# Hypothesis B - Approximate Hit Count: Random Strings

Luís Fonseca nº 89066

***Resumo* – No presente relatório é feito um estudo a 2 tipos de contadores probabilísticos, no sentido de perceber como estes funcionam e em que situações são mais bem aplicados. Neste contexto os contadores abordados são um contador com probabilidade fixa de 1/16 e um contador logarítmico de base 2. Estes são examinados em função de duas variáveis principais, de forma a facilitar a analise e eventual comparação entre os mesmos.**

***Abstract* – This report studies two probabilistic counters seeking to have a better understanding on how they work and when they are a best fit for a situation. Two main variables are object of study to ease the analysis and comparation between the counters.**

## I. CONTEXT

Algoritmos de contagem aproximada, mais comumente conhecidos por contadores probabilísticos, permitem a contagem de um grande número de eventos usando uma pequena quantidade de memória. Como o nome indica usam técnicas probabilísticas para aumentar o valor de um contador, permitindo obter teoricamente valores bastante próximos dos reais. É uma técnica com inúmeras áreas de aplicação atualmente (motores de pesquisa, compressão de informação, inteligência artificial, etc.), pois assenta na necessidade de se trabalhar com quantidades enormes de *data* incapazes de caber em memória.

A nível de funcionamento são conceptualmente bastante simples. Primeiramente é atribuída uma probabilidade a um contador, podendo esta ser estática – isto é, não sofre qualquer variação – ou dinâmica – ou seja, varia em função de algum fator. De seguida, começam-se a contar os valores: quando um evento acontece, só um certo número de vezes (dependendo da referida probabilidade) é que esse evento é efetivamente contado/incrementado. Quando seja necessário saber o valor real, basta realizar algum tipo de operação, mais ou menos complexa, sobre o valor atual, de forma a alcançar o valor real, aproximado.

## II. CONSIDERATIONS

Para o estudo dos seguintes contadores, foram analisados diversos parâmetros, como se pode ver na figura 1.

Contudo, com o intuito de simplificar a analise e tendo em conta o contexto do relatório, foram apenas considerados 2 parâmetros principais: *mean accuracy* e *mean relative error* – onde por vezes se omitirá a notação de "*mean*" no sentido de tornar a leitura mais fluida.

```
Exact Counter:
{'a': 2, 'l': 3, 'd': 2, 'n': 2, 'e': 1}
1/16 Counter:
Trials:  10000 ; Len:  10
```

| Letter | ExpVal | Mean-AE | Mean-RE | Max-RE | Min-RE | Mean-AcR | SmallVal | BigVal | MeanVal | MAD | StdDev | MaxDev | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 2 | 3.522 | 176.1% | 1500.0% | 100.0% | 100.6% | 0 | 32 | 2.0 | 3.532 | 29.99 | 5.453 | 29.74 |
| l | 3 | 4.904 | 163.5% | 1500.0% | 100.0% | 97.8% | 0 | 48 | 2.9 | 4.86 | 45.07 | 6.68 | 44.62 |
| d | 2 | 3.592 | 179.6% | 1500.0% | 100.0% | 105.1% | 0 | 32 | 2.1 | 3.669 | 29.9 | 5.587 | 31.22 |
| n | 2 | 3.489 | 174.4% | 1500.0% | 100.0% | 98.3% | 0 | 32 | 2.0 | 3.463 | 30.03 | 5.421 | 29.39 |
| e | 1 | 1.825 | 182.5% | 1500.0% | 100.0% | 94.2% | 0 | 16 | 0.9 | 1.774 | 15.06 | 3.767 | 14.19 |

**Figura 1 - Exemplificação de Output do Programa**

Como principal justificação para esta aproximação, está o facto de se procurar trabalhar com as métricas de Accuracy - auto-explicativa - e de Precisão, como descrito em [1]. Assim, *Accuracy* servirá para indicar se nos é possível aproximar do valor real por aproximação através da média, e *Precision*, que estando relacionado com "Repeatability", ou seja, a variação de valores quando a mesma operação é repetida nas mesmas condições, indicará o quão perto os valores obtidos estão uns dos outros. A figura seguinte apresenta uma visualização que pode facilitar a interpretação destes conceitos.

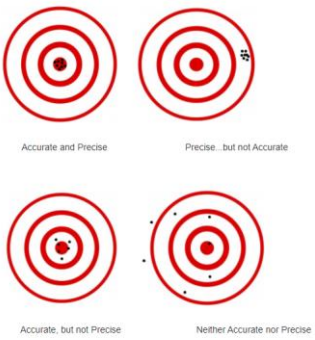


**Figura 2 - Visualização de Accuracy e Precision**

Assim, para o cálculo da Accuracy temos:

$$acc = \frac{valor\ estimado\ pelo\ contador}{valor\ exato}$$

E o de Erro Relativo:

$$acc = \frac{|valor\ estimado - valor\ exato|}{valor\ exato}$$

Eventualmente outros parâmetros são referidos, aquando da tentativa de justificação de dado fenómeno, mas de forma bastante breve.

## III. FIXED COUNTER - 1/16 PROBABILITY

O primeiro contador a ser estudado é um contador fixo, de probabilidade 1/16 (por vezes nomeado "prob16", para simplificação). Fixo, como vimos, no sentido em que a probabilidade do mesmo é estática, ou seja, não varia sobre nenhuma condição.

Conceptualmente funciona de forma bastante simples: aquando da possibilidade de contar um evento, este é apenas contado 1/16 das vezes.

### A. *Variating Trials*

#### I. Small Sequence

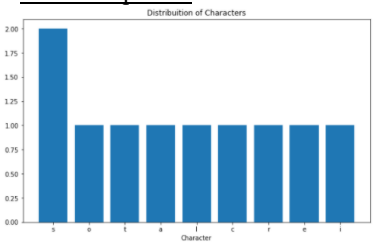


**Figura 3 - Distribuição de Caracteres para Sequencia Pequena**

O primeiro aspeto que se procurou estudar, foi o efeito das tentativas, ou seja, execuções que se faziam e a sua influência nos resultados obtidos. Desta forma, inicialmente começou-se com uma sequência de tamanho bastante reduzido (10 caracteres de comprimento), tendo sido então incrementado o valor de *trials* sucessivamente. A distribuição de caracteres da sequência gerada encontra-se no gráfico 3, que como se pode observar, têm maioritariamente uma ocorrência apenas.

*Accuracy*

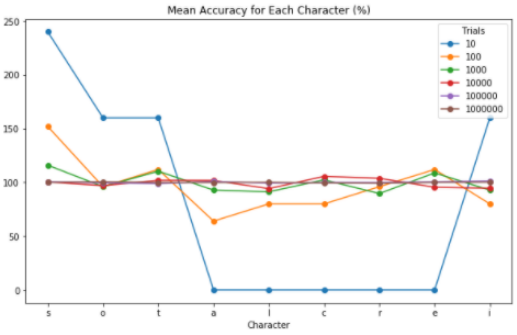


**Figura 4 - Accuracy Média para Sequencia Pequena**

Note-se que tendo em conta que a *Accuracy* corresponde à razão entre o valor obtido pelo valor esperado, espera-se que os melhores resultados estejam à volta dos 100%.

Assim, observando o gráfico anterior, consegue-se perceber rapidamente que com o aumento do número de *trials*, a *accuracy* (média) tende cada vez mais para os 100%, ou seja, vai-se aproximando do valor real.

*Relative Error*

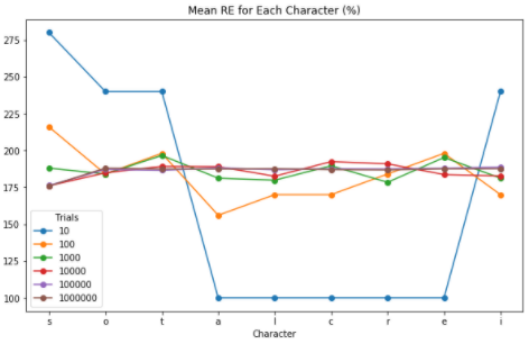


**Figura 5 - Erro Relativo Médio para Sequencia Pequena**

Contudo, apesar da *accuracy* tender para valores próximos do considerado "ideal", outro parâmetro apresentou um comportamento peculiar: a média do Erro Relativo, não tendia para 0, mas sim para os 187%. Interessante salientar aqui o facto do seguinte gráfico ser, no que diz respeito ao comportamento das "linhas", igual ao anterior, divergindo apenas nos valores do eixo das ordenadas e na variável em questão.

Ora, como explicado anteriormente [1][2] usa-se o Erro Relativo como principal medida de Precisão. Desta forma, tendo em conta que os valores do erro relativo estão bastante acima dos 0% (e até mesmo dos 100%), este contador, para um número reduzido de eventos é em geral pouco preciso e apenas para valores relativamente grandes de *trials* é que apresenta accuracy muito boa.

Uma justificação por detrás deste fenómeno, está no facto de que, uma vez que o contador considera os fenómenos 1/16 vezes, na maioria dos casos apenas apresentará os valores 0 ou 16. Como visto no gráfico 3, as letras aparecem maioritariamente 1 vez apenas, pelo que entre 1 e 16 encontramos uma razão considerável.

#### II. Big Sequence

De seguida, estudou-se o comportamento para um sequencia de tamanho maior, no caso, 10000 caracteres de comprimento.

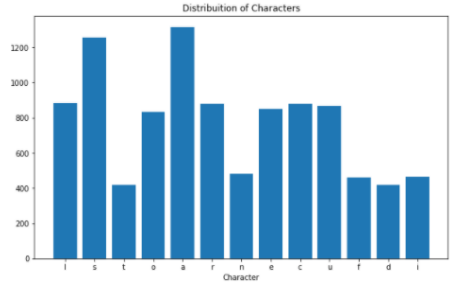


**Figura 6 - Distribuição de Caracteres para Sequencia Grande**

A sequencia gerada apresenta a distribuição do gráfico 6. Em jeito de comparação, apresenta-se também a distribuição de caracteres da sequência "luiscarlosduartefonseca" a partir do qual todas as sequências são geradas, no gráfico 7.

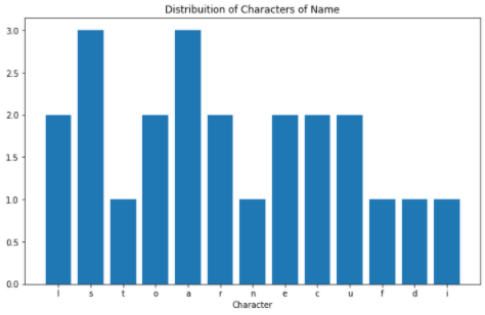


**Figura 7 - Distribuição de Caracteres do Nome Gerador**

Como expectável e comparando as 2 visualizações, para este número já considerável de caracteres, a distribuição dos mesmos já se apresenta idêntica à do nome, com os caracteres mais frequentes deste mais representados na sequência gerada, e vice-versa.

*Accuracy*

Observando o gráfico seguinte, à primeira vista tem-se a ideia de que os resultados são piores que no caso de menos caracteres. Contudo atentando ao eixo das ordenadas verifica-se que não é o caso. Estas oscilações bruscas que se podem ver, ocorrem em intervalos bastantes pequenos de valores, pelo que se tornam um pouco negligenciáveis.

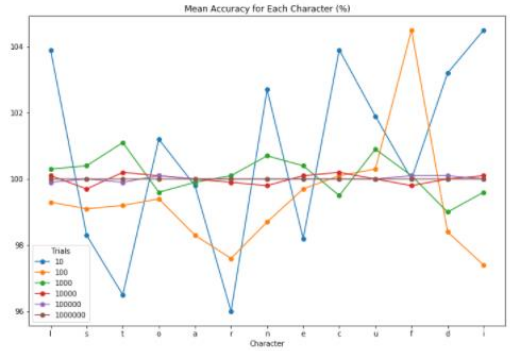


**Figura 8 - Accuracy Média para Sequencia Grande**

Foi então feita uma tentativa de justificar os picos notórios para trials=10, e apesar de em bastantes casos os picos ocorrerem em caracteres de menor frequência, não é algo consistente o suficiente para retirar a responsabilidade à aleatoriedade associada.

Assim, de maneira geral, para uma *string* de maior comprimento, a *Accuracy* tende para o valor ideal mais rapidamente, uma vez que mesmo para um pequeno número de *trials*, já estamos bastante próximos desse valor. Tal parece indicar que se certa forma um aumento no comprimento da sequência resulte num comportamento semelhante ao do aumento de *trials*.

*Relative Error*

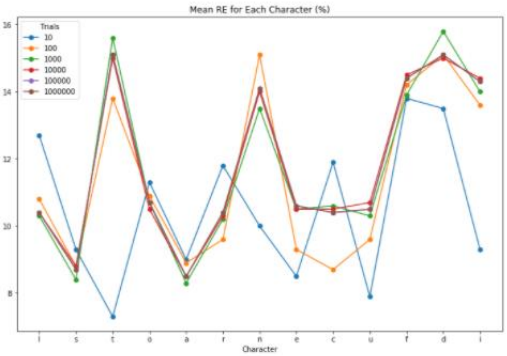


**Figura 9 - Erro Relativo Médio para Sequencia Grande**

No que respeita ao erro relativo, este apresenta-se também muito menor que para o caso de menor comprimento. Apesar das oscilações observadas, estas ocorrem entre intervalos de valores bastante reduzidos, pelo que se podem novamente considerar negligenciáveis.

Ainda algo interessante aqui é o facto de para trials ≥10, os valores apresentarem mais ou menos o mesmo comportamento. Atendendo a que caracteres correspondem, consegue-se perceber que os mínimos "globais" se encontram nas 2 letras com mais ocorrências, e de forma análoga, os máximos "globais" ocorrem nas letras de menor frequência (ou seja, "d", "n", "t").

Importante, contudo, reforçar que se está a comparar a nível de Erro Relativo. Isto porque a nível de erro absoluto, para o caso de menor comprimento, a média do mesmo é menor (atingindo apenas o 6), enquanto para o caso de maior comprimento o valor é obviamente maior (chegando à centena).

Assim, justificando estes valores bastante mais pequenos de RE estará o facto de que ao se estar a trabalhar com resultados maiores, a razão entre os valores experimentais e real suavize, pelo que se obtêm valores menores.

Relembrando o gráfico 2, seria como se os pontos se aproximassem uns dos outros.

### *B. Variating Trials*

Finalmente, o estudo final para este contador, passa por observar o comportamento do mesmo em função do tamanho da sequência. Assim, fixou-se o número de trials em 10000, por se considerar um valor capaz de conseguir resultados satisfatórios em tempo útil.

*Accuracy e RE*

Analisando a figura 10, desde logo é possível observar que para o gráfico de comprimento = 10, com este número elevado de trials, a *accuracy* do mesmo já é bastante próxima dos 100%, o que vai ao encontro do estudado anteriormente.

Numa perspetiva mais geral, consegue-se perceber que não é apenas o número de trials que aumenta a accuracy das contagens; O próprio aumento do número de ocorrências/eventos também causa uma convergência do valor de Accuracy, de tal forma que para strings de tamanho superior a 1000 esta aproxima-se muito dos 100%.

O erro relativo, também apresenta um comportamento semelhante, no sentido em que, à medida que o comprimento da sequência aumenta, este também diminui.

Relembrando que se está a trabalhar com termos relativos, o aumento do comprimento da sequência faz então com que este valor tenda a diminuir, uma vez que o erro associado a este contador vai-se tornando menos significativo.

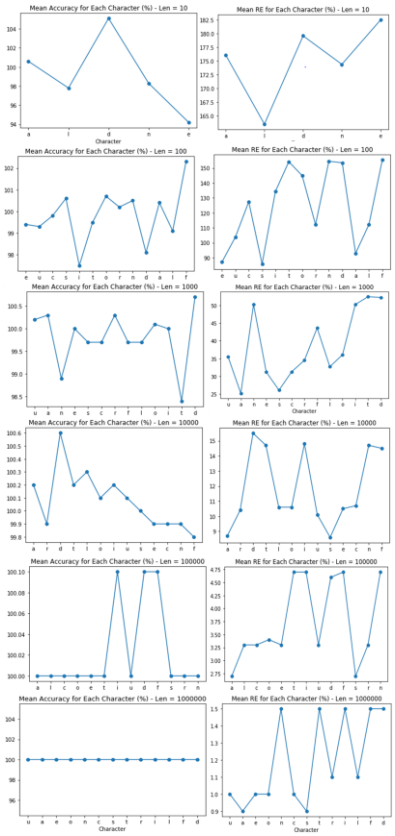

**Figura 10 - Accuracy e ER Médios em Função do Tamanho da Sequência**

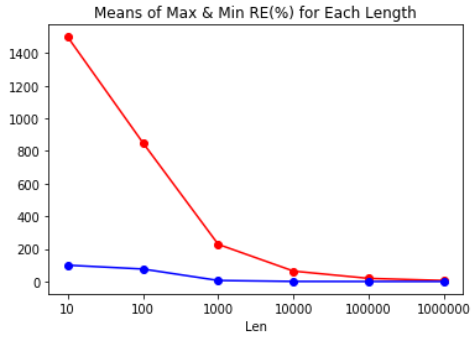


**Figura 11 - Médias de Accuracy Máxima e Mínima em Função do Tamanho da Sequência**

O gráfico 11, ilustra a variação da média do erro relativo máximo e mínimo para os vários comprimentos, e ajuda a perceber o referido anteriormente. Tendo em conta que este contador irá resultar em valores múltiplos de 16, consegue-se perceber que para valores cada vez maiores diferenças de pequenos múltiplos de 16 vão se tornando cada vez menos significativas relativamente ao valor real, produzindo esta diminuição no ER.

## IV. DECREASING COUNTER – LOG2 PROBABILITY

O segundo e último contador a ser estudado é o logarítmico em base 2 (por vezes abreviado de log2, para simplificação). Ao contrário do anterior, que era fixo, este é "dinâmico" no sentido em que a probabilidade de registar um evento diminui com o número de vezes que o mesmo já foi incrementado. Noutras palavras, à medida que o contador é incrementado, este é incrementado com menos probabilidade.

Como explicado numa das apresentações da cadeira [3]:

- If counter has value k
  - Increment it with probability $1 / 2^k$
  - Do not increment it with probability $( 1 – 1 / 2^k )$

Como o valor de *k* naturalmente aumenta, em probabilidade de incrementar diminiu com o mesmo.

- How to estimate the number of events from the counter value k ?
  - Compute $2^k – 1$

### *A. Variating Trials*

#### I. Small Sequence

*Accuracy*

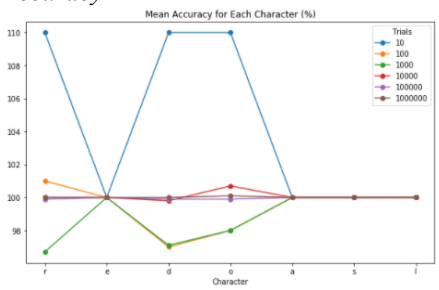


**Figura 12 - Accuracy Média para Sequencia Pequena**

Relativamente ao contador anterior, este apresenta muito melhor resultados mesmo para um número diminuto de *trials* (10).

Tendo em conta o funcionamento do mesmo, rapidamente se percebe o porquê de isto acontecer: o primeiro evento é sempre contado, e o seguinte tem uma "generosa" meia (1/2) probabilidade de ser também contado. Assim, neste contexto, onde cada letra aparece praticamente 1 vez apenas, faz sentido que a *Accuracy* seja à partida melhor, mesmo sem a suavização causada pela média ao longo dos *trials*. Daí as piores *accuracies* serem nos caracteres que aparecem 2 vezes, já que o cálculo da estimativa dá *skip* no 2, "saltando" do 1 para o 3, causando este erro.

*Relative Error*

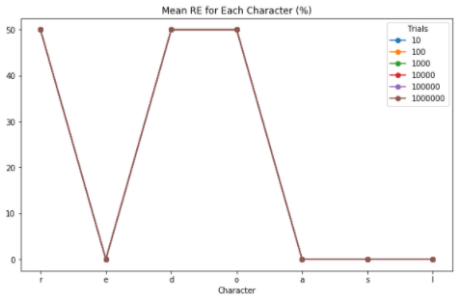


**Figura 13 - Erro Relativo Médio para Sequencia Pequena**

No que diz respeito ao erro relativo, neste é nos apresentado um gráfico bastante interessante, uma vez que para qualquer que seja o número de trials, os valores são os mesmos.

Isto parece indicar que, para este caso, o Erro Relativo não é tao influenciado pela média ao longo dos *trials*, evidenciando a precisão do contador. Cuidando quais os caracteres que apresentam maior ER, rapidamente de percebe que são os de maior frequência (no caso, 2). Ora, como já referido, um contador deste tipo - nesta base – fará com que se dê *skip* ao valor 2 durante a estimativa, passando de 1 para 3. Assim, os caracteres com 1 ocorrência não possuem erro relativo, pois como referido acima a contagem é sempre certa, e os de 2 ocorrências incorrem nesta nuance, em que são contados como 1 ou 3 aparições.

II. Big Sequence

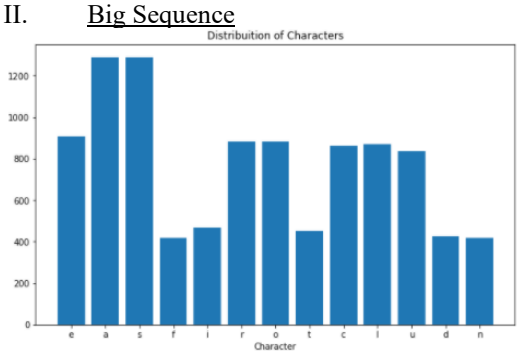


**Figura 14 - Distribuição de Caracteres para Sequencia Grande**

Mais uma vez foi gerada uma sequência de 10000 caracteres de comprimento, com a distribuição apresentada, e novamente se observa que esta vai ao encontro da distribuição das letras do nome gerador.

*Accuracy*

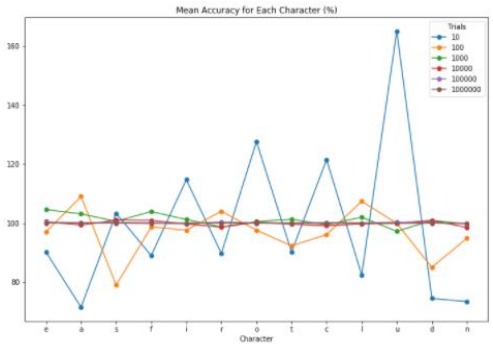


**Figura 15 – Accuracy Média para Sequencia Grande**

À semelhança dos análogo para prob16, é possível ver que com um aumento de *trials*, a accuracy acaba por convergir para 100%, devido a eventual suavização causada pela média. Contudo, um detalhe em relação ao prob16 é que este parece oscilar entre amplitudes de valores maiores (em modulo), para trials menores (10). Tal pode ser indicativo do facto de o contador não ser capaz de obter valores próximos dos reais, acabando por obter resultados (bastante) superiores ou inferiores.

*Relative Error*

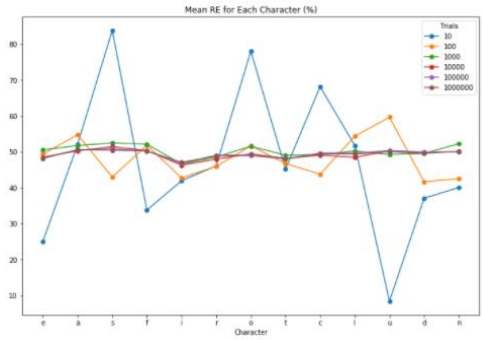


**Figura 16 - Erro Relativo Médio para Sequencia Grande**

Analisando este gráfico 16, percebe-se que o ER tende a convergir para os 50%, ao contrário do análogo anterior onde se obtinham resultados a rondar os 12%. Tendo em conta o referido anteriormente, tal parece indicar que não se estão a obter valores precisos - ou seja, estes encontram-se relativamente afastados entre si, possível consequência do facto do contador não atingir os "melhores" valores de forma consistente.

*B. Variating Trials*

Tal como no contador anterior, também se fez uma análise em função do comprimento da sequência gerada, tendo feito um estudo sobre as mesmas condições.

*Accuracy e RE*

Não surgindo como uma surpresa, observa-se na figura 17, que para qualquer tamanho da sequência gerada, a *Acc* tende a rondar os 100%, aproximando-se deste valor à medida que o tamanho da sequência aumenta. Salienta-se por isso a importância de atentar ao eixo das ordenadas, uma vez que é lá que se consegue perceber a tendência da *Accuracy*.

No que respeita o ER, ao contrário do outro contador observa-se o fenómeno oposto: enquanto que no anterior à medida que se aumentava o comprimento da sequencia, o erro relativo diminuía, tendendo cada vez mais para 0, aqui parece haver uma convergência para os 50%, sendo que apenas no caso de uma string pequena é que se obtém algum RE de 0% (tendo a justificação para este valor já sido referida anteriormente, e corresponde a uma situação negligenciável uma vez que não tem utilidade pratica um contador para um numero tão pequeno de eventos de todas as formas).

Observando o gráfico 18, percebe-se que ao contrário do análogo para prob16, aqui há uma tendência para o ER (máximo) aumentar. Tal deve-se ao facto de que a contagem não funciona de forma linear como no anterior, e devido a este crescimento exponencial, a diferença nos valores obtidos torna-se considerável o suficiente. Contudo como já vimos, em termos de valores médios a performance é muito melhor, sendo estes extremos casos

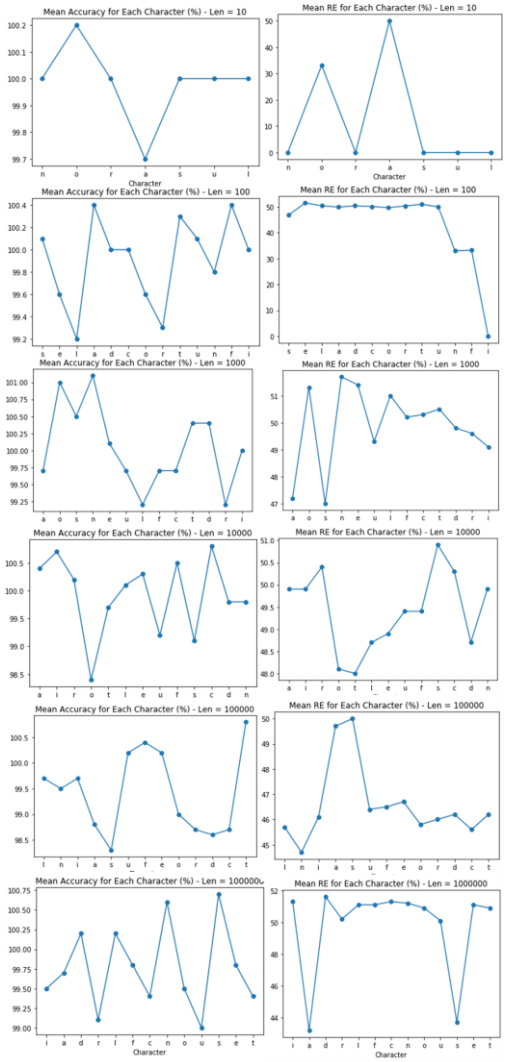


**Figura 17 - Accuracy e ER Médios em Função do Tamanho da Sequência**

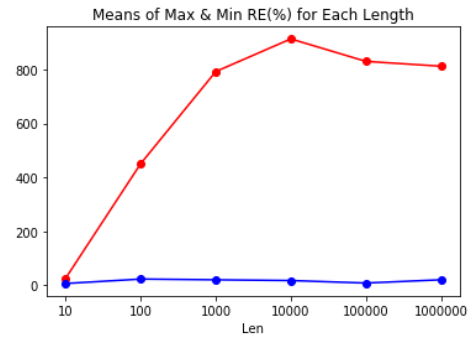


**Figura 18 - Médias de Accuracy Máxima e Mínima em Função do Tamanho da Sequência**

pontuais, que servem para indicar que não é um contador tao consistente.

## V. CONCLUSION

Uma vez que prob16 parecia ter melhores resultados (diferenciando-se pelo ER uma vez que a nível de *mean accuracy* não foi conclusivo), decidiu-se verificar os resultados para valores de sequências ainda maiores, com o objetivo de avaliar principalmente a performance do log2.

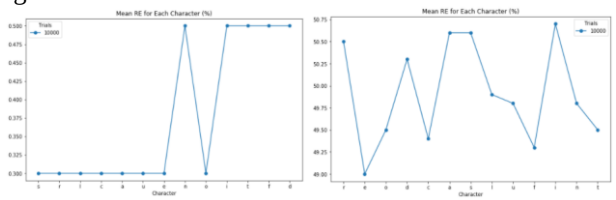


**Figura 19 - Erro Relativo Médio para Contador prob16 e log2, respetivamente, para Sequencia de Maior Tamanho**

Isto leva-nos aos 2 gráficos anteriores (figura 19), que revelam o *mean* RE para prob16 e log2 para uma sequência de 10 Milhões de caracteres.

Como se pode observar, log2 mantém a tendência de apresentar um ER maior que prob16, à volta dos 50%, enquanto que prob16 se aproxima de valores próximos de 0.

Relembrando o objetivo destes contadores, facilmente se entende que a presença de algum erro é uma inevitabilidade. À primeira vista, parece que log2 é um contador pior que prob16, pelo menos para valores maiores, já que no caso da sequência pequena é o único capaz de obter valores reais sem a influencia da média ao longo dos *trials*.

Contudo, este é um pensamento bastante errado. Primeiramente, como até já referido, estes contadores não têm utilidade pratica para sequencias (eventos) reduzidas, já que contadores exatos teriam a performance ótima sem apresentar o problema de memória. Nesta linha de pensamento, é preciso notar que log2 permite alcançar contagens muito mais elevadas que prob16, pelo que o facto de apresentar em geral maior erro relativo é um dos custos para tal vantagem. Para referência, se *n* for o número de bits, o número de eventos que se podem contar com um contador probabilístico simples de $\frac{1}{2}$ é $2^n$. Se a probabilidade do contador for de $\frac{1}{2^k}$ com k=4 para o caso de prob16, então o número de eventos que se podem contar é de $2^n \times 2^k = 2^{n+k} = 2^{n+4}$. Por outro lado, para o caso de log2, após *n* eventos, o valor em memória encontra-se em $log\left(log(n)\right)$ bits.

Dependendo do contexto em que se trabalhe, o erro que se está disposto a aceitar pode ser bastante tolerável: valores *exatos* demasiado grandes podem não ser relevantes, sendo um valor aproximado uma solução satisfatória o suficiente, pois permite perceber em geral à volta de que valores um evento se encontra.

Assim é necessário ponderar o *tradeoff* que se tem entre atingir valores maiores e obter maior erro na contagem.

Desta forma, relativamente aos 2 contadores em estudo, percebe-se que para contagens que se prevejam de grandezas enormes, o contador logarítmico se apresente como a melhor opção. Por outro lado, para valores

também elevados, mas de menor grandeza, um contador de probabilidade fixa de 1/16 seria a melhor opção.

## REFERENCES

[1] [ONLINE]. AVAILABLE: HTTPS://WWW.THOUGHTCO.COM/DEFINITION-OF-RELATIVE-ERROR-605609.

[2] [ONLINE]. AVAILABLE: HTTPS://WWW.STATISTICSHOWTO.COM/RELATIVE-ERROR.

[3] [ONLINE]. AVAILABLE: HTTPS://ELEARNING.UA.PT/PLUGINFILE.PHP/2931377/-MOD_RESOURCE/CONTENT/0/AA_09_PROBABILISTIC