2nd AA Project — Approximate Counting

Rui Fernandes

***Abstract* –This report seeks to present how the author coded and tested approximate counters both with fixed and decreasing probability. The chosen coding language was Python(3.7), and the algorithms will be shortly explained before diving into an exposition of results.**

***Keywords* –Counter, Approximate Counting, Fixed Probability, Decreasing Probability**

I. Introduction

The problem presented for this project and that this report seeks to explain is the application of approximate counters and comparison with an exact counter.

The corpus used for testing the counters, consisted in the manuscript of "The Odyssey" by Homer, in 3 different languages: english, spanish and french.

To test the counters, they were run 10000 times each per language file.

With the report comes the code files used and the result files obtained, in the form of the exact counts of letters and 2 sets of files per language: '_results' files that have the letter counting of each trial done, and '_stats' files that include all the statistics of the corresponding results.

To run the code one should first get the exact count of a file, then run the desired counter(changing of iteration number is only possible manually). And to compile the statistics one needs to run the 'build_stats.py' script with the first argument as the original file used, the second as the results file, and the third as 'F' or 'D' in relation to fixed or decreasing probability

For example:

```
$ python exact_counter.py 'Odyssey.txt'
$ python fixed_probability_counter.py
'Odyssey.txt'
$ python build_stats.py 'Odyssey.txt'
'Odyssey_fixed_prob_results.txt' 'F'
```

II. The Problem

This problem is simple in nature, and the counters weren't a challenge to make, rather it's more of a collecting and comparing of data, to verify the validity of each counter type on the texts used.

Regardless, the problem that approximate counters seek to resolve is lack of memory, in the present memory is abundant, so it isn't a problem for this scale of experiment, but for Big Data or statistics on massive databases for example, approximate counters might be a necessity.

By scaling down the amount of times the counters go up, it reduces the stress on memory, allowing us to count more data. In the case of fixed probability, a determined probability is used to decide to count or not, in the project's case it was 1/64. As for decreasing probability, the more we count on the counter, the less the probability that we count again, in this case, the following function was used:

$$1/\sqrt{2}^{k}$$

where k is the current count.

III. Results Discussion

For the results, first will be noted the interesting findings here, and after all the graphs used to gain this information.

As before stated the statistics for each counter-language pairings are stored in files ending in '_stats' and all the information exposed here comes from there.

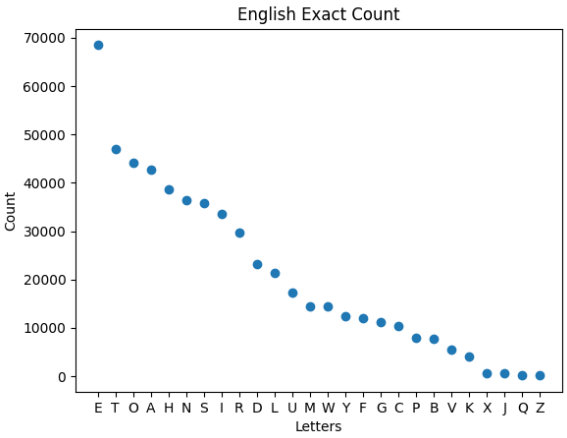
The statistics are for the general count of characters, and then some provided for the specific characters.

The most obvious note is that the fixed probability counter managed to get in mean, extremely accurate results, this is also because 10000 trials were done, if it was simply 100 it wouldn't be as spot on. On the other hand, the decreasing probability counter did quite poorly, there is the possibility that a miss calculation occurred, but there's the influence from the fact that the text analyzed wasn't big enough data to warrant the decreasing probability counter useful, as most times it doesn't even count once for every character.

Some stats not exposed on here specifically are the ones relating to keeping the correct letter frequency order, this was done by simply getting the letters ordered by frequency for each trial and saving them. With them we can see that, for the english and fixed counter example, 0.26% of trials managed to register the exact order of frequencies all the way through, that doesn't sound impressing, but for the less frequent letters, there's a lot of volatility. Looking only at the top 5 more frequent letters, it actually manages to be correct 46.43% of the time.

Another interesting observation is that throughout the 3 languages, the letter 'E' is always the most frequent, but there are some differences in the top 5, 'O' is very prominent in english and spanish, but not quite in french for example.

IV. GRAPHS

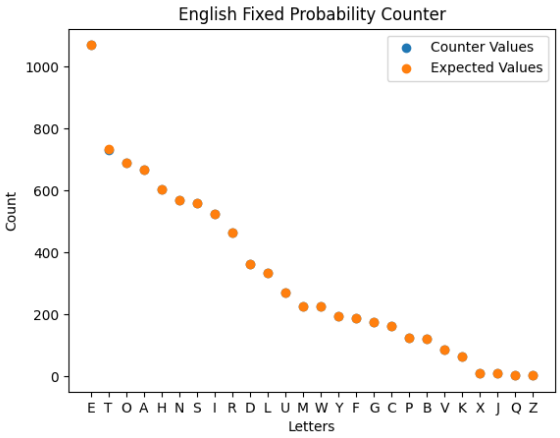


Counter Statistics For French

	Fixed	Decreasing
Expected Value	4496.266	8.431
Variance	2248.133	4.216
Standard Deviation	8.382	2.053
Mean Absolute Error	52.949	0.993
Mean Relative Error	1.178%	11.776%
Mean Accuracy Ratio	100.009%	396.934%
Smallest Counter Value	4216	30
Largest Counter Value	4735	39
Mean Counter Value	4496.678	33.467
Mean Absolute Deviation	1.455e-14	2.182e-15
Standard Deviation	66.351	1.209
Maximum Deviation	280.678	5.533
Variance	4402.393	1.463

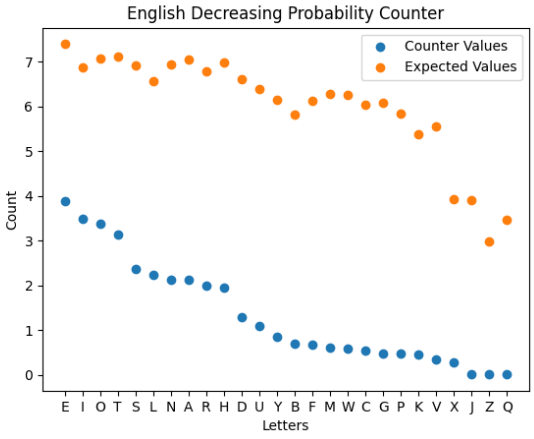
Counter Statistics For English

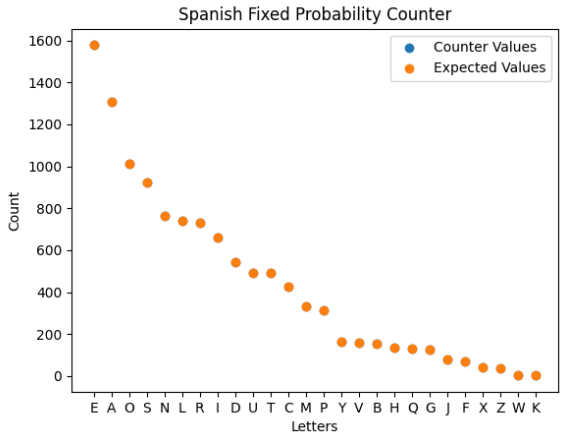
	Fixed	Decreasing
Expected Value	8442.188	8.886
Variance	4221.094	4.443
Standard Deviation	11.485	2.108
Mean Absolute Error	73.685	0.989
Mean Relative Error	0.873%	11.132%
Mean Accuracy Ratio	99.997%	397.113%
Smallest Counter Value	8122	31
Largest Counter Value	8806	40
Mean Counter Value	8441.898	35.287
Mean Absolute Deviation	4.685e-13	1.432e-15
Standard Deviation	92.017	1.219
Maximum Deviation	364.102	4.713
Variance	8467.197	1.486

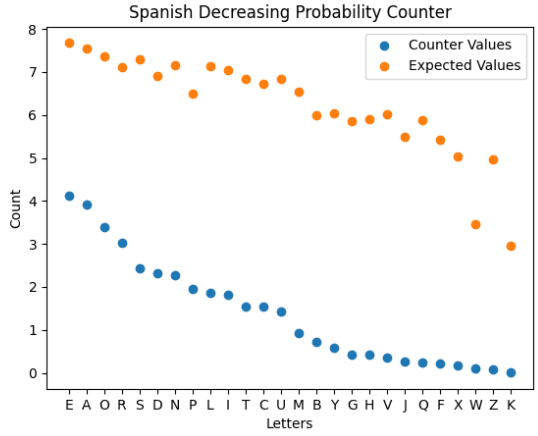


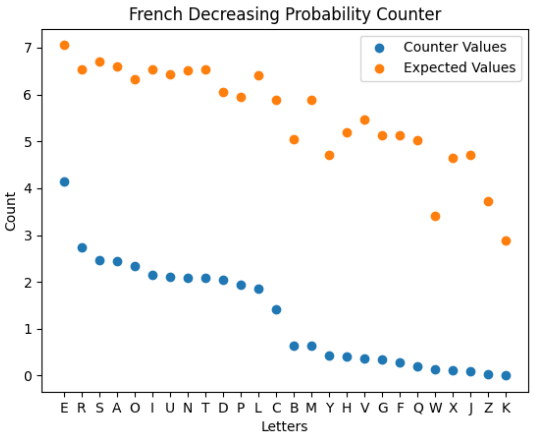
Counter Statistics For Spanish

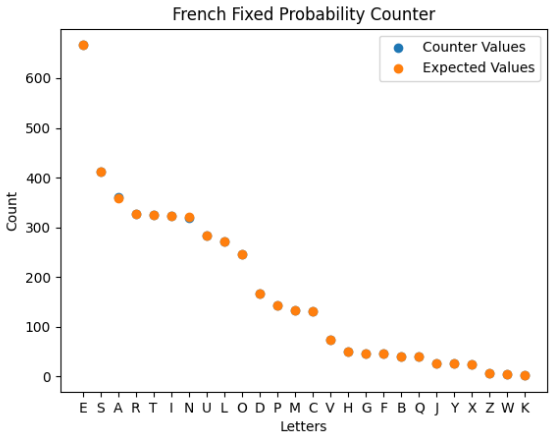
	Fixed	Decreasing
Expected Value	11412.469	9.103
Variance	5706.234	4.552
Standard Deviation	13.354	2.133
Mean Absolute Error	84.313	0.951
Mean Relative Error	0.739%	10.446%
Mean Accuracy Ratio	100.002%	397.007%
Smallest Counter Value	11046	32
Largest Counter Value	11823	41
Mean Counter Value	11412.739	36.141
Mean Absolute Deviation	8.265e-13	4.810e-16
Standard Deviation	105.935	1.212
Maximum Deviation	410.261	4.859
Variance	11222.132	1.47




Spanish Fixed Probability Counter
Counter Values
Expected Values
Count
Letters
E A O S N L R I D U T C M P Y V B H Q G J F X Z W K


Spanish Decreasing Probability Counter
Counter Values
Expected Values
Count
Letters
E A O R S D N P L I T C U M B Y G H V J Q F X W Z K


French Decreasing Probability Counter
Counter Values
Expected Values
Count
Letters
E R S A O I U N T D P L C B M Y H V G F Q W X J Z K


French Fixed Probability Counter
Counter Values
Expected Values
Count
Letters
E S A R T I N U L O D P M C V H G F B Q J Y X Z W K