大学生のための統計学入門

ーテーマ2. 代表的なグラフー

京都大学大学院医学研究科

臨床統計学/生物統計家育成コース　田中司朗

テーマ2. 代表的なグラフ

- 分布とグラフ
- **度数分布表の活用事例**
- ヒストグラムの作り方

羊皮紙の暗号を解く

- 難破船の残骸のある海辺で, 黄金虫と羊皮紙を拾った. 羊皮紙を偶然ろうそくにかざすと熱反応で髑髏の模様が浮かび上がった. 更に加熱すると山羊の模様が現われ, そしてそれらの間には以下のような暗号が現われた

53‡‡†305))6*;4826)4‡.)4‡);806*;48†8¶60))85;1‡(;:‡*8†83(88)5*†;46(;88*96*?;8)*‡(;485);5*†2:*‡(;4956*2(5*4)8¶8*;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4)485†528806*81(‡9;48;(88;4(‡?34;48)4‡;161;:188;‡?;

黄金虫. エドガー・アラン・ポー著

質問

- この暗号をどうやって解きますか?
- ヒント
  - アルファベットを記号・数字に置き換えて作った

53‡‡†305))6*;4826)4‡.)4‡);806*;48†8¶60))85;1‡(;:‡*8
†83(88)5*†;46(;88*96*?;8)*‡(;485);5*†2:*‡(;4956*2(5
4)8¶8;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4
)485†528806*81(‡9;48;(88;4(‡?34;48)4‡;161;:188;‡?;

度数分布表による推理

- 最もよく用いられるアルファベットは e だ
- それ以降は, aoidhnrstuycfglmwbkpqxzの順だ
- 暗号の中でどの記号がよく用いられているか集計してみよう

記号	度数	割合	記号	度数	割合
8	33	20.8%	†、1	8	5.0%
;	26	16.4%	0	6	3.8%
4	19	11.9%	9、2	5	3.1%
‡、)	16	10.1%	:、3	4	2.5%
*	13	8.2%	?	3	1.9%
5	12	7.5%	¶	2	1.3%
6	11	6.9%	-、.	1	0.6%

度数分布表による推理

- 最も多い記号は8なので, これが eを表わすと仮定する
- 英語ではeが二つ続く頻度は非常に高い
- そこで, 8が連続して現われるかどうかを調べてみると, 五回も重なって出てくる

記号	度数	割合	記号	度数	割合
8	33	20.8%	†、1	8	5.0%
;	26	16.4%	0	6	3.8%
4	19	11.9%	9、2	5	3.1%
‡、)	16	10.1%	:、3	4	2.5%
*	13	8.2%	?	3	1.9%
5	12	7.5%	¶	2	1.3%
6	11	6.9%	-、.	1	0.6%

本当にeaoid…の順なのか
別のサンプルで確かめてみよう

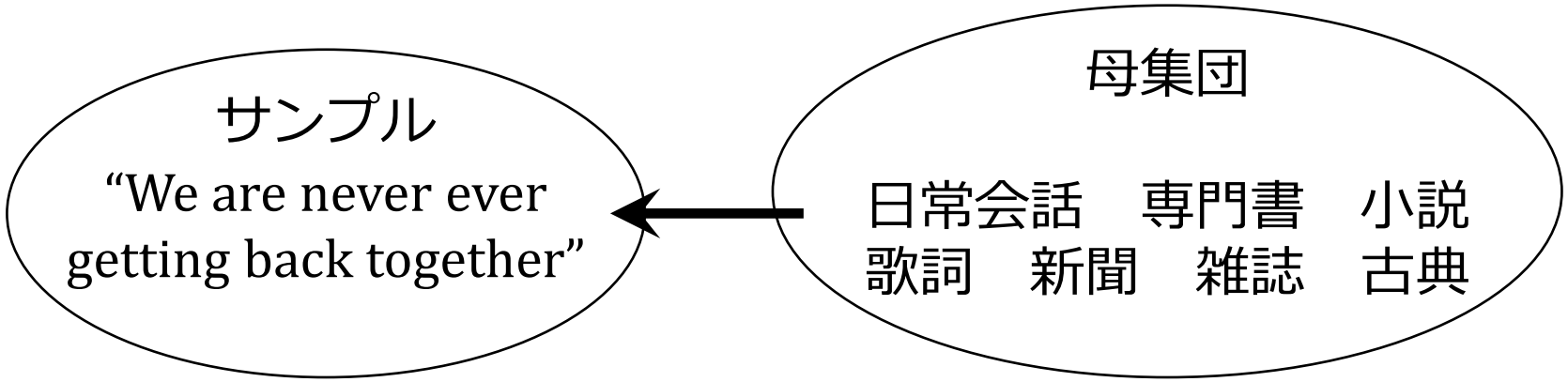
We are never ever getting back together

作詞 Taylor Swift/Max Martin/Shellback, 作曲 Taylor Swift/Max Martin/Shellback

歌詞（サビ） アルファベット使用頻度

記号	度数	割合	記号	度数	割合
e	38	23.3%	c	3	1.8%
t	19	11.7%	h	3	1.8%
r	19	11.7%	l	3	1.8%
g	10	6.1%	u	3	1.8%
a	9	5.5%	w	3	1.8%
o	9	5.5%	y	3	1.8%
v	9	5.5%	d	2	1.2%
n	8	4.9%	f	2	1.2%
k	6	3.7%	m	2	1.2%
l	5	3.1%	s	2	1.2%
b	4	2.5%	v	1	0.6%

We are never ever getting back togetherより引用
作詞Taylor Swift/Max Martin/Shellback, 作曲: Taylor Swift/Max Martin/Shellback

なぜ二番目以降が違ったのか?

- **サンプル**が**母集団**を代表していなかった
  - “We are never ever getting back together”はラブソングという特殊なジャンル
  - 英語文章一般については選び方が偏っていた
- 統計学では, 一部のサンプルを調べて, 全体のことを推測する, という考え方をすることがある