Probabilistic Counters

ABLORDEPPEY Prosper

***Abstract* –The goal is to count the number of occurrences of letters in text files and, for instance, identify the most common ones. Three types of counters were analyzed. The *Exact Counter*; which provides the exact count or frequency of each letter present in the text, the *Fixed Probability Counter*, with probability $p = \frac{1}{2}$, whose counter value has a 50% chance of being increased on future occurrences. The last counting method considered, being the *Decreasing Probability Counter* where on future occurrence of a letter, the probability of increasing the counter value decreases. Given the $k'th$ occurrence, this probability is defined as $\left(\frac{1}{\sqrt{2}^k}\right)$.**

Project Outline

The outline of the project in order is given as.

- Introduction
- Project Directory Structure
- A Midsummer Nights Dream
- Conclusion

I. Introduction

The literary works considered in this project are two (2) works of **Shakespeare** with each, translated in three (3) different languages *English, French* and *German*. The titles of these works in English are;

1. A Midsummer Nights Dream
2. Hamlet

II. Project Directory Structure

The folder structure for this project is graphically given in Fig. (1). From the representation, the project has three (3) directories and the *ProbabilityCounters* python notebook file. Detailed description is presented below;

- **shakespeare directory** :- This folder hosts the two (2) literary works of Shakespeare considered in this project. In total, the folder has $2 \cdot 3 = 6$ literary works since each work has translations in English, French and German. For instance, Hamlet has variations, *hamlet-english, hamlet-french* and *hamlet-german.*

- **output directory** :- This directory contains the counter and estimate registers in *csv* format, for each of the two (2) count methods (exact, fixed and decreasing prob). Using each count method, multiple experiments were performed for each of the six (6) literary works. Fifty (50) experiments were performed for fixed and decreasing probability counters and only one (1) experiment was performed for the exact counter since the result for multiple experiments will be same. The counter and estimate values were recorded and presented in the *counter* and *estimate* named csv files respectively. Thus, there are $6 \cdot 3 \cdot 2 = 36$ csv files in this directory. For instance, the files as shown for the output directory in Fig. 1 is the output registers for hamlet-english literary work only.

  Each output counter and estimate register have columns presented in TABLE I and TABLE II respectively.

```
Project-Two(2)
├─ shakespeare
│   ├─ hamlet-english.txt
│   ├─ hamlet-french.txt
│   ├─ hamlet-german.txt
│   └─ ...
├─ output
│   ├─ hamlet-english_fixed_counter.csv
│   ├─ hamlet-english_fixed_estimate.csv
│   ├─ hamlet-english_decrease_counter.csv
│   ├─ hamlet-english_decrease_estimate.csv
│   ├─ hamlet-english_exact_counter.csv
│   ├─ hamlet-english_exact_estimate.csv
│   └─ ...
├─ combined
│   ├─ hamlet-english_counter_combined.csv
│   ├─ hamlet-english_estimate_combined.csv
│   ├─ hamlet-french_counter_combined.csv
│   ├─ hamlet-french_estimate_combined.csv
│   └─ ...
└─ ProbabilityCounters.ipynb
```

Fig. 1 - Project Directory Structure

- **combined directory**:- This folder holds the combined counters and estimates registers for each literary work.

TABLE I
OUTPUT COUNTER REGISTER

letter	minimum	STD
exp 1	maximum	VAR
· · ·	average	estimate
exp 50	average_percent	

TABLE II
OUTPUT ESTIMATE REGISTER

letter	minimum	STD
exp 1	maximum	VAR
· · ·	average	
exp 50	average_percent	

Since there are two (2) outputs for each literary work, there are in all $6 \cdot 2 = 12$ csv registers in csv format. Each *counter* csv register contains headings as outlined in TABLE III. These tabulations resulted from counter registers in the output directory.

TABLE III
COMBINED COUNTER REGISTER

letter	fixed_minimum	decrease_minimum
exact	fixed_maximum	decrease_maximum
exact_percent	fixed_average	decrease_average
exact_bits	fixed_average_percent	decrease_average_percent
	fixed_bits	decrease_bits
	fixed_estimate	decrease_estimate
	fixed_STD	decrease_STD
	fixed_VAR	decrease_VAR
	fixed_meanAbsError	decrease_meanAbsError
	fixed_meanRelError	decrease_meanRelError

Similarly, each *estimate* csv register contains headings as outlined in TABLE IV. These tabulations resulted from estimate registers in the output directory.

TABLE IV
COMBINED ESTIMATE REGISTER

letter	fixed_minimum	decrease_minimum
exact	fixed_maximum	decrease_maximum
exact_percent	fixed_average	decrease_average
exact_bits	fixed_average_percent	decrease_average_percent
	fixed_bits	decrease_bits
	fixed_STD	decrease_STD
	fixed_VAR	decrease_VAR
	fixed_meanAbsError	decrease_meanAbsError
	fixed_meanRelError	decrease_meanRelError

- **ProbabilityCounters.ipynb**
  This is the python notebook file generating the *output* and *combined* registers as detailed above. This file implements three (3) classes. Namely
  - BaseCounter (IO File Operations)
  - ApproximateCounters (Counters Implementation)
  - Combiner (Combined File Generations)

III. A Midsummer Nights Dream

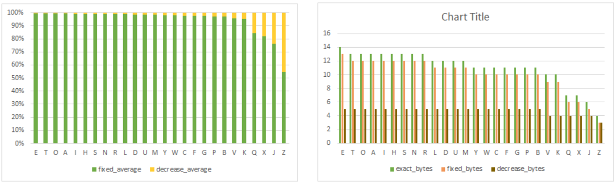
A. *English*

A.1 *counters*

Fig. 2 shows distribution of the counter values for the exact, fixed and decreasing probability counters. It can be observed that the fixed probability counter values are approximately half of the exact counter values. This is due to the fact that the fix probability of increasing counter value on new occurrence is $p = \frac{1}{2}$. For larger occurrences, the decreasing probability counter slowly updates the counter value as observed in Fig. 3(a). This explains why such counter values are insignificant as compared to the fixed probability counter values.

Again, the space complexity for the decreasing probability counter is the smallest, followed by the fixed probability counter. The minimum amount of memory (bits) required to store each counter value is also presented in Fig. 3(b). Though the minimum memory required to store the exact and fixed probability counter values are approximately equal but relatively higher for the exact counter, the decreasing probability counter has slowly increasing memory size, as shown.

Fig. 2 - A Midsummer Nights Dream - English Counters



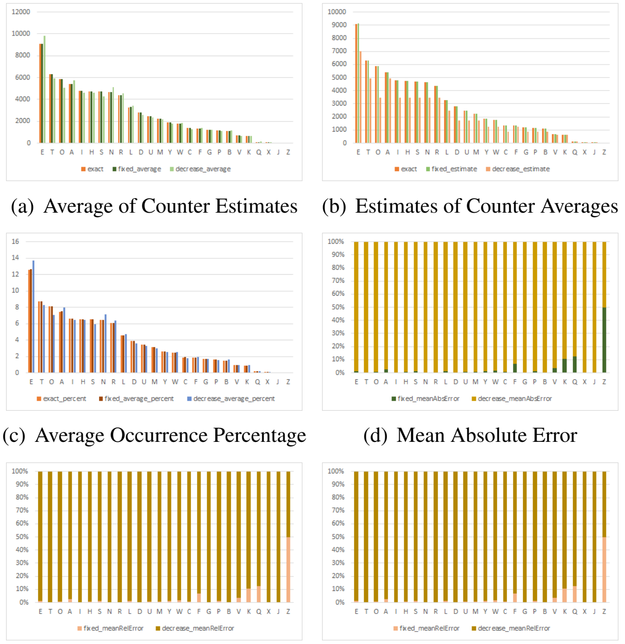
(a) Counter Average (b) counter minimum space (bits)

A.2 *estimates*

Fig. 3 presents the distribution of the estimates. As shown in Fig. 4(a), and Fig. 4(b), estimates computed from taking the average of the counter estimates for all experiments (thus, 4(a)) tends to be more closer to the exact number of occurrence for the decreasing probability counter. Both approaches provide similar result for the fixed probability counter. The order of the counters is preserved in both cases.

Fig. 3 shows the average percentage of occurrence for each letter for the various count methods. Using the exact values, the top four (4) most occurring letters are *E-12.5%, T-8.7%, O-8.1%* and *A-7.5%*. This ranking is preserved for the fixed and decreasing probability counters. Although the decreasing probability estimates recorded slightly higher occurrence percentages, the distribution of these estimates is similar to that of the exact and fixed probability counters.

Fig. 3 - A Midsummer Nights Dream Estimates - English Estimates



(a) Average of Counter Estimates (b) Estimates of Counter Averages

(c) Average Occurrence Percentage (d) Mean Absolute Error

(e) Mean Relative Error (f) Variance

These percentages are consistent with the known standards.

Just as with the mean absolute errors, we observe larger mean relative errors for the decrease probability counter as compared to the fixed probability counter. This is so because, the decreasing probability counter used very less memory for counting as opposed to the fixed probability counter, thereby, recording larger errors.

We also observe very high variation in estimates for the decreasing probability counter as opposed to the fixed probability counter, as shown in Fig. 3.

B. French

Fig. 4 presents the distribution of the counter values, and Fig. 4 shows the corresponding estimates distribution.

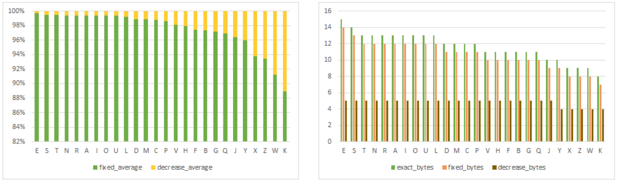
As shown in Fig. 4, estimates computed from taking the average of the counter estimates for all experiments also happens to be better estimates than the estimates computed from the average counter values. Both approaches provide similar result for the fixed probability counter. Although there are a few disagreement with the ranking of the most occurring letters for the decrease probability counter estimates, on average, the order is preserved.

Fig. 5 shows the average percentage occurrence for each letter for the various count methods. Using the exact values, the top four (4) most occurring letters are *E-15%, S-7.8%, T-7.2%, N-7.1%* and *R-7%*. The least occurrence being X, Z, W, K. These rankings are preserved for the fixed and decreasing probability counters for all letters. Although the decreasing probability estimates recorded slightly higher occurrence percentages, the distribution of these estimates is similar to that of the exact and fixed probability counters. These rankings are consistent with known standards.

Just as the mean absolute errors, Fig. 5 shows larger mean relative errors for the decrease probability counter estimates as compared to the fixed probability counters. This is so because, the decreasing probability counter used very less memory for counting as opposed to the fixed probability counter, thereby, recording larger errors.

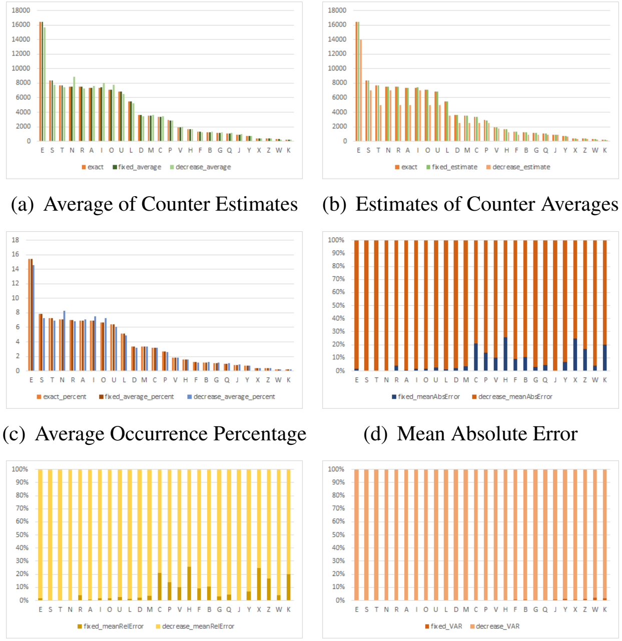
We also observe very high variation in estimates from the decreasing probability counter as opposed to the fixed probability counter as shown in Fig. 5.

Fig. 4 - A Midsummer Nights Dream Counters - French Counter



(a) counter average (b) counter minimum space (bits)

Fig. 5 - A Midsummer Nights Dream - French Estimates



(a) Average of Counter Estimates (b) Estimates of Counter Averages

(c) Average Occurrence Percentage (d) Mean Absolute Error

(e) Mean Relative Error (f) Variance

C. German

Fig. 6 shows the distribution of the counter values, and Fig. 7 shows the corresponding estimates distribution.

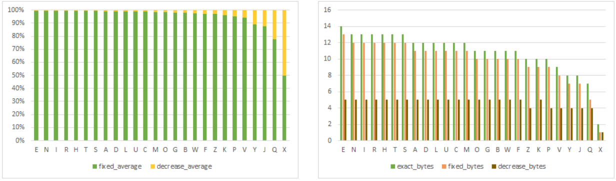
As shown in Fig. 6, estimates computed from taking the average of the counter estimates for all experiments also happens to be better estimates than the estimates computed from the average counter values. Both approaches provide similar result for the fixed probability counter. Although there are a few disagreement with the ranking of the most occurring letters for the decrease probability counter estimates, on average, the order is preserved.

Fig. 7 shows the average percentage occurrence for each letter for the various count methods. Using the exact values, the top four (4) most occurring letters are *E-15.6%, N-9.7%, I-8.7%, R-6.8%* and *H-6.3%*. The least occurrence being Y, J, Q, X. These rankings are preserved for the fixed and decreasing probability counters for all letters. Although the decreasing probability estimates recorded slightly higher occurrence percentages, the distribution of these estimates is similar to that of the exact and fixed probability counters. These rankings are consistent with known standards.

Just as the mean absolute errors in Fig. 7 shows larger mean relative errors for the decrease probability counter as compared to the fixed probability counters. This is so because, the decreasing probability counter used very less memory for counting as opposed to the fixed probability counter, thereby, recording larger errors.

We also observe very high variation in estimates from the decreasing probability counter as opposed to the fixed probability counter as shown in Fig. 7.

Fig. 6 - A Midsummer Nights Dream Counters - German Counter



(a) counter average (b) counter minimum space (bits)

IV. Hamlet

A. *English*

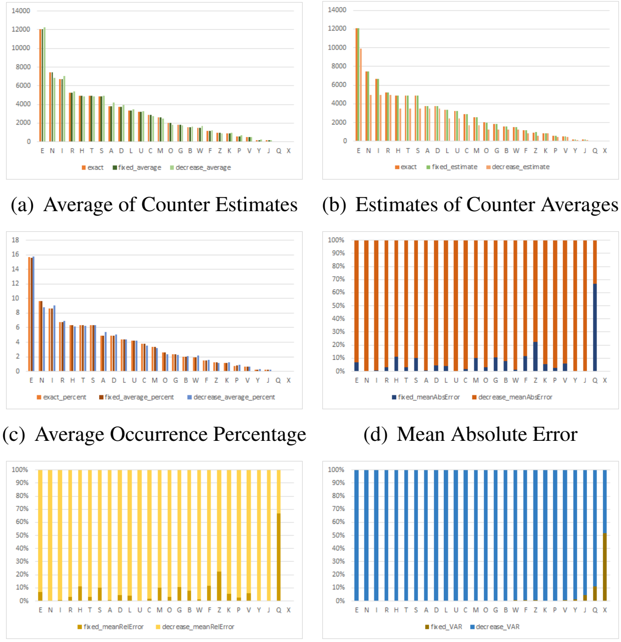
Fig. 8 shows the distribution of counter values for each letter, and Fig. 9 also shows the distribution of the estimates for each count method. *E,A,O,S,R* happens to be the top five (5) most occurring letters, with percentages *13.6%, 11.1%, 9.3%, 7.8* and *6.9%* respectively.

There is not significant variation between the estimates from the fixed and decreasing probability estimates. Letters K and W are have similar variations for both counters.

The estimates computed from taking the average of estimates for all experiments seem to be a better estimate than those computed via the counter averages, although on average, they record slightly higher values compared to the exact counters values.

The relative errors observed for the fixed probability counter are generally low compared to those from the exact counter, but we see an abnormally high error recorded for letter V for the fixed probability counter.

Fig. 7 - A Midsummer Nights Dream - German Estimates



(a) Average of Counter Estimates (b) Estimates of Counter Averages

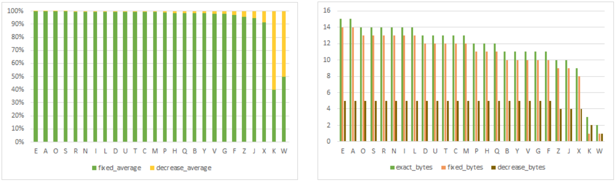
(c) Average Occurrence Percentage (d) Mean Absolute Error

(e) Mean Relative Error (f) Variance

From the above, we confirm from the graph the high variation in the estimates for the decreasing probability counter. For letters K and W which occurs the least, we notice similar variation of the computed estimates.

Again, the memory size in bits required to store the counters for the decreasing probability counter is very small and increases slowly for larger occurrences.

Fig. 8 - Hamlet Counters - English Counter



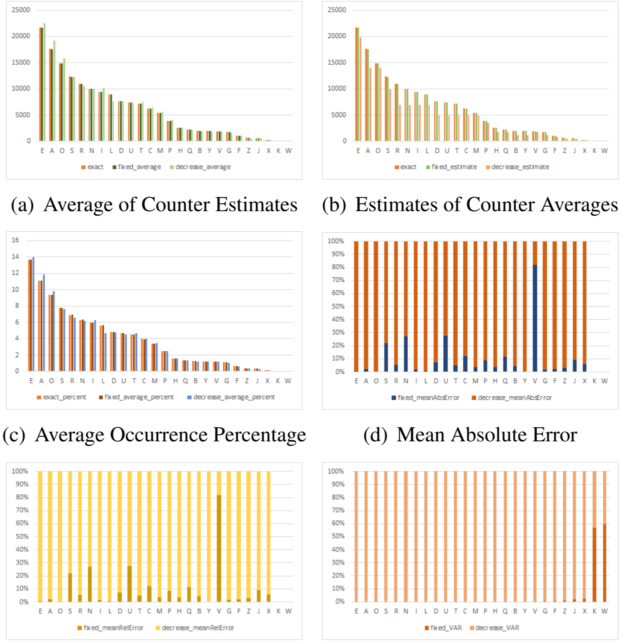
(a) counter average (b) counter minimum space (bits)

B. *French*

Fig. 10 shows the distribution of counter values for each letter. The top five (5) most occurring letters are *E,S,A,N,T*, happens to occur with percentages *16%, 8.3%, 7.4%, 7.3%* and *7.2%* respectively. The least five occurring letters are *X, Z, Y, K, W* .The estimates closer to the exact counter values are those computed from the average of the counter estimates as observed in Fig. 11(a). Most variation in the fixed probability counter estimates was recorded for estimate for letter F.

Variation in the decrease probability counter estimates is very high for all records.

Fig. 9 - Hamlet - English Estimates

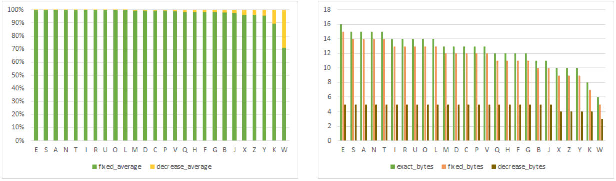


(a) Average of Counter Estimates (b) Estimates of Counter Averages

(c) Average Occurrence Percentage (d) Mean Absolute Error

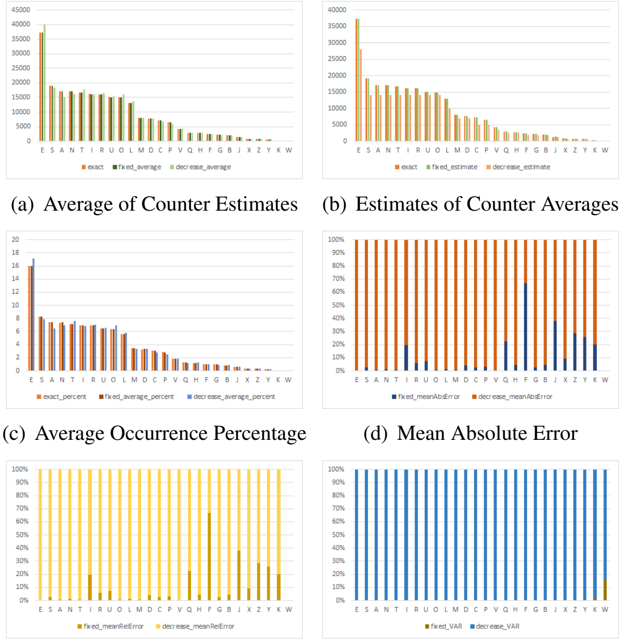
(e) Mean Relative Error (f) Variance

Fig. 10 - Hamlet Counters - French Counter



(a) counter average (b) counter minimum space (bits)

Fig. 11 - Hamlet - French Estimates

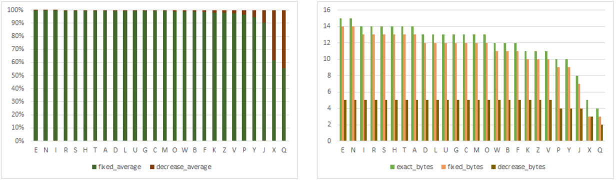


(a) Average of Counter Estimates (b) Estimates of Counter Averages

(c) Average Occurrence Percentage (d) Mean Absolute Error

(e) Mean Relative Error (f) Variance

Fig. 12 - Hamlet Counters - German Counter



(a) counter average (b) counter minimum space (bits)

C. German

Fig. 12 shows the distribution of counter values for each letter. The top five (5) most occurring letters are *E,S,A,N,T*, happens to occur with percentages *16%, 8.3%, 7.4%, 7.3%* and *7.2%* respectively. The least five occurring letters are *W, K, X, J, Z* .The estimates closer to the exact counter values are those computed from the average of the counter estimates as observed in Fig. 13. Most variation in the fixed probability counter estimates was recorded for the letter W.

The counter averages for the decreasing probability counter increases slowly for higher occurrences, as shown in Fig. 12. The minimum memory (bits) is also constant for large occurrences.

V. Conclusion

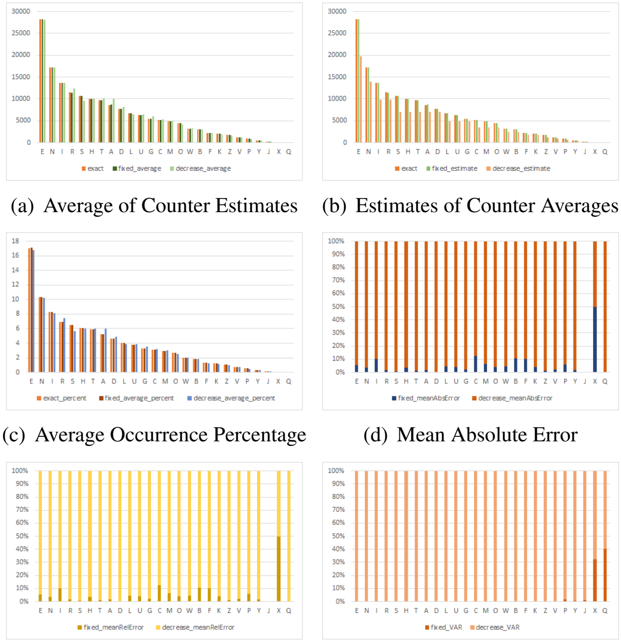
From the above results and analysis, we arrive at the following conclusion;

- Estimates computed by taking the average of all estimates yields better result as compared to those obtained from average of the counter values.
- The top (10) most frequent English letters in a standard text documents are $\{E,T,A,O,I,N,S,R,H,L\}$, confirmed by our results in this presentation [1].
- Also, the first ten (10) most frequent French letters as observed in the analysis are $\{E,S,A,N,T,I,R,O,U,L\}$. These conforms with the standard most frequent letters used in a typical French text.
- Finally, the first (10) most frequent German letters in a standard text in German are $\{E,N,I,S,R,A,T,D,H,U\}$. These are also confirmed by the presentation in this project.
- The minimum memory (bits) as realized in this presentation to store decreasing probability counter values is very small as compared to the fixed probability counter and the exact counter. The fixed probability counter considered in this implementation requires minimum space approximately equal to that of the exact counter, just marginally lower.

References

[1] letterfrequency.org. letter-frequency-by-language.

Fig. 13 - Hamlet - German Estimates



(a) Average of Counter Estimates

(b) Estimates of Counter Averages

(c) Average Occurrence Percentage

(d) Mean Absolute Error

(e) Mean Relative Error

(f) Variance