

数字・英字・かな文字を用いた暗証番号における安全性評価

リスク工学グループ演習1班
石田若菜 岡宏樹 品川和雅 野貴泰
アドバイザー教員 金山直樹

1 背景

現在、本人だけがサービスを受けられるようにするための認証技術の1つとして、暗証番号が広く用いられている。その中でも4桁の暗証番号は、利便性の観点から広く普及しており、金融口座やクレジットカードの暗証番号、携帯電話のロック解除など、さまざまな場面で活用されている。サービス利用時に暗証番号の入力を求めることによって、暗証番号を知らない攻撃者（不正な利用者）は、十分小さな確率でしかサービスを利用することが出来ない。しかし、実際に用いられている暗証番号は、誕生日の日付などを元に設定するケースが多いことが報告されている [1]。暗証番号として日付が用いられている場合、その番号を推測することは1/10000の確率よりも高い確率で可能であり、暗証番号のこのような利用の安全性については問題視されている [2]。本稿では、暗証番号の安全性とランダム性の評価を行い、暗証番号を用いるシステムの利用と設計に関して有益な知見を得ることを目指す。

2 目的

本稿では、「暗証番号の安全性が高い」ことを、「攻撃者が暗証番号を推測できる確率が低い」ことであるとして論じる。

暗証番号の安全性を向上させる方法の一つとして、使用できる番号の総数を増やす方法がある。しかしながら、番号の総数を2倍に増やしたからといって、安全性も2倍に向上するとは言えない。例えば、使用できる番号の総数を10000通りから20000通りに増やしたとしても、利用者が選ぶ番号の分布が偏っている場合、暗証番号を推測できる確率が単に1/2倍されるとは限らないからである。本研究では、使用できる番号の総数を増やすために、暗証番号に使用できる文字の種類を変えたときの安全性について、定性的・定量的な評価を行うことを目的とする。具体的には、数字（10種類）、英字（26種類）、英字+数字（36種類）、かな文字（46種類）を用いた暗証番号における安全性の評価を行う。

3 評価方法・実験方法

3.1 評価方法

ここでは、4桁の暗証番号について、ランダム性の高さを情報量比によって評価し、安全性の高さを Guess Work によって評価する。また、各暗証番号方式の特徴については、出現回数を元に考察を行う。

ランダム性が高いほど、その確率分布から得られる情報量は大きいため、情報量を用いてランダム性を計ることは妥当な尺度であると考えられる。最も情報量が大きくなるのは、確率分布が一様分布のときであるが、その場合は N 個のデータに対して情報量は $\log_2 N$ となる。ここで、一様分布に近い確率分布をランダム性の高い分布であると評価する方が妥当であると考えられる。そこで、単純に情報量を比較するのではなく、一様分布の情報量との比（情報量比）を用いて、ランダム性を評価する。情報量比は次のように定義される。

定義 3.1 (情報量比) N 通りの暗証番号が存在する認証方式について、暗証番号を出現頻度の高い順に並べたものを x_1, x_2, \dots, x_N とする。すなわち、 x_i の出現頻度を p_i とすると、 $p_1 \geq p_2 \geq \dots \geq p_N$ である。このとき、以下の測定量を情報量比と定義する。

$$H = -\frac{1}{\log_2 N} \cdot \sum_{i=1}^N p_i \cdot \log_2 p_i$$

Guess Work は、「攻撃者が暗証番号を推測するために必要な試行回数の期待値」を意味し、既存研究でも安全性の尺度として用いられている [3]。最も安全性の低い暗証番号は Guess Work の値が1になり、暗証番号の安全性が高くなるほど Guess Work の値は大きくなる。Guess Work は以下のように定義される。

定義 3.2 (Guess Work (1)) 定義 3.1 と同様に、暗証番号 x_i の出現頻度を p_i とし、全部で N 個の暗証番号が存在するものとする。このとき、以下の測定量を Guess Work と定義する。

$$G_1 = \sum_{i=1}^N p_i \cdot i$$

上記が一般的な Guess Work の定義であるが、本稿ではこれを変形した次の定義を用いる。

定義 3.3 (Guess Work (2)) L 桁の暗証番号認証方式について、各桁の *Guess Work (1)* が g_1, g_2, \dots, g_L であるとき、*Guess Work (2)* を次のように定義する。

$$G_2 = \prod_{i=1}^L g_i$$

第 i 桁の *Guess Work (1)* とは、攻撃者が第 i 桁を推測するときに必要な試行回数の期待値である。したがって、 L 桁の暗証番号を推測するときの試行回数の期待値は、各桁の *Guess Work (1)* の総積であると考えられるため、*Guess Work (2)* は、*Guess Work (1)* の自然な変形である。

本稿で *Guess Work (2)* を用いる理由は、最大で $N = 46^4$ 通りの暗証番号について解析するのに対し、サンプル数が 212 個であるため、*Guess Work (1)* を用いたときと比べて意味のある結果が得られやすいと判断したためである。

3.2 実験方法

数字 (10 種類)、英字 (26 種類)、英字+数字 (36 種類)、かな文字 (46 種類) を用いた暗証番号の実態を知ることが目的として、アンケート調査を実施した。調査方法は、筑波大学学群生を対象にアンケート用紙を授業前に配布し、10 分程度でその場で回答するアンケート形式である。回答者数は、218 名であった。ここで調査対象文字、調査項目、調査結果の整理を説明する。調査対象文字は、4 桁数字 (0~9)、4 文字のアルファベット (a~z)、4 文字のひらがな (あ~ん)、アルファベットと数字 (a~z, 0~9) を組み合わせた 4 文字の 4 種類とする。なお、ひらがなは濁点・半濁点 (例: が、ず、ば) や小さい文字 (例: つ、や、よ) 以外の文字とする。調査項目は、第一に覚えておくことができ、なるべく他人に推測されない 4 桁 (以下、暗証番号的な 4 桁と表記)、第二にでたらめな 4 桁 (以下、でたらめな 4 桁と表記) である。なお、本研究では、暗証番号的な 4 桁の結果を主に考え、でたらめな 4 桁の結果は適宜必要に応じて使うものとする。最後に、「覚えておくことができ、なるべく他人に推測されないものを回答するとき、何を意識したか」を 4 種類の文字ごとに質問する。調査項目は以下の 8 設問および簡単なアンケートであった。

設問 1 覚えておける 4 桁の数字で、なるべく他人に推測されないものを書いてください。

設問 2 覚えておける 4 桁の英字で、なるべく他人に推測されないものを書いてください。

設問 3 覚えておける 4 桁のかな文字で、なるべく他人に推測されないものを書いてください。

設問 4 覚えておける 4 桁の英数字で、なるべく他人に推測されないものを書いてください。

設問 5 でたらめな 4 桁の数字を書いてください。

設問 6 でたらめな 4 桁の英字を書いてください。

設問 7 でたらめな 4 桁のかな文字を書いてください。

設問 8 でたらめな 4 桁の英数字を書いてください。

アンケートは、設問 1~4 の回答において、どのような意図で選んだかを質問するものである。調査結果は、有効な回答を選定するため、判別不明な文字が含まれていたものは除外した。その結果、有効回答数 (有効回答率) は、212 名 (97.2%) であった。

4 結果・考察

4.1 節では、情報量比と *Guess Work* を用いた、ランダム性と安全性に関する結果についての考察を行う。4.2 節以降では、数字・英字・かな文字・英数字のそれぞれの設問から得られた出現回数の特徴について、考察を行う。

4.1 ランダム性と安全性の結果

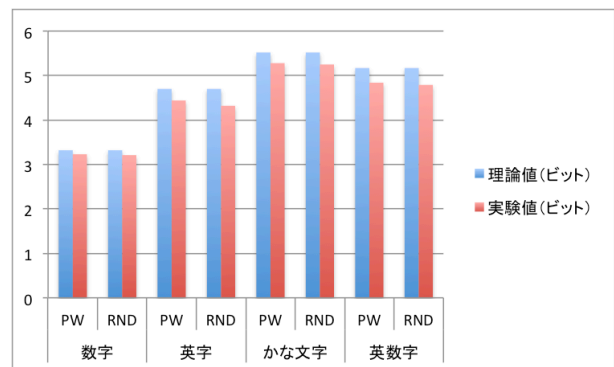


図 1: 情報量の比較

情報量比の比較 数字・英字・かな文字・英数字の文字数はそれぞれ $N = 10, 26, 46, 36$ であるため、暗証番号が一様分布に従って選ばれている場合、各桁の情報量は $H = \log_2 N = 3.32, 4.70, 5.52, 5.17$ ビットである。表 1 の上部および図 1 に、実験によって得られた情報量 (情報量比) を掲載する。図 1 において、青線は理論値、赤線は実験値を表し、PW および RND はそれぞれ暗証番号的な 4 桁とでたらめな 4 桁を表す。この実験データから以下のような事実を読み取ることが出来る。

1. 暗証番号的な 4 桁において、情報量比は、数字、かな文字、英字、英数字の順に高い。
2. すべての実験において、でたらめな 4 桁よりも、暗証番号的な 4 桁の方が情報量比が大きい (ランダム性が高い)。

表 1: 情報量および Guess Work の比較表

	数字		英字		かな文字		英数字	
	PW	RND	PW	RND	PW	RND	PW	RND
○ 情報量								
理論値 (最大値) [ビット]	3.32	3.32	4.70	4.70	5.52	5.52	5.17	5.17
実験値 [ビット]	3.23	3.21	4.44	4.37	5.28	5.25	4.84	4.79
情報量比	0.973	0.967	0.945	0.919	0.957	0.951	0.936	0.926
○ Guess Work								
Guess Work (回)	414	412	5043	3790	38307	36943	11449	11094

(1)の実験結果から、かな文字は文字数およびランダム性の両方の観点から、英字・英数字よりも勝っていることが分かる。一方、数字は最もランダム性が高いが、文字数が少ないため、暗証番号の安全性はかな文字に劣ると考えられる。

(2)の実験結果は我々も全く予想していなかった結果である。「でたらめな4桁」と質問するよりも、「覚えておいて推測されにくい4桁」と質問した方がよりランダムな4桁を得られるという現象は非常に興味深い。このような現象が起こる原因の解明は今後の課題とする。

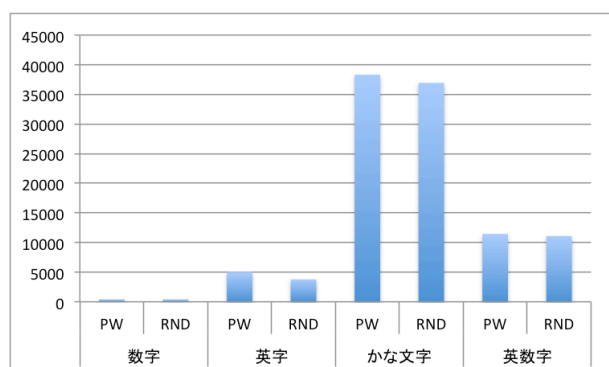


図 2: Guess Work の比較

Guess Work の比較 表 1 の下部に実験によって得られた Guess Work (定義 3.3) を掲載する。暗証番号的な 4 桁において、かな文字と数字の Guess Work を比較すると、90 倍以上の差がある。すなわち、攻撃者が暗証番号を推測する際、必要な試行回数は、数字 4 桁の場合よりもかな文字の場合の方が 90 倍以上も多いことが期待される。よって、かな文字 4 桁の方が数字 4 桁よりも安全性が 90 倍以上高いと考えられる。

4.2 数字の特徴

図 3 と図 4 に、それぞれ設問 1 (暗証番号的な 4 桁) と設問 5 (でたらめな 4 桁) における各数字の出現回数を掲載する。

- 第一に設問 1 は 0 から 2 の発生が 4 桁中最も多く、1 桁目から順に約 58%, 41%, 49%, 36% となっている。一方、設問 5 は 1 から 3 の発生が 4 桁中最も多く、1 桁目から順に約 50%, 47%, 43%, 42% となっている。特に設問 1 は 0 の発生が多く、設問 3 は 0 が少なく 3 が多い。
- 第二に 4 桁のうち、1 桁目と 2 桁目、3 桁目と 4 桁目の数字の 2 桁の組み合わせが日単位 (1 ~ 31) の範囲内である個数は、設問 1 では 1 桁目と 2 桁目のペアが 124 個 (占有率 58%)、3 桁目と 4 桁目のペアが 115 個 (占有率 54%) であり、どちらも半数を超えていることがわかる。一方、設問 5 では 1 桁目と 2 桁目のペアが 77 個 (占有率 36%)、3 桁目と 4 桁目のペアが 74 個 (占有率 34%) であり、設問 1 に比べて少ない。

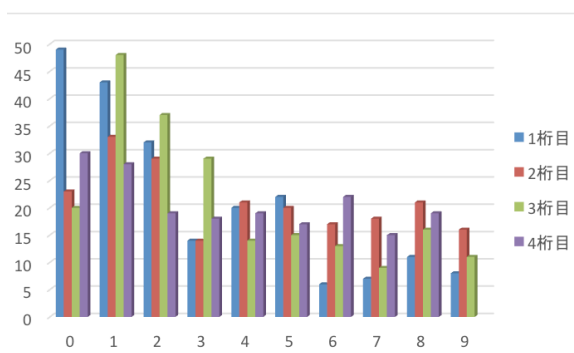


図 3: 暗証番号的な数字 4 桁 (設問 1) の出現回数

以上から上記の第一では設問 1 の「覚えておくことができ、なるべく他人に推測されないもの」と質問されると被験者は、0~9 までの数字のうち、0 に近いものを「覚えやすい数」として認識する傾向があると考えられる。一方、設問 5 の「でたらめなもの」では、0 以外の小さい数字を選ぶ傾向がある。暗証番号的な 4 桁またはでたらめな 4 桁、どちらを認知しても小さい数字を選ぶ傾向があると考えられる。上記の第二では、設問 1 の「覚えておくことができ、なるべく他人に推測されないもの」と質問されると被験者は、日単位の個数が多い結果から、誕生日などの日付を選択する

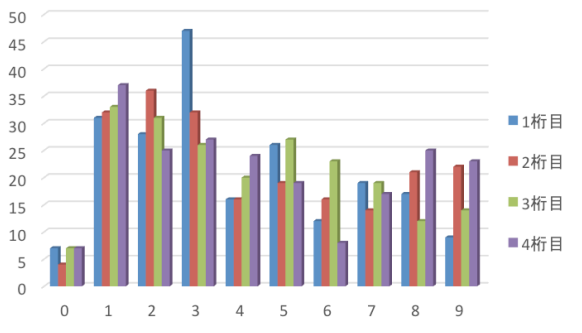


図 4: でたらめな数字 4 桁 (設問 5) の出現回数

傾向が強いことが考えられる。一方、設問 5 の「でたらめなもの」では、日単位の個数が設問 5 より約 20% 少ないことから、でたらめと認知すると誕生日など日付は使わない傾向があると考えられる。暗証番号の作成にあつては、被験者は小さい数字を選ぶ傾向があることから、注意が必要である。また、暗証番号に設定されやすい誕生日などの日付は、「でたらめ」を認識することで偏りを小さくできると考える。

4.3 英字の特徴

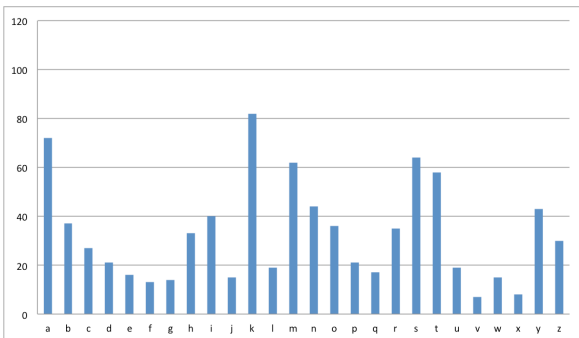


図 5: 暗証番号的な英字 4 桁 (設問 2) の出現回数

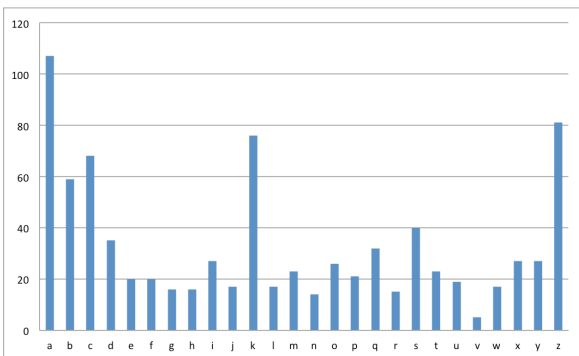


図 6: でたらめな英字 4 桁 (設問 6) の出現回数

図 5 と図 6 にそれぞれ設問 2 (暗証番号的な 4 桁) と設問 6 (でたらめな 4 桁) における各英字の出現回数を掲載する。実験結果から、以下のことが分かる。

- 暗証番号的な 4 桁において、k、a、s、m、t の順で出現回数が多いことがわかる。ここで、文献 [3] において、日本語会話における単音節の出現頻度においても、k、a、s、m、t の出現頻度は高いことが明らかになっている。したがって、暗証番号的な 4 桁は、日常会話で用いられている単語などが使用されやすい可能性がある。
- でたらめな 4 桁において、a、z、k、c、b の順で出現回数が多いことがわかる。このような結果になった理由として、アルファベットの最初と最後の文字は思いつきやすく、でたらめな 4 桁に設定する人が多かったためだと考えられる。k については、設問 2 において出現回数が最も多いため、設問 6 においても出現回数が多くなったと考えられる。

以上より、設問 2 と設問 6 において、それぞれの文字の出現回数の分布が異なる結果となった。暗証番号的な 4 桁を設定する際、覚えやすくするため、日常会話で用いられているような、親しみのある単語を使用する傾向があると考えられる。

4.4 かな文字の特徴

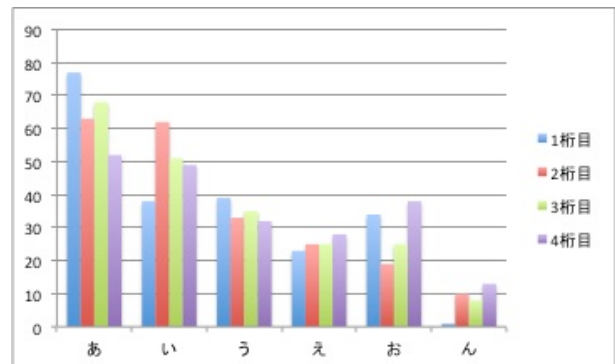


図 7: 暗証番号的なかな文字 4 桁 (設問 3) の母音の出現回数

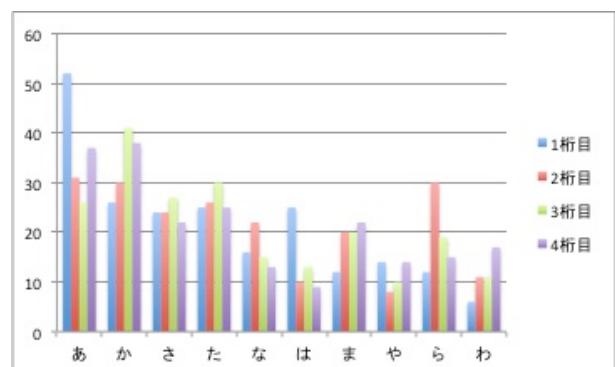


図 8: 暗証番号的なかな文字 4 桁 (設問 3) の子音の出現回数

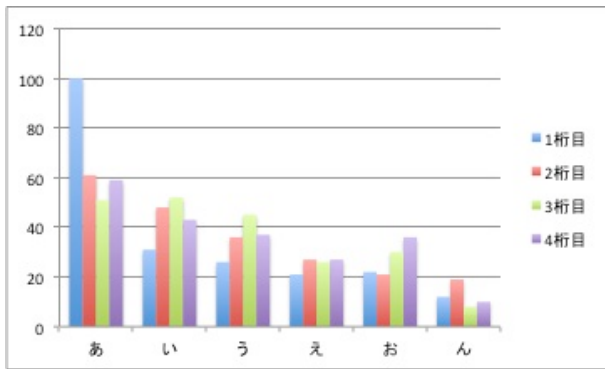


図 9: でたらめなかな文字 4 桁 (設問 7) の母音の出現回数

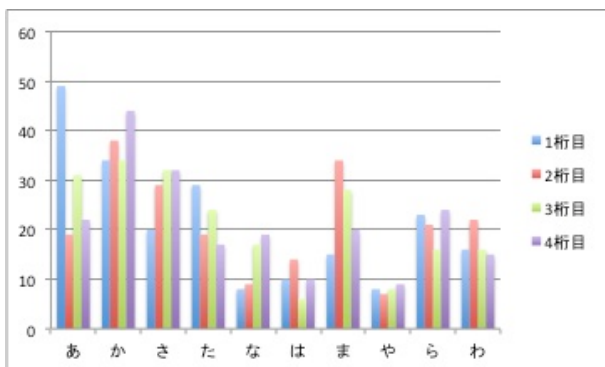


図 10: でたらめなかな文字 4 桁 (設問 7) の子音の出現回数

各桁の出現回数 図 7 と図 8 にそれぞれ設問 3 (暗証番号的な 4 桁) と設問 7 (でたらめな 4 桁) における各桁の母音・子音の出現回数を掲載する。同様に、図 9 と図 10 にそれぞれ設問 3 と設問 7 おける各桁の母音・子音の出現回数を掲載する。ただし、母音については「ん」のみ特別扱いし、子音については「わをん」を全て「わ行」として統一してある。暗証番号的な 4 桁とでたらめな 4 桁の間に共通する特徴として、次のようなものが挙げられる。

- 1 桁目は「あ」を選ぶ確率が高い。この現象は、「あ」がかな文字の先頭であり、最も連想しやすい音素であることが原因と思われる。
- 母音について、「あ段」の母音が最も出現回数が高い。「い段」、「う段」、「お段」と続き、「え段」が最も出現回数が低い。
- 子音について、「あ・か・さ・た」の行が、ほぼ「あ」、「か」、「さ」、「た」の順に出現回数が高い。この事実、日本人が五十音を連想するとき「あかさたなはまやらわ」のように唱和することに影響している可能性を示唆している。

日常会話の音節分布との相関係数 文献 [3] では、日常会話に登場する音節の出現頻度が一覧表として掲載されている (以下、この分布を日常会話の音節分布と

呼称する)。今回調査した暗証番号的な 4 桁と、でたらめな 4 桁のそれぞれについて、文献 [3] の日常会話の音節分布との相関係数を算出した。その結果、暗証番号的な 4 桁と日常会話の音節分布は、相関係数が 0.558、でたらめな 4 桁と日常会話の音節分布は、相関係数が 0.434 であり、どちらも正の (中間的強さの) 相関であることが分かった。特に注目すべき点として、暗証番号的な 4 桁はでたらめな 4 桁よりも情報量が (わずかに) 小さかったのに対し、日常会話の音節分布との相関はより強くなっていることが挙げられる。この事実、次のように解釈することが出来る。

- 暗証番号的な 4 桁の方が日常で用いられる単語など (自然言語) が使用されやすい。
- でたらめな 4 桁のとき、被験者はなるべく意味のない文字を選ぼうとするが、無意識に自然言語に近い分布から選んでしまう。
- でたらめに選ぼうとした場合よりも、(暗証番号的な 4 桁のように) 自然言語ベースで選んだ場合の方がランダム性が高くなる場合がある。

4.5 英数字の特徴

設問4 (暗証番号的)	出現回数	設問8 (でたらめ)	出現回数
英英数数	47	英数数英	27
英数英数	27	英英数数	27
英数数英	23	英数英数	23
英英英数	23	数数英英	17
英数英英	17	数英数英 英英英数	16

図 11: 英数字 (設問 4、8) の出現回数 (上位)

	1桁目	2桁目	3桁目	4桁目
設問4 (暗証番号的)	a (24)	1 (22)	1 (30)	2 (24)
	t (23)	k (19)	2 (18)	1 (16)
	k (15)	0 (13)	0 (14)	3 (15)
	s (14)	2 (13)	k (13)	4 (14)
	j (12)	b (13)	4,8,b,s (9)	0,5,7 (11)
設問8 (でたらめ)	a (30)	2 (25)	2 (22)	9 (17)
	2 (19)	9 (14)	b (15)	1 (14)
	0 (15)	k (14)	1 (13)	8 (14)
	j (12)	a (13)	a (13)	b (14)
	3,c,k (11)	3,b (12)	9 (12)	3,5 (13)

図 12: 英数字 (設問 4、8) の並びの出現回数

図 11 に、設問 4 (暗証番号的な 4 桁) と設問 8 (でたらめな 4 桁) のそれぞれにおける各桁の出現回数の多い文字を、図 12 に、英数字の並びの出現回数の多いものを掲載する。まず図 11 において、設問 4 では、1 桁目には英字が、3、4 桁目には数字が使用される頻度が非常に高かった。またその英字には a, t, k, s の出

現回数が多く、これは英字のみの暗証番号における結果と類似していた。数字の特徴としては、3桁目には0, 1, 2が他に比べ多く出現しているが、4桁目では目立った偏りは無く、ばらついていた。設問8では、他の桁に比べ4桁目に数字が使用される頻度が高かったが、その他に目立った偏りは無かった。

次に図12において、設問4では、「英字 英字 数字 数字」の並びが最も多く、全体の約1/4を占めていた。設問8では、設問4に比べ並びの偏りは小さく、突出したものは無かった。

上記の結果を踏まえ、「英字 英字 数字 数字」となるものの中で、3桁目と4桁目を組み合わせた2桁の数字が日単位(01~31)の範囲内である個数を調べると、設問4では42個(占有率89%)、設問8では14個(占有率52%)であった。

以上より、英数字を使用して設問4の「覚えておくことができ、なるべく他人に推測されない」4桁の番号を作る時、1,2桁目には日本語会話における単音節の出現頻度の高い英字が、3,4桁目には2桁合わせて日単位となる数字が使用されていることから、名前などの言葉と、誕生日などの日付を組み合わせて作っている傾向があると考えられる。また、設問8のように「でたらめ」とした場合は、設問4に比べ文字の並びや使用される文字に偏りが少なかったため、言葉や日付は意識していないと考えられる。

5 結論

純粹に安全性(Guess Work)のみを考慮した場合、かな文字、英数字、英字、数字の順に暗証番号として用いられることが望ましい。(すなわち、用いることのできる暗証番号の総数が多ければ多いほど安全性が高い。)一方、一様ランダムな分布との近さは、数字、かな文字、英字、英数字の順である。すなわち、かな文字・英字・英数字は、数字に比べて安全性は高いものの、ランダム性は低い。

興味深いことに、「用いることのできる暗証番号の総数」が多いほどランダム性が下がるわけではなく、最も総数の多いかな文字は、英字・英数字に比べてランダム性が高い。したがって、安全性とランダム性を同時に考慮した場合、かな文字の暗証番号か数字の暗証番号が良い性質を持っていると結論づけられる。

6 今後の課題

本稿では、異なる文字を用いた暗証番号の安全性とランダム性についてのみ議論を行ったが、以下のような課題が残っている。

- アンケートの設問において、入力範囲を明示的に書き、記入例を与えたが、それらがバイアスと

なっている可能性がある。そのため、いくつかの設問を用意する等、バイアスを考慮した実験を行う。

- 今回は日本語を母語とする学生を対象としたアンケート調査を行ったが、日本語以外の母語を持つ被験者で同じ実験を行うと、ランダム性などに変化が出る可能性がある。
- 反対に、日本語を母語とする学生に対して、ハングル文字やアラビア文字など、母語以外の文字を用いて実験を行うことも考えられる。

参考文献

- [1] Joseph Bonneau, Sören Preibusch and Ross J. Anderson, "A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs," Financial Cryptography and Data Security - 16th International Conference, FC 2012, Kralendijk, Bonaire, February 27-March 2, 2012, Revised Selected Papers, p25-40, 2012.
- [2] 独立行政法人情報処理推進機構: パスワード - もっと強くキミを守りたい -, <http://www.ipa.go.jp/security/keihatsu/munekyun-pw/>
- [3] 小寺 一興, 平石 光俊, "日本語会話における単音節の出現頻度: 語音明瞭度検査の語表構成の検討," Audiology Japan, 03038106, p73-78, 1998.

付録

アンケートご協力のお願い

私たちの在籍するシステム情報工学研究科リスク工学専攻では「グループ演習」という必修科目があり、数人の学生がグループを作り、与えられた課題について研究調査をします。

私たちのグループで取り組んでいる課題では多くの方に「数字などの文字列を選んでもらう」という作業をお願いする必要があり、学群生の皆様にご協力を頂きたいと思い、今回アンケートをお願いする次第です。

このアンケート調査は、リスク工学専攻のグループワーク課題を調査するために実施するものです。調査は無記名で回答いただくものであり、調査結果は上記目的以外に絶対に使用しません。

学群生の皆様ならびに担当科目の先生方にはご迷惑をおかけしますがご協力いただけると幸いです。何卒よろしく願いいたします。

システム情報工学研究科リスク工学専攻開設科目「グループ演習」第1班
第1班班長 品川和雅（リスク工学専攻博士前期課程1年）
第1班指導教員 金山直樹（システム情報系情報工学域・助教）

設問 1

覚えておくことのできる 4桁の数字 (0～9) で、なるべく他人に推測されないものを書いてください。

(記入例) 4284、5031

--	--	--	--

設問 2

覚えておくことのできる 4文字のアルファベット (a～z) で、なるべく他人に推測されないものを書いてください。

(記入例) kszi、ccaz

--	--	--	--

設問 3

覚えておくことのできる 4文字のひらがな (あ～ん) で、なるべく他人に推測されないものを書いてください。ただし、濁点・半濁点 (例：が、ず、ぱ) や小さい文字 (例：っ、ゃ、ょ) 以外の文字を選んでください。

(記入例) うきんう、せまとけ

(悪い例) がすりこ (濁点が入っている)

--	--	--	--

設問 4

覚えておくことのできる 4文字のアルファベットと数字 (a～z、0～9) で、なるべく他人に推測されないものを書いてください。

(記入例) j31b、bkk2

--	--	--	--

次のページに進んでください。

設問 5

4桁の**数字**（0～9）をでたらめに書いてください。

（記入例）4284、5031

--	--	--	--

設問 6

4文字の**アルファベット**（a～z）をでたらめに書いてください。

（記入例）kszi、ccaz

--	--	--	--

設問 7

4文字の**ひらがな**（あ～ん）をでたらめに書いてください。ただし、濁点・半濁点（例：が、ず、ぱ）や小さい文字（例：っ、ゃ、よ）以外の文字を選んでください。

（記入例）うきんう、せまとけ

（悪い例）がすりこ（濁点が入っている）

--	--	--	--

設問 8

4文字の**アルファベット**と**数字**（a～z、0～9）をでたらめに書いてください。

（記入例）j31b、bkk2

--	--	--	--

次のページに進んでください。

