally scored good reviews, then having a feature who’s reviews are around 3 star would be a sign that this feature isn’t positively received, whereas if an app generally has poor reviews, then a 3 star feature might be seen as a positive feature of an otherwise underwhelming app. Calculating this is straightforward, namely:

*Rating difference = (Mean rating for apple pay) – (Mean rating for NOT apple pay)*

where *(Mean rating for NOT apple pay)* means all ratings which do not mention Apple Pay.

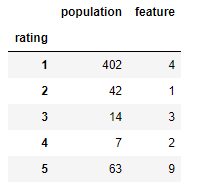
Now we have a number which we think captures how well a rating is received; a negative value of rating difference means the feature was generally badly received, whereas a positive value of rating difference means that the feature was generally well received. The next question we want to ask the data is, is this rating difference *significant*?

Let’s take an extreme hypothetical example where only one person wrote a review about the Apple Pay feature, and they left a 5 star review. This would give a rating difference – calculated through the method above – of 3.35. We would be quite confident in saying that this rating difference, even though it is quite large, is not significant. That is, we know that one person’s view alone isn’t enough to give a significant rating difference.

But what do we do in situations outside of this trivial case, how do we decide if ratings differences we calculate are significant? Statistically speaking, the question we are asking is “Is the distribution of ratings for reviews which mention apple pay significantly different from the distribution of ratings which do not mention apple pay for reviews of this app?”. We are on firm statistical footing with this question, and all that’s left for us to do is to find the most appropriate statistical test to check for this difference between distributions.

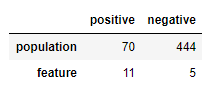
In searching for the correct test, we must first look at the type of data we have at hand. The ratings for a review are ordinal variables, as they have a clear ordering (e.g. a three star rating is more positive than a two star rating), but no well-defined distance between the categories (e.g. a four star rating is not necessarily twice as good as a two star rating). When we aggregate such reviews we’re left with *categorical data*. This is easy to forget, as the fact that the star ratings are given by numbers means that we can calculate some statistics which are associated with continuous data (e.g. the mean rating we calculated above) on them.

Our next consideration is sample size. Some tests only provide valid results in the limit of large numbers, and quasi-normal distributions, whereas others are quite happy in the sparsely populated, distribution-free world of small sample sizes. Let’s have a look at the star ratings for Apple Pay:



On the right we see the number of reviews for each rating of reviews involving the term “apple pay”. In the left column are the respective values for reviews which do not mention this term. It’s clear looking at the right column that we’re in the small-sample size regime; we have 19 observations split over 5 categories.

Our requirements are now set; we need a test for the difference between two distributions of categorical variables which behaves itself in the limit of small sample sizes. A test which satisfies these requirements is Fisher’s exact test ([Fisher's exact test - Wikipedia](https://en.wikipedia.org/wiki/Fisher%27s_exact_test)). So all that’s left for us to do is to calculate the RxC contingency table for our feature of interest, just like we did for Apple Pay in the table above, then put this into a fishers exact calculator! However, sadly for us in the scipy library in python, only 2x2 contingency tables are allowed[[1]](#footnote-1), so we have to perform an extra step of converting our rating table to a positive/negative review table. The way we do this is by classifying any 1 or 2 star reviews as negative and 4 or 5 star reviews as positive, so the above table becomes:



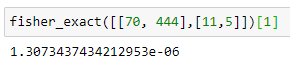
Once we’ve made this table we can use the fisher\_exact function from scipy to run the fisher exact test. The test returns a p-value. Let’s talk about what the p-value means in the context of the above table, where we see there are a total of 514 reviews which do not mention apple pay (labeled “population”) and 16 which do mention apple pay (labelled “feature”). We consider the population as being drawn from some “review probability distribution”. This probability distribution is very simple for the above 2x2 table, it’s just specified by the density of reviews which are negative or positive.

Perhaps it’s useful to think of the probability distribution we’d have for a fair coin, there we’d have 0.5 probability for heads, and 0.5 probability for tails. If we were to take a series of coin flips, what we’re essentially doing is taking multiple samples from this “0.5 heads 0.5 tails” probability distribution. It doesn’t mean that when we flip a coin 514 times we get 257 heads and 257 tails, but it means that that outcome is far more likely than getting 514 heads for example. Similarly, we might judge from our data that the probability distribution for a review about bunq bank (that doesn’t mention apple pay) is something like 0.14 (≈ 70/514) probability for positive and 0.86 (≈ 444/514). This actual distribution is unknown, and is not directly measurable, but the more reviews we have, the closer the population approximates this distribution.

Now we ask the question, if we were to sample 16 times from this distribution of reviews, what is the probability that we would get 11 positive and 5 negative, or a situation more extreme (e.g. 12 positive 4 negative)? This is exactly the p-value which is being calculated. So a high p-value (p ≈ 1) tells us, there’s a very high probability that the 16 reviews for the feature were drawn from the same distribution as those from the population. Conversely a low value (p ≈ 0) tells us that it’s very unlikely that these 16 feature reviews were drawn from the population distribution, i.e. the feature reviews are of a fundamentally different nature (= drawn from a different distribution) from the population reviews.

Clearly this number can vary anywhere between 0 and 1, so we have to choose a – somewhat arbitrary! – point on this line and so we go with convention and choose 0.05. If the p-value calculated from the fisher exact test is greater than 0.05, then we say that we are not confident enough that there is a difference between the ratings probability distribution for ratings involving the feature and the distribution for ratings not involving the feature, and therefore any difference in mean ratings is not significant. That is, there may well be a difference in mean ratings between feature and population, but this deviation could fairly be expected to be the result of the random nature of sampling from two equal populations. Conversely, if the p-value from the fisher exact test is less than 0.05, then we say that we are confident that there is a difference between the distribution of ratings probability distribution for ratings involving the feature and the distribution for ratings not involving the feature, and therefore any difference in mean ratings *is significant*.

For the concrete example of apple pay, calculating the fisher’s exact test gives a p-value:



And, given that 1.307e-6 is less than 0.05, we say that the mean rating difference is significant. That is, we conclude that, to the 0.05 significance level we are confident that reviews about apple pay are different from reviews not about apple pay (and it turns out this difference is positive because the mean review about apple pay is around 1.86 higher).

1. Fisher’s test is valid for arbitrary RxC table sizes, this is just the way scipy decided to implement it. If you’re comfortable with R, there are better stats packages there which don't have this 2x2 restriction [↑](#footnote-ref-1)