受験番号	

2023 年度　一橋大学大学院ソーシャル・データサイエンス研究科（修士課程）
二次選考（筆記試験）

統計学・情報学　入試問題

2023 年 1 月実施

【 注 意 事 項 】

１.「解答はじめ」というまで開いてはいけない。試験時間：10 時 30 分～11 時 30 分。

２. 問題は本文 16 ページ、解答用紙は２枚である。下書き用紙２枚は自由に使ってよい。
　試験開始後、直ちに確認し、ページ数・枚数が異なる場合は、手を挙げなさい。

３. 試験開始後、問題の表紙、解答用紙、下書き用紙に受験番号を正確に記入しなさい。氏名は記入しないこと。
　問題、解答用紙、下書き用紙は一切持ち帰らないこと。

４. 統計学①　統計学②　情報学①　情報学②の４つの問題から２つを選んで、それぞれ別々の解答用紙に日本語または英語で解答を記入しなさい。
　解答用紙上段の「問題名」の中から、各解答用紙で解答する問題名１つを選んで、〇で囲みなさい。１枚の解答用紙で２つ以上の問題名を〇で囲んでいる場合や、問題名を〇で囲んでいない場合は得点を与えないので、以下の記入例をよく確認すること。

《記入例》 統計学① 及び 情報学① を解答する場合

解答用紙1枚目

問題名	以下の中から解答する問題名1つを選んで、〇で囲みなさい。 もう一方の解答用紙で選択した問題を選ぶことはできません。
	統計学① ／ 統計学② ／ 情報学① ／ 情報学②

解答用紙2枚目

問題名	以下の中から解答する問題名1つを選んで、〇で囲みなさい。 もう一方の解答用紙で選択した問題を選ぶことはできません。
	統計学① ／ 統計学② ／ **情報学①** ／ 情報学②

５. 解答を記入する際は、解答する問の番号（問１、問１.(1)等）を必ず記入すること。

統計学①

以下の問1–問7から4問を選び答えなさい．選んだ問を明記し，4問より多く解答しないこと．さらに，問8に答えなさい．

問 1

分布の再生性とは何か，またその例を作成せよ．

問 2

分布収束するが，確率収束しない例を作成せよ．

問 3

一致推定量，不偏推定量についてそれぞれ説明せよ．

問 4

フィッシャー情報量とクラメル・ラオの下限について説明せよ．

問 5

p値について説明せよ．

問 6

線形回帰モデルにおける変数選択の方法を 1 つ取り上げて説明せよ．

問 7

線形回帰モデルにおける多重共線性と，それにより生じる問題について説明せよ．

問 8

2次元データ$(y_1, x_1), \ldots, (y_n, x_n)$は以下の線形回帰モデルに従うとする．

$$y_i = \alpha + \beta x_i + u_i, \quad i = 1, \ldots, n.$$

ただし，説明変数x_iは非確率変数列であり，誤差項u_iは平均ゼロの何らかの確率変数列とする．回帰係数(α, β)は未知とする．

(1) 回帰係数(α, β)の最小二乗推定量$(\hat{\alpha}, \hat{\beta})$を導出せよ．
(2) 誤差項u_iに適当な仮定を置いたうえで，$\mathrm{Var}(\hat{\beta})$を導出し，その推定量を構成せよ．
(3) 仮説の組

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0$$

についての検定問題を考える．誤差項u_iに適当な仮定を置いたうえで，検定統計量を適切に構成し，有意水準αで検定する手順を説明せよ．

統計学②

以下の問1–問4から2問を選び答えなさい．選んだ問を明記し，2問より多く解答しないこと．

問 1

標本空間Ωの事象$A, A_1, A_2, \ldots$について，次の公理1 – 3を満たす関数$\mathbb{P} : \Omega \to \mathbb{R}$を確率という：

公理1. $\mathbb{P}(A) \geq 0$;　　公理2. $\mathbb{P}(\Omega) = 1$;

公理3. 事象$A_1, A_2, \ldots$が互いに排反であるとき，$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

以下の(1)–(3)に答えよ．

(1) 事象$A_1, A_2, \ldots$が互いに排反であるとはどういうことか，その定義を述べよ．

(2) 確率の公理1 – 3を用いて次のa) – c)を示せ．

a) $\mathbb{P}(\emptyset) = 0$;

b) 事象A, Bが$B \subset A$を満たすとき$\mathbb{P}(B) \leq \mathbb{P}(A)$;

c) ブールの不等式：任意の自然数nについて

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(A_i\right).$$

(3) ある帰無仮説の列$H_{0,1}, \ldots, H_{0,m}$に対応するp値をそれぞれ$p_1, \ldots, p_m$とする．仮説の組

$$H_0 : H_{0,1} \cap \cdots \cap H_{0,m} \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

に対する有意水準αの検定を構成せよ．

問 2

2次元データ$(y_1, x_1), \ldots, (y_n, x_n)$は以下の潜在変数モデルに従うとする．

$$y_i^* = x_i\beta + u_i, \quad y_i = \begin{cases} 1 & \text{if} \quad y_i^* \geq 0, \\ 0 & \text{otherwise}, \end{cases}$$

ただし，説明変数x_iは非確率変数列，係数βは未知のパラメータであり，誤差項u_iはロジスティック分布

$$\mathbb{P}(u_i \leq t) = \Lambda(t) := \frac{\exp(t)}{1 + \exp(t)}$$

からの独立標本とする．以下の(1)–(4)に答えよ．

(1) 対数尤度関数とスコア関数を導出したうえで，βの最尤推定量$\hat{\beta}$を定義せよ．

(2) フィッシャー情報量を導出したうえで，$\hat{\beta}$の漸近分布を答えよ．（漸近分布の導出過程は必要ない．）

(3) 仮説の組

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0$$

に対する有意水準αの検定を構成せよ．

(4) $\beta = 1$のときの検出力について議論せよ．

問 3

線形回帰モデルにおける予測誤差評価を考える．まず，i番目のデータを抜いた上で推定した$\boldsymbol{\beta}$の最小二乗推定量$\hat{\boldsymbol{\beta}}_{[i]}$を用いて，予測誤差$(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{[i]})^2$を評価する．ただし，$\boldsymbol{x}_i$は$(p+1)\times 1$の説明変数の列ベクトルとする．これを$i = 1, \ldots, n$に対して繰り返し行い，以下の平均予測誤差を計算する．

$$\mathrm{MPE}_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{[i]})^2$$

これは，以下の簡単な形に書き直せることが知られている．

$$\mathrm{MPE}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}{1 - \boldsymbol{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}\boldsymbol{x}_i}\right)^2$$

ただし，$\mathbf{X}$は$n \times (p+1)$の説明変数のフル・ランク行列とする．

(1) MPE_2の表現に書き直せることにより，平均予測誤差の計算にどのようなメリットがあるか議論せよ．

(2) $\mathbf{A} \in \mathbb{R}^{n\times n}$は$n \times n$の正則行列，$\boldsymbol{u}$，$\boldsymbol{v} \in \mathbb{R}^{n\times 1}$をそれぞれ$n \times 1$の列ベクトルとする．この時，Sherman-Morrison-Woodburyの公式は以下のように与えられる．

$$(\mathbf{A} - \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top \mathbf{A}^{-1}}{1 - \boldsymbol{v}^\top \mathbf{A}^{-1}\boldsymbol{u}}$$

ただし，$1 - \boldsymbol{v}^\top \mathbf{A}^{-1}\boldsymbol{u} \neq 0$とする．左辺を$\mathbf{W}$，右辺を$\mathbf{V}$と定義し，$\mathbf{W}^{-1}\mathbf{V} = \mathbf{I}_n$，$\mathbf{V}\mathbf{W}^{-1} = \mathbf{I}_n$を示せ．ただし，$\mathbf{I}_n$は$n$次元の単位行列とする．

(3) $\mathbf{X}_{[i]}$，$\mathbf{Y}_{[i]}$は，それぞれ$\mathbf{X}$，$\mathbf{Y}$のi行目を削除した行列とする．上記公式を用いて，以下を示せ．

$$(\mathbf{X}_{[i]}^\top \mathbf{X}_{[i]})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - \boldsymbol{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}\boldsymbol{x}_i}$$

また，$\hat{\boldsymbol{\beta}}_{[i]}$を$\hat{\boldsymbol{\beta}}$を用いて導出せよ．導出の過程も明記すること．

(4) $\mathrm{MPE}_1 = \mathrm{MPE}_2$を示せ．過程も明記すること．

問 4

(1) 確率変数X，Vを以下のように定義する．

$$X|V = v \sim N(0, v), \quad V \sim \mathrm{Exp}\left(\frac{1}{2a^2}\right)$$

ただしVの密度関数は，aを正の定数として

$$f_V(v) = \frac{1}{2a^2}\exp\left(-\frac{v}{2a^2}\right) \quad \text{for} \quad v > 0$$

で与えられる． この時，ラプラス分布は正規分布と指数分布のmixtureで表現できることを示せ．つまり，以下の等式が成立することを示せ．

$$\frac{1}{2a}\exp\left(-\frac{|x|}{a}\right) = \int_0^\infty f_{X|V=v}(x) f_V(v) dv$$

ただし，必要に応じて以下の逆ガウス分布(平均パラメータμ, 形状パラメータλ)を利用せよ．

$$f(v; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi v^3}}\exp\left(-\frac{\lambda(v-\mu)^2}{2\mu^2 v}\right) \quad \text{for} \quad v > 0$$

(2) 以下の線形回帰モデルに対して，$\boldsymbol{\beta}$の事前分布にラプラス分布を考える．つまり，

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$
$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{1}{2a\sqrt{\sigma^2}}\exp\left(-\frac{|\beta_j|}{a\sqrt{\sigma^2}}\right)$$

とする．(1)の表現に基づき，βの事後分布が解析的に導出できるように$p(\boldsymbol{\beta}|\sigma^2)$を変形せよ．必要に応じて，自分でパラメータを定義すること．

(3) (2)で得られた表現を用いて，β_jの事後分布を解析的に導出せよ．

(4) $p(\boldsymbol{\beta}|\sigma^2)$にガウス分布を適用する場合に比べて，ラプラス分布を適用することでどのような違いが生じうるか議論せよ．縮小推定との関連も議論せよ．

情報学①

C言語によるデータ構造を用いたグラフ探索のプログラムについて，以下の問いに答えよ．なお，各問において，プログラムのソースコードを記述する際には，処理内容が判別可能な擬似コードであれば，厳密なC言語に従わなくても可とする．

問 1

スタック，キューと呼ばれるデータ構造の両方がどのようなデータ構造であるかを合わせて5行程度で説明せよ．

問 2

以下のプログラム 1のようにC言語でスタックを実装し，以下のように関数を定義した．

- `push()`:スタックへの要素の追加
- `pop()`:スタックの先頭要素の削除
- `top()`:スタックの先頭要素へのアクセス

また，`MAX_SIZE`は定数でスタックの最大のサイズを表すマクロである．`push()`の関数定義にある空欄 ① に入るプログラムを1–3行程度で，`pop()`の関数定義にある空欄 ② に入るプログラムを2–6行程度で書け．

プログラム 1　C言語によるスタックの実装例

```
1  #define MAX_SIZE 32
2
3  struct stack {
4      int top;
5      int data[MAX_SIZE];
6  };
7
8  void init(struct stack *stk) {
9      stk->top = 0;
10 }
11
12 void push(struct stack *stk, int item) {
13     if (stk->top >= MAX_SIZE) {
14         printf("stack is full\n");
15     } else {
16             ①
17     }
18 }
19
20 void pop(struct stack *stk) {
21         ②
22 }
23
24 int top(struct stack *stk) {
25     int item;
26     if (stk->top == 0) {
27         return -1;
28     } else {
29         item = stk->data[stk->top];
30         return item;
31     }
32 }
```

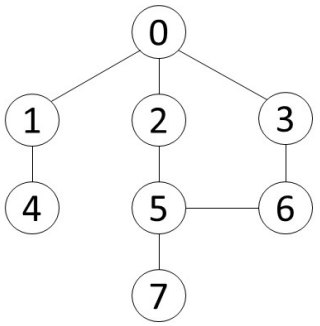
問 3

プログラム 2のような関数の再帰呼び出しによるグラフ探索のプログラムを作成した．adjは探索を行うグラフの隣接行列 ($N \times N$の要素を持つ, 説明は下記) で，FALSEで初期化された要素数Nの配列visitedは，各頂点訪問済みであるかを記録するものとする．また，グラフの頂点には0から$N-1$までの非負の整数で番号が振られており，始点は0である．ここで書かれているアルゴリズムと同様の手続きで，図 1の0番の頂点を始点としてグラフを探索する場合の可能な頂点の訪問順序の例をひとつ書け．

隣接行列: グラフを表す正方行列で，各要素は対応する頂点間の辺の有無をあらわす．隣接行列adjのi行j列の要素adj[i][j]は，頂点i, j間に辺が存在するとき1，辺が存在しないとき0である．プログラム 2では隣接行列は2次元配列で表されている (例は，図 1を参照).

プログラム 2　再帰を用いたグラフ探索

```
1  void recsearch(int adj[N][N], int visited[N], int start) {
2      visited[start] = TRUE;
3      for (int dst = 0; dst < N; ++dst) {
4          if (adj[start][dst] == 1) {
5              if (visited[dst] == FALSE) {
6                  visited[dst] = TRUE;
7                  recsearch(adj, visited, dst);
8              }
9          }
10     }
11 }
```



左のグラフをあらわす隣接行列

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

図 1　グラフと隣接行列

問 4

問2に示したスタックを用いたプログラム 3は，上記のプログラム 2と同様にグラフの全ての頂点を訪問することができる．プログラム中の空欄 ③ に入るプログラムを1–3行で，④ に入るプログラムを2–5行程度で書け．

プログラム 3 スタックを用いたグラフ探索

```
1  void dfsearch(int adj[N][N]) {
2      struct stack S;
3      init(&S);
4
5      int visited[N];
6      for (int i = 0; i < N; ++i) {
7          visited[i] = FALSE;
8      }
9
10     visited[0] = TRUE;
11     push(&S, 0);
12     while (top(&S) != -1) {
13         ③
14         for (int dst = 0; dst < N; ++dst) {
15             ④
16         }
17     }
18 }
```

問 5

問3ならびに問4に関して，再帰を用いたアルゴリズムよりもスタックを用いたアルゴリズムが計算時間とメモリ使用量の両面で効率的であることが多い．その理由について200文字以内で論ぜよ．

情報学② − A

注意: 「情報学②」を選択する場合には，**A, Bのいずれか一方のみ**について解答すること．

今，互いに区別可能なN枚のカードをランダムにシャッフルするプログラムを作成したい．このようなシャッフルの方法にはFisher–Yatesのアルゴリズムが知られており，以下のステップでシャッフル後のカードの順序を決定する．なお，以下の説明では，シャッフル前のカードの束を$\mathcal{I}$，シャッフル後のカードの束を$\mathcal{O}$と表す．

Fisher–Yatesのアルゴリズム

- $\mathcal{I}$のN枚のカードからランダムに1枚を選び，それを$\mathcal{O}$の1番目のカードとする．
- $\mathcal{I}$の残りの$N-1$枚のカードからランダムに1枚をえらび，それを$\mathcal{O}$の2番目のカードとする．
- 以下，i番目の操作 $(i = 1, 2, \ldots, N)$においては，$\mathcal{I}$の残り$N-i+1$のカードからランダムに1枚のカードを選び，それを$\mathcal{O}$のi番目のカードとする操作を$i = N$となるまで繰り返す．

このような，カードのランダムシャッフルにおいて，「完全なランダムシャッフル」は，$\mathcal{I}$の任意のカードについて，そのカードが$\mathcal{O}$においてi番目のカードとなる確率が，全てのiについて$\frac{1}{N}$となることを指す．これを踏まえて以下の問いに答えよ．

問 1

$\mathcal{I}$のN枚のカードのうちの1枚に注目したとき，そのカードが$\mathcal{O}$の1番目のカードとなる確率はいくらか．同様に2番目のカードになる確率はいくらか．

問 2

Fisher–Yatesのアルゴリズムによって，$\mathcal{I}$をシャッフルして$\mathcal{O}$を得たとき，このシャッフルが「完全なランダムシャッフル」となることを示せ．

問 3

今，$\mathcal{I}$に含まれていたN枚のカードを完全にランダムにシャッフルすることができたが，ジョーカー1枚を入れてシャッフルするのを忘れてしまった．この場合，もう一度全てのカードをシャッフルすることなく，できるだけ少ない操作でジョーカーを含む「完全なランダムシャッフル」を得るにはどうすれば良いか．これに関して，以下の問に答えよ．なお，ジョーカー挿入後の$N+1$枚のカードの束を以下では$\mathcal{O}'$と表わす．

(1) ジョーカー1枚を$\mathcal{O}$のN枚のカードの先頭，末尾，あるいはカードの間の$N+1$箇所のいずれに挿入する操作を考える．この時，ジョーカーが$\mathcal{O}'$の中でi番目$(i=1,2,\ldots,N+1)$に来る確率はいくらか．

(2) ジョーカーが挿入された後に，$\mathcal{O}$のj番目$(j=1,2,\ldots,N)$のカードが何番目のカードになるかを考える．これは，ジョーカーがj番目のカードより前に挿入されて$j+1$番目になるか，後に挿入されてj番目のまま変わらないかのいずれかとなるが，それぞれが起きる確率はいくらか．

(3) 上記(1), (2)を踏まえて，(1)で行ったような，ジョーカーをランダムな位置に挿入する操作を行うと，$\mathcal{O}'$が完全なランダムシャッフルとなることを示せ．

問 4

今，「完全にランダムシャッフル」された二つのカードの束があり，それぞれはこれまで同様N枚のカードから成る．これらを$\mathcal{O}_1$ならびに$\mathcal{O}_2$を表わす．$\mathcal{O}_1$，$\mathcal{O}_2$を上手く混ぜ合わせて，$2N$枚の「完全にランダムシャッフル」されたカードの束$\mathcal{O}''$を得たい．このとき，$2N$枚のカードを再度シャッフルし直すことなく，完全なランダムシャッフル結果を得るにはどうすれば良いか．ただし，二つのカードの束は絵柄が異なっており，互いの束に含まれるカードは相互に区別可能であるものとする．

特に，計算機において上記のようなランダムシャッフルを行う際には，乱数を生成するのに比較的大きな計算コストが生じるため，乱数生成の回数が最小となるような方法がどのような方法であるかを述べ，その方法で完全なランダムシャッフルを実現できることを示せ．

情報学② − B

注意: 「情報学②」を選択する場合には，**A, Bのいずれか一方のみ**について解答すること．

S学部のあるゼミでは社会課題に対して情報科学の技術で解決する方法について議論を行っている．本日の議題は「ディープ・フェイクニュース (深層学習技術によって自動生成された誤った情報) の抑制」である．

そこで，ゼミ生の α さんは提案を行った．「ディープ・フェイクニュースであるための条件は [①] ですよね．であれば，まずはプログラムで自動生成された文か人手で作成された文か見分ける方法について議論してはどうでしょうか?」その意見はゼミ生から賛同を得た．

続けて同じくゼミ生の β さんが発言した．「文を単語列として見たときに，人間にとって不自然な並びの単語列になっていればプログラムで自動生成された文だと判定できないでしょうか?」その意見に先生は満足した顔持ちでこう言った．**「では，文の自然さを評価する方法をみんなで考えてみようか」(発言②)**

問 1

①に該当する項目を下記の選択肢 (a)–(d) から選択せよ．

(a) プログラムで自動生成された文
(b) 誤った情報
(c) プログラムで自動生成された文，かつ，誤った情報
(d) プログラムで自動生成された文，もしくは，誤った情報

発言②に関して，文の自然さを評価する方法の一案として下記が検討された．文を単語分割して取り出された単語列 $S = (t_1, t_2, \cdots, t_N)$ において，Sの自然さスコア$F_{\mathrm{nat}}(S)$を，文中で連続するに単語の並びの自然さを表す事後分布の積で以下のように表す．

$$F_{\mathrm{nat}}(S) = \prod_{i=1}^{N} p(t_i|t_{i-1}) \tag{1}$$

ただし，先頭から i 番目の単語 t_i の自然さ値は直前の単語 t_{i-1} の後に出現する条件付き確率$p(t_i|t_{i-1})$で表すこととする．なお，t_0 は文頭 (BOS = beginning of sentence)とし，先頭の単語 t_1 の自然さ値は， t_1 が文頭に出現する確率 $p(t_1|t_0)$ として算出されることとする．

例えば，文S_{sample}: “John loves potatoes” の自然さスコア $F_{\mathrm{nat}}(S_{sample})$を考えてみる．文を単語列に分割すると， “John”, “loves”, “potatoes” が得られるから，$F_{\mathrm{nat}}(S_{\mathrm{sample}})$ は，

$$F_{\mathrm{nat}}(S_{\mathrm{sample}}) = p(\text{“John”}|\mathrm{BOS}) \cdot p(\text{“loves”}|\text{“John”}) \cdot p(\text{“potatoes”}|\text{“loves”}) \tag{2}$$

で算出される．表 1 の条件付き確率を用いて算出した結果，

$$\begin{aligned} F_{\mathrm{nat}}(S_{\mathrm{sample}}) &= 0.25 \cdot 0.20 \cdot 0.20 \\ &= 0.01 \end{aligned} \tag{3}$$

が得られた．

表 1　$p(t_i|t_{i-1})$ の条件付き確率 (例文)

t_i \ t_{i-1}	BOS	“John”	“loves”	“potatoes”
“John”	0.25	0.00	0.10	0.00
“loves”	0.07	0.20	0.01	0.00
“potatoes”	0.15	0.01	0.20	0.01

問 2

ここまでの説明を踏まえ，表 2に示した条件付き確率の値を用いて，下記の文 S_1から文 S_5 のうち最も自然さスコアの高い文とその値を答えよ.

- 文S_1の単語列: “A”, “B”, “C”
- 文S_2の単語列: “D”, “E”
- 文S_3の単語列: “A”, “B”, “A”, “B”
- 文S_4の単語列: “C”, “B”, “A”
- 文S_5の単語列: “C”, “A”, “D”, “B”, “E”, “E”

表 2　$p(t_i|t_{i-1})$ の条件付き確率 (問2)

t_i \ t_{i-1}	BOS	“A”	“B”	“C”	“D”	“E”
“A”	0.07	0.00	0.04	0.00	0.05	0.01
“B”	0.01	0.04	0.00	0.01	0.12	0.03
“C”	0.04	0.00	0.05	0.00	0.07	0.07
“D”	0.01	0.02	0.07	0.03	0.00	0.05
“E”	0.01	0.05	0.01	0.07	0.07	0.00

問 3

式 (1) を用いて文の自然さスコアを計算したところ，文の長さ (文に含まれる単語数) が大きくなるほどスコアが小さくなり，自然な長文よりも不自然な短文の方がスコアが高くなるという現象が観測された．この課題に関し，文の長さの影響を受けない自然さスコアの計算式を示し，その計算式が文の長さの影響を受けない理由を述べよ． ただし，計算式は下記の 2 つの条件を満たすこと．

- 文の全単語に関する条件付き確率 $p(t_i|t_{i-1})$ $(i = 1, 2, \cdots, N)$を用いる
- 任意のiについて $p(t_i|t_{i-1}) > 0$であるとき，十分大きなNに対して$F_{\mathrm{nat}} > 0$を満たす

問 4

ディープ・フェイクニュースの特徴を，人手で作成するフェイクニュースと比較しながら，質と量の観点から 300 文字以内で論ぜよ．