統計入門

公衆衛生活動で困らないために

平成30年度　行政歯科保健担当者研修会

鶴見大学　歯学部探索歯学講座

野村義明

Contents

Contents

統計でできること

偶然でない差があるか。

偶然でない関連があるか。

→ 有意差
有意な関連

↓

簡単に言えば有意確率を求めるために分析に使う道具

発展型

複雑なデータから意味のあるものを見つける
統計モデルを作製しパターン予測、将来予測

統計でできることの例

体重は男と女で差があるか。

身長と体重は関連があるか。

差、関連はある。その差、関連が偶然かそうでないのかを示すものが有意確率

簡単に言えば**有意確率**を求めるために分析に使う道具

統計学の考え方

- 全国の都道府県を対象に人口と出生率の関連について調査した。
  その結果、人口と出生率の間に有意な相関がみられた。

1. 回答率は10%であった。
2. 回答率は50%であった。
3. 回答率は100%であった。

致命的な誤りはどれ？

歯科医師国家試験問題　110回　C-59

59　A小学校で2001～2015年の間、全児童を対象に実施した齲蝕予防事業の評価結果を表に示す。この結果からでは、本事業に齲蝕予防効果があったと言えない。

年度	児童数	DMF歯数(平均値)	DMF歯数(平均値)の差の検定結果
2000	100	2.3	$p < 0.01$
2015	50	1.5	

その理由はどれか。1つ選べ。

a　児童数が同じでない。

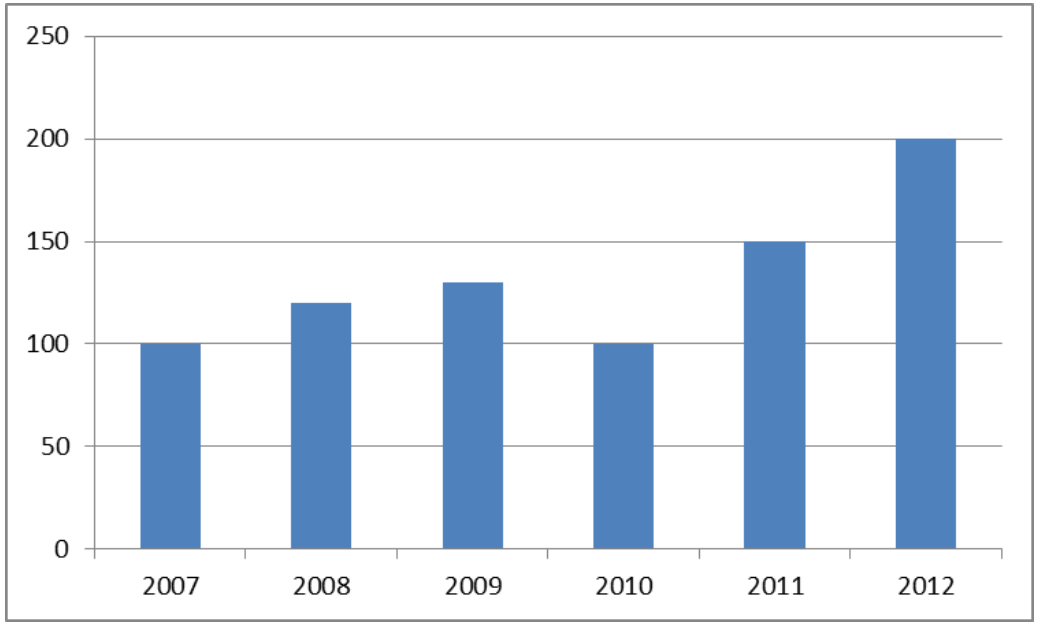
b　ランダム化比較試験でない。

c　DMF歯数(平均値)の差が小さい。

d　齲蝕のない児童の割合が分からない。

e　事業を実施していない他校のデータがない。

- 警察官の犯罪は増加傾向にある。



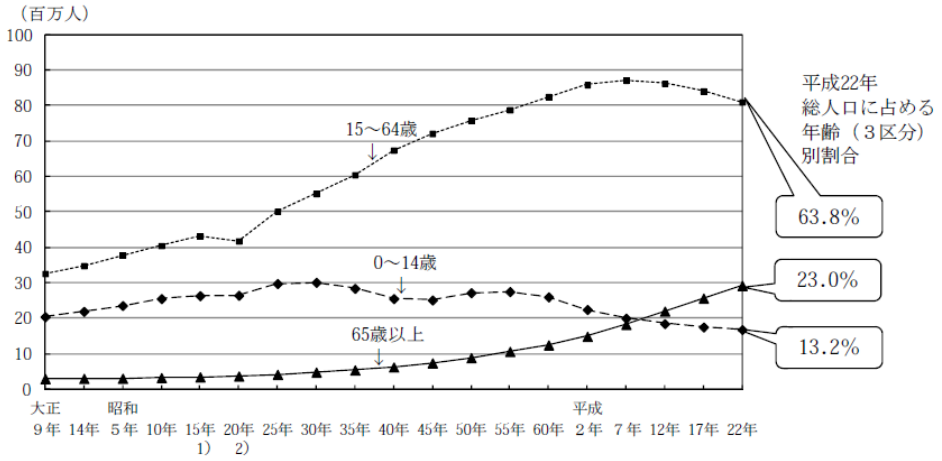
仮想データ

国勢調査ではなぜ統計学的検定がされていないのか？

1　全国の人口

65歳以上人口は13.9%増，総人口に占める割合は20.2%から23.0%に上昇
15～64歳人口は3.6%減，割合は66.1%から63.8%に低下
15歳未満人口は4.1%減，割合は13.8%から13.2%に低下

図Ⅱ－１－１　年齢（３区分）別人口の推移－全国（大正９年～平成22年）

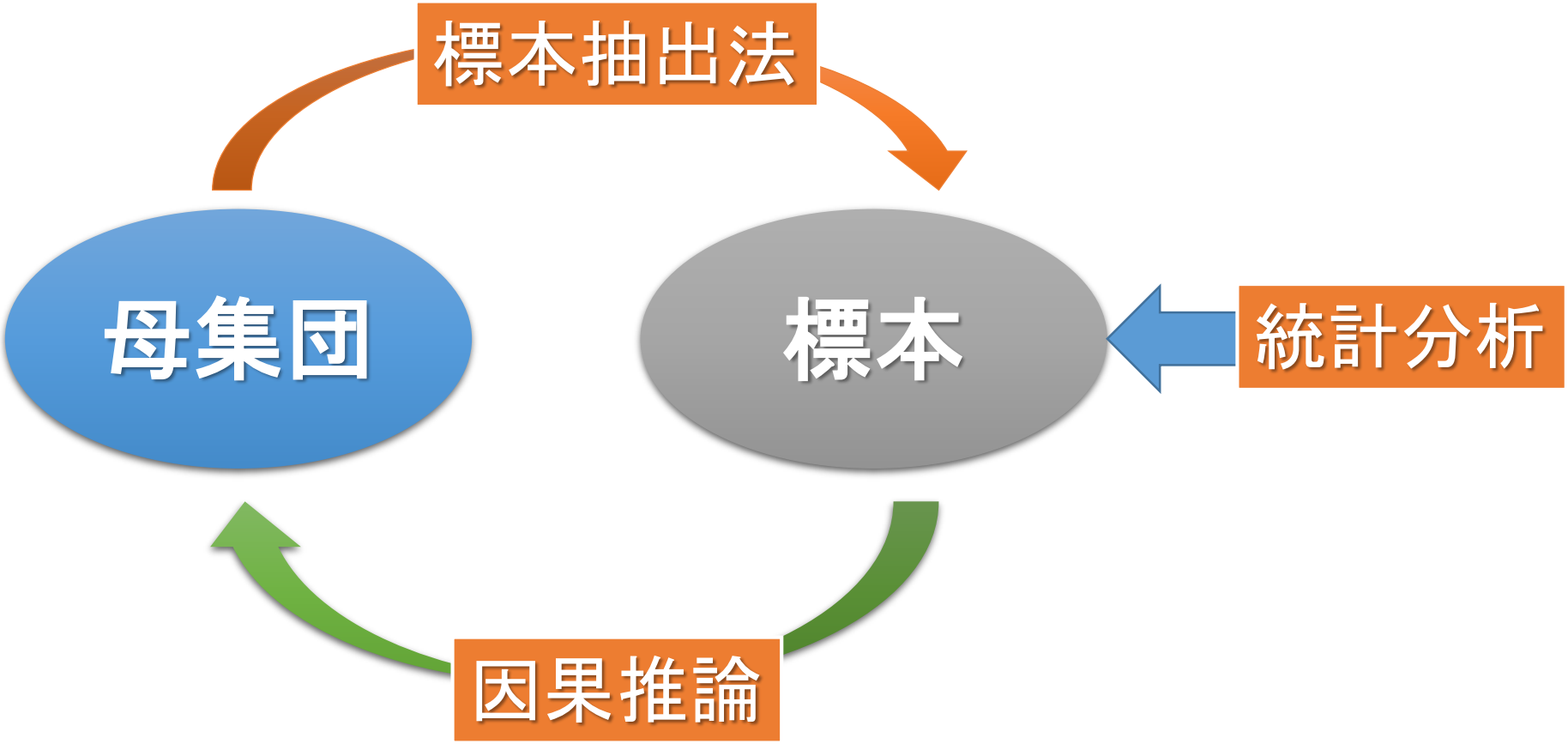


（注）昭和20年は人口調査結果による。
1）朝鮮，台湾，樺太及び南洋群島以外の国籍の外国人（39,237人）を除く。
2）沖縄県を除く。

国民健康栄養調査

2．糖尿病に関する状況

「糖尿病が強く疑われる者」の割合は、男性 19.5%、女性 9.2%である。平成 18 年以降でみると、男女とも有意な変化はみられなかった。



時系列分析などの例外も多くあるが基本はこれ！

- 全国の都道府県を対象に人口と出生率の関連について調査した。
  その結果、人口と出生率の間に有意な相関がみられた。

1. 回答率は10%であった。
2. 回答率は50%であった。
3. 回答率は100%であった。

致命的な誤り

- 統計は標本調査から得られたデータを分析し差や関連が母集団で当てはまるかを検討するもの。

母集団

全数調査が不可能な場合標本調査によって
母集団を推定する。

- 製品の破壊検査
- 麺のゆで具合
- 化粧品のサンプル提供
- 食料品売り場の試食

偏った標本から推定することの危険性

- 相撲取りから日本人の平均体重を予測
- 小児のう蝕も患者数も減っていない。（教授の発言）
- 平成教育委員会の正答率

バイアス(英：bias)

斜め、または偏りや歪みを意味し、転じて偏見や先入観という意味をもつ。

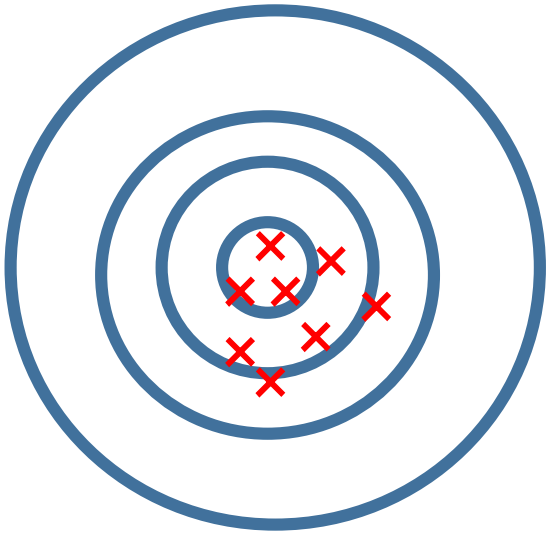
心理学や社会学などの統計から一般論を導く分野で使われることが多い

系統的な誤差のこと

（ランダムな誤差をばらつきという）

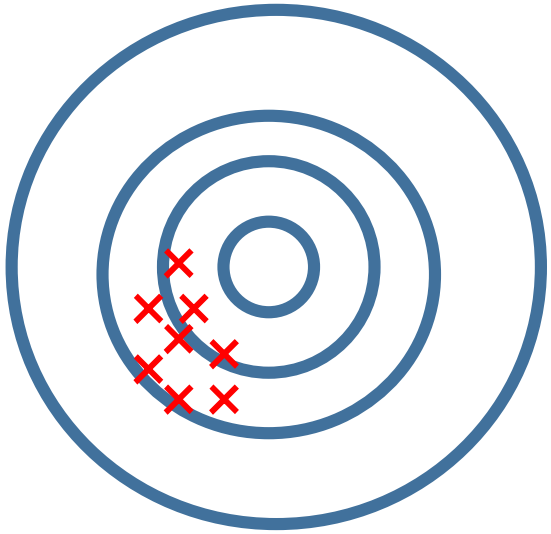
ばらつきとバイアス

ばらつき



ランダムな誤差
（偶然誤差）

バイアス



系統的な誤差
（系統誤差）

測定値には必ず測定誤差を含む
測定誤差は一般的に正規分布にした学ことが知られている
誤差を含む測定値から一定の傾向を見いだすのが統計学

母集団の分布等が明らかになっている場合などを除き
統計学の技法では見いだすことが困難

系統誤差を制御する方法

フィッシャーの3原則

- 無作為化
- 反復
- ブルック化

実験計画法の基本

疫学では必ずしも適用できない

標本抽出法

- （単純）無作為抽出

  無作為に母集団から抽出する方法

- 系統抽出

  はじめの一つを無作為に抽出後、等間隔に抽出する方法

- 層化抽出法

  母集団の性質に併せて母集団を区分しそこから無作為に抽出する方法

- 集落抽出法

  母集団を無作為に区分し、その幾つかを調査

Contents

Part I　Basics

- 統計の基本的な考え方
- 統計学的検定の基本

Part II　Advanced

- 回帰分析の注意点

尺度、変数

1. （男＋女）÷2=
2. 1位と3位の平均順位は?
3. 3位-1位＝
4. 宝くじの1等-2等＝
5. 宝くじの4等-5等＝
6. 30℃は10℃の何倍暑い？
7. 30℃-20℃＝
8. 30㎝は10㎝の何倍？

尺度

- 名義尺度　名前、性別、血液型
- 順序尺度　1位、2位
- 間隔尺度　気温、体温など
  原点が任意に決められる
  0に意味がない
- 比例尺度　身長、体重、時間、お金
  （比尺度）　原点が決められている
  0に意味がある

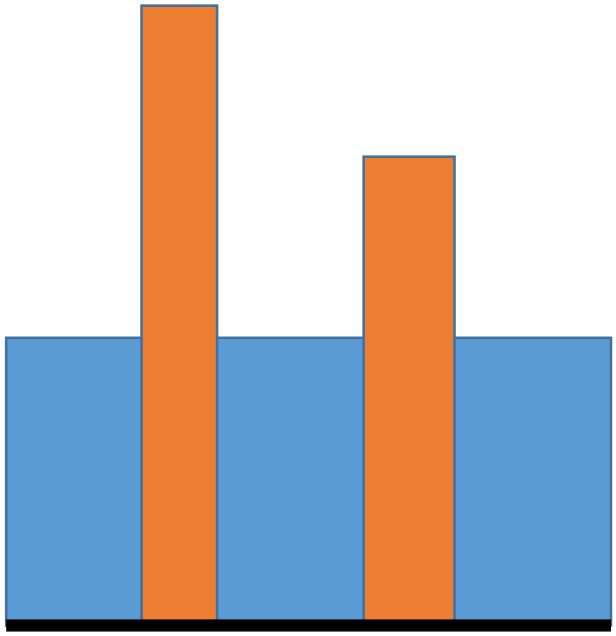
尺度の基本

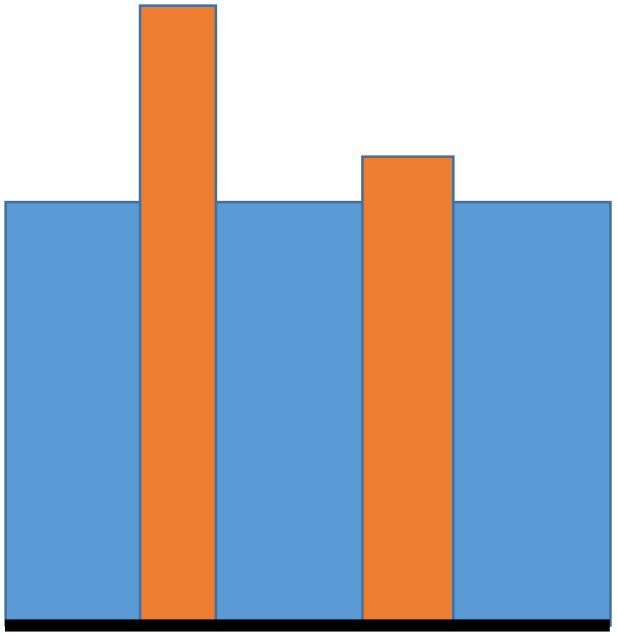
- 名義尺度　分類のみ。　順序もない。
- 順序尺度　等間隔ではない。
- 間隔尺度　原点が任意、乗除は不可。
- 比尺度　　原点が一意。

歯科の指標と尺度

CPIは順序尺度
平均値を計算してはいけない。

間隔尺度と比尺度





差？
倍？

統計で使用する変数と尺度のずれ

尺度	統計で使用する変数
名義尺度	カテゴリー変数 質的変数
順序尺度	
間隔尺度	連続変数 量的変数
比尺度	

統計学的検定方法の選択
基本パターン

統計処理方法は変数の組み合わせで決まる

- 質的変数と質的変数　　χ^2検定
- 質的変数と量的変数　基本はt検定
- 量的変数と量的変数　相関分析

パラメトリック検定とノンパラメトリック検定

- データが正規分布にしたがう場合
- パラメトリック検定

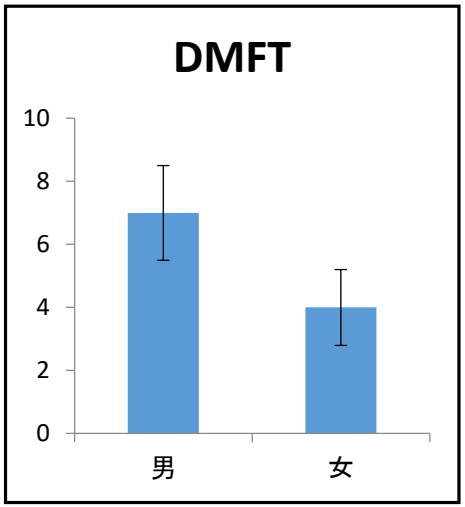
- データが正規分布に従わない場合
- ノンパラメトリック検定

データの整理の方法で統計学的検定方法は決まる

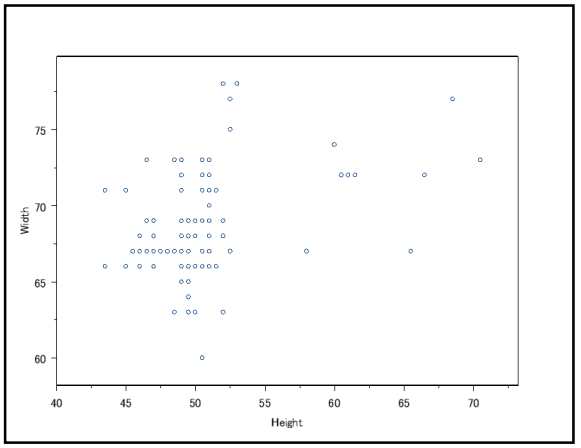
	DMFあり	DMFなし	合計
男	50	50	100
女	30	70	100
合計	80	120	200

カイ2乗検定

Advance
・対応のある分割表
McNemar検定
・交絡の調整をしたいとき
Cochran–Mantel–Haenszel検定
・どこに有意なセルがあるか知りたいとき
対数線型モデル

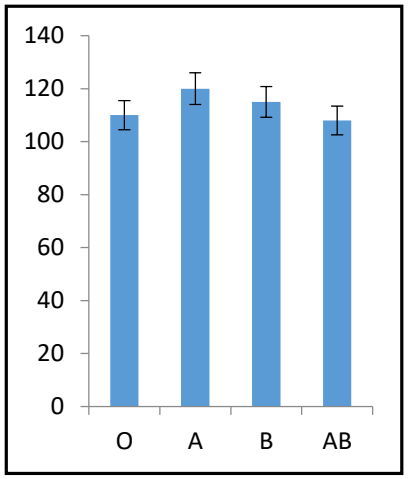


二標本検定
Mann　WhitneyのU検定



相関分析

Advance
- 重回帰分析
- 一般化線型モデル
- 線形混合モデル



一元配置分散分析
Kruscal Wallis検定

Advance
- 線型モデルへの拡張
- 実験計画法
- 多重比較

統計処理方法のまとめ

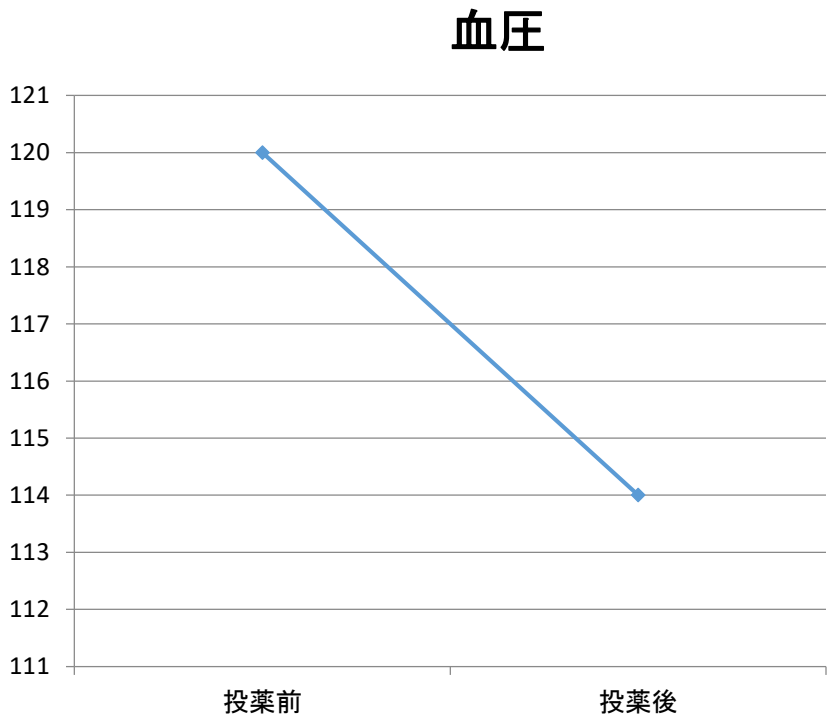
質的変数と質的変数	X^2検定	分割表 クロス集計表
量的変数と量的変数	ピアソンの相関係数 相関分析	相関図

		パラメトリック	ノンパラメトリック
対応なし	2群	二標本t検定	Mann WhitnyのU　test
	3群以上	一元配置分散分析	Kruscal Wallis検定
対応あり	1群の前後	対応のあるt検定	Wilcoxsonの符号検定

変数の種類が見分けられれば分析可能

注：Wilcoxsonの符号付き順位和検定は　Mann WhitneyのU検定と同義
Wilcoxsonの符号検定とは別物

前後比較をする場合



対応のあるt検定

注：同じ人の前後で統計処理する。

Contents

Contents

Part I　Basics
- 統計の基本的な考え方
- 統計学的検定の基本

Part II　Advanced
- 回帰分析の注意点
  - 確率分布
  - 一般化線形モデル
  - 線形混合モデル
  - アンバランスなデータ
  - 0が多いモデル

統計の基本は分布です。

離散分布

- 一様分布
- ベルヌーイ分布
- 二項分布
- ポアソン分布
- 幾何分布
- 負の二項分布
- 超幾何分布

連続分布

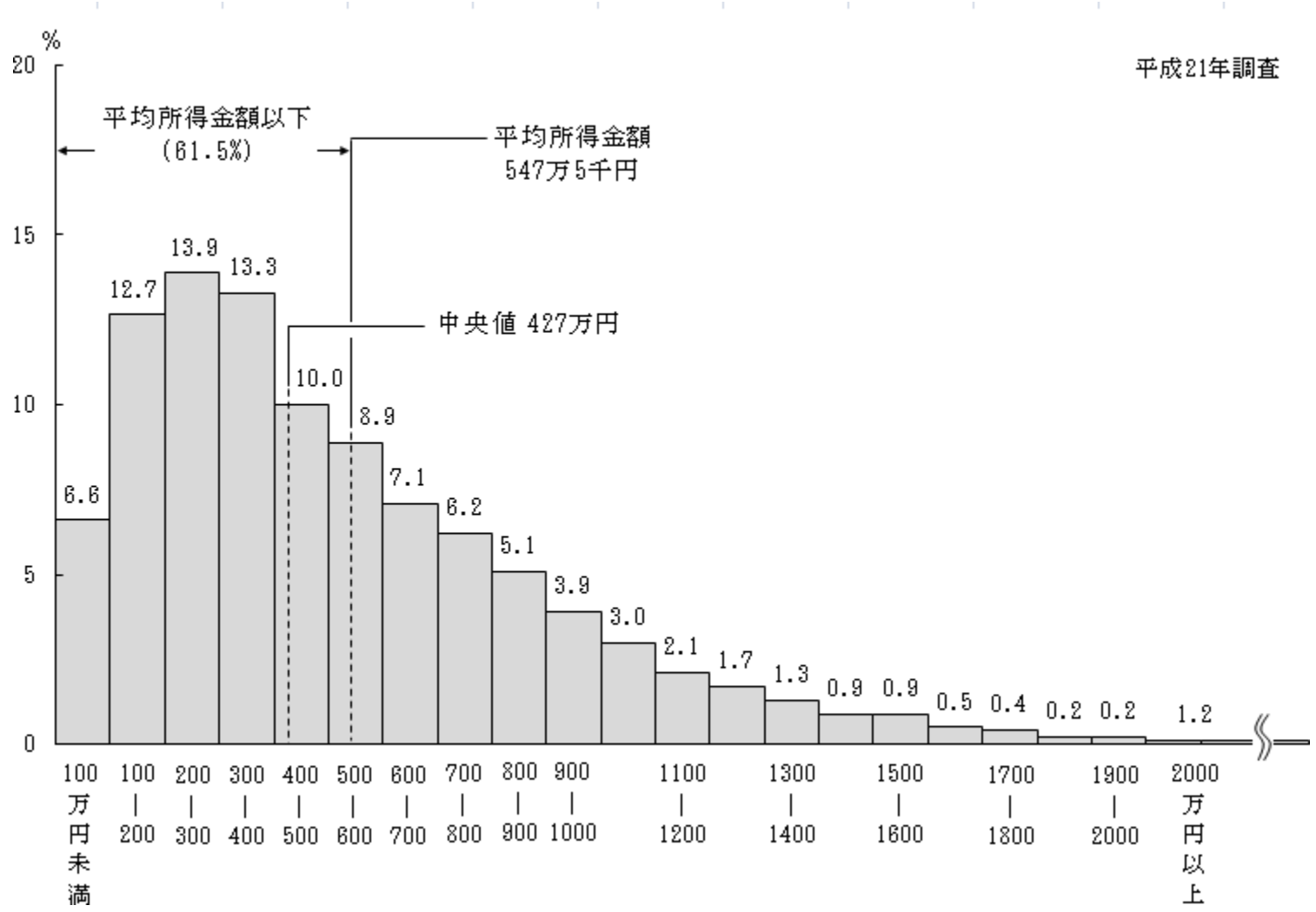
- 一様分布
- 正規分布
- ガンマ分布
- 指数分布
- ベータ分布

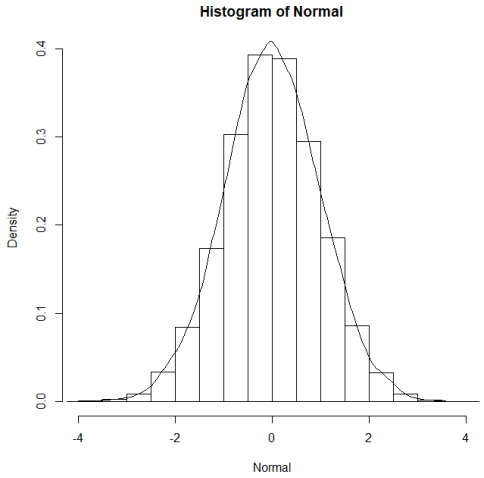
標本分布

- X^2 分布
- t 分布
- F分布

その他：ラプラス分布、アーラン分布、ヂレクリ分布、ガンベル分布

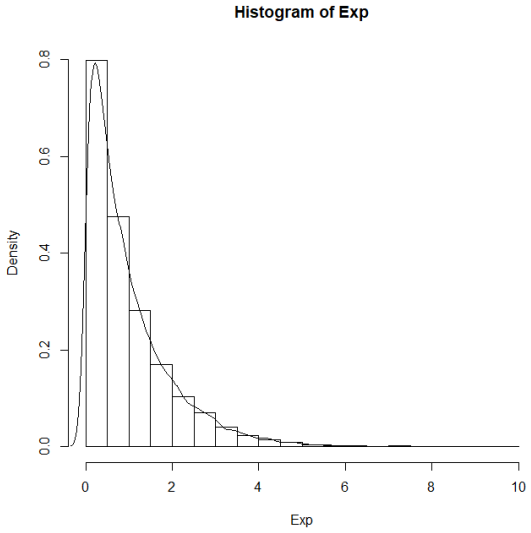
所得金額階級別にみた世帯数の相対度数分布



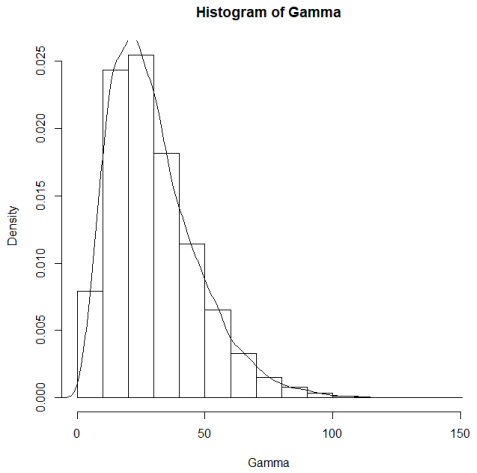


正規分布

$$P(X = x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-} \frac{(x-\mu)^2}{\sigma^2} dx$$



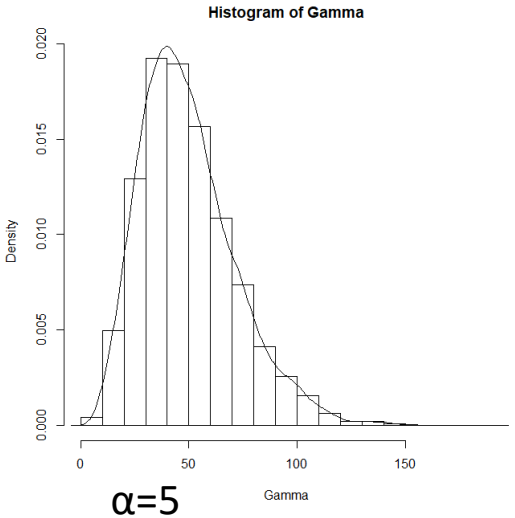
α=1



α=3

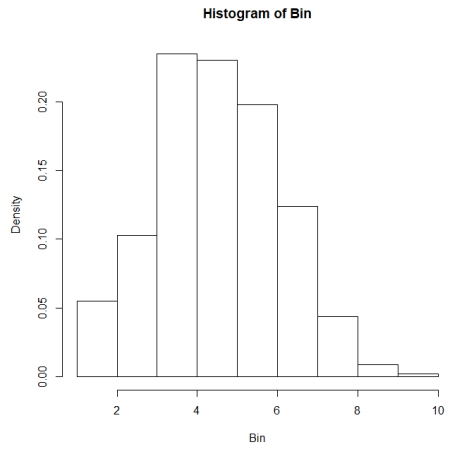
ガンマ分布

$$P(X = x|\alpha, \beta) = \int_0^{-1} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{-\alpha} e^{-\frac{1}{\beta}} dx$$



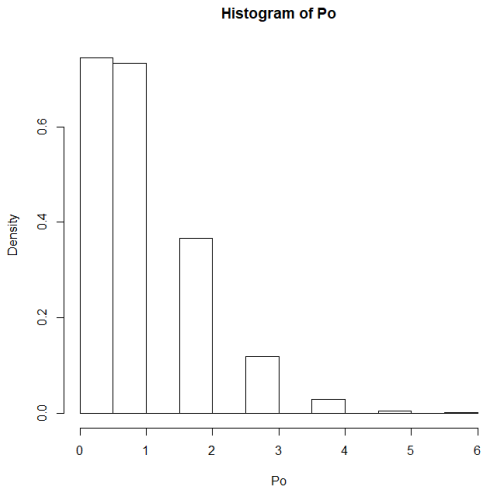
α=5

2項分布



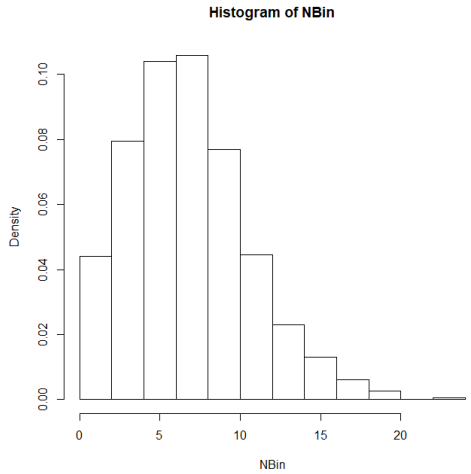
$$Bin(n,p) = \sum_{0}^{\infty} \binom{n}{p} p^k (1-p)^{n-k}$$

ポアソン分布



$$P_o(x) = \sum_{0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}$$

負の2項分布



$$NBin(k,r) = \sum_{0}^{\infty} \binom{k+r-1}{r} p^r (1-p)^k$$

$$= \sum_{0}^{\infty} \binom{-\alpha}{r} p^{\alpha} (-p)^k$$

Contents

Part I　Basics
- 統計の基本的な考え方
- 統計学的検定の基本

Part II　Advanced
- 回帰分析の注意点
  - 確率分布
  - 一般化線形モデル
  - 線形混合モデル
  - アンバランスなデータ
  - 0が多いモデル

線形モデル

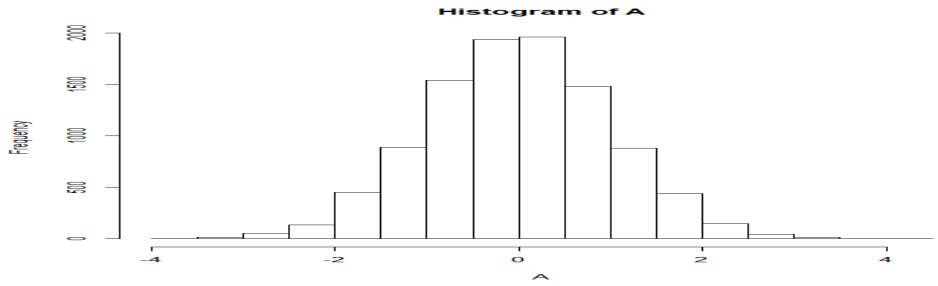
- 単純だけと知識がないと誤った結論を導きやすい

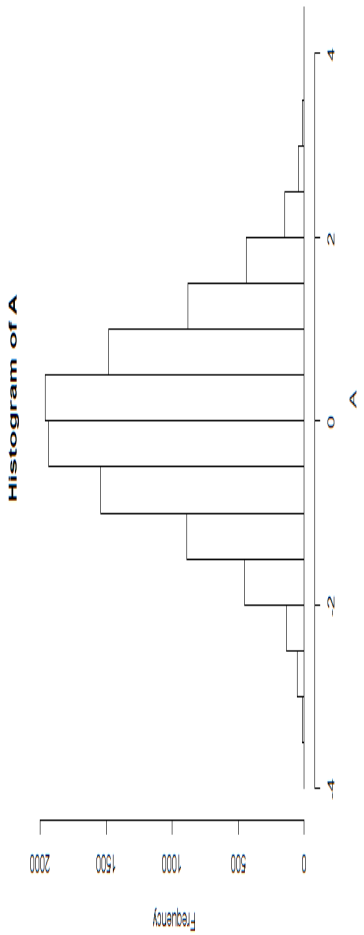
単回帰

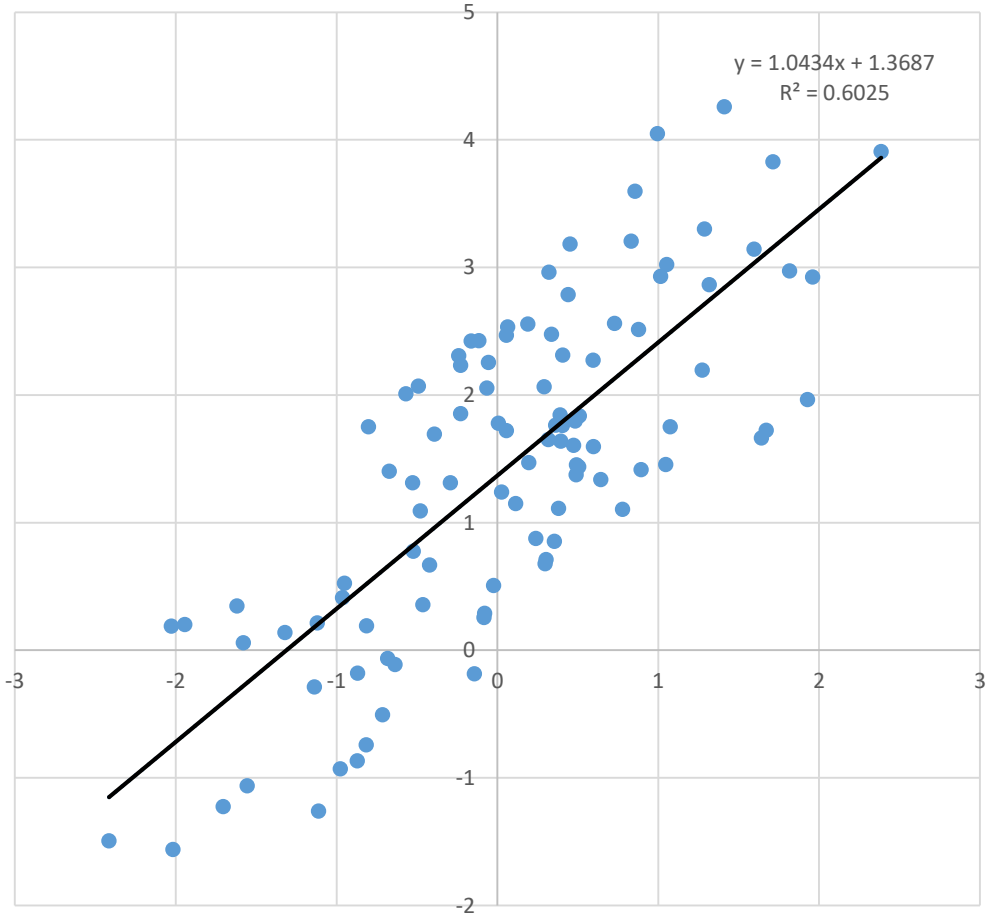
$Y=ax+b$

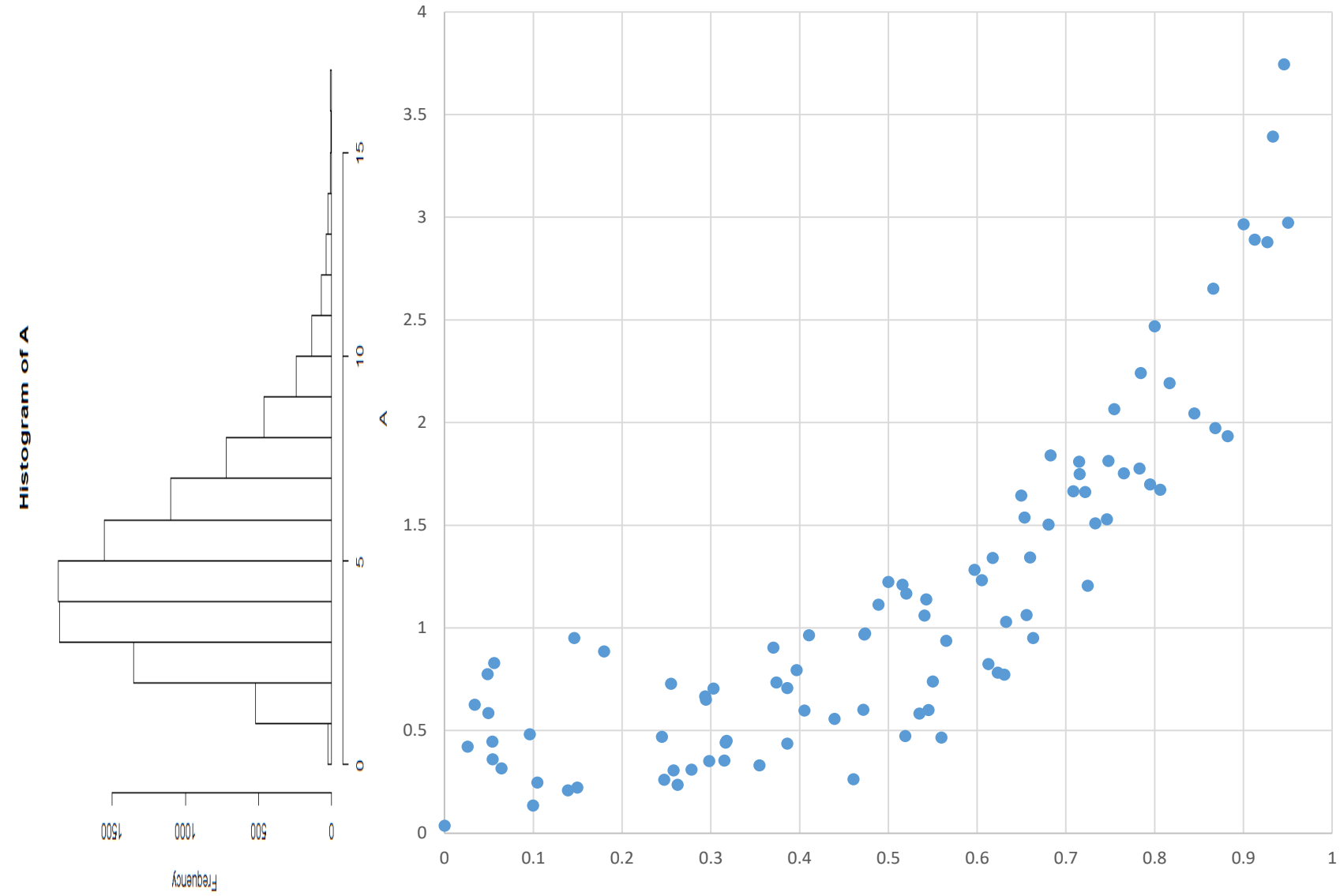
重回帰

$Y=ax_1+ax_2+・・・・・・・・+b$


Histogram of A
Frequency
0
500
1000
1500
2000
-4
-2
0
2
4
A


Histogram of A
A
4
2
0
-2
-4
0
500
1000
1500
2000
Frequency


y = 1.0434x + 1.3687
R² = 0.6025
5
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3


Histogram of A
Frequency
0
500
1000
1500
0
5
10
15
A
0
0.5
1
1.5
2
2.5
3
3.5
4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

一般化線形モデル

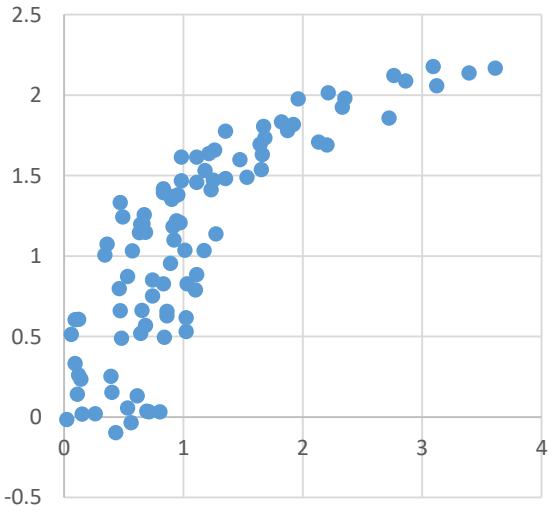
$$y = \alpha x + b$$

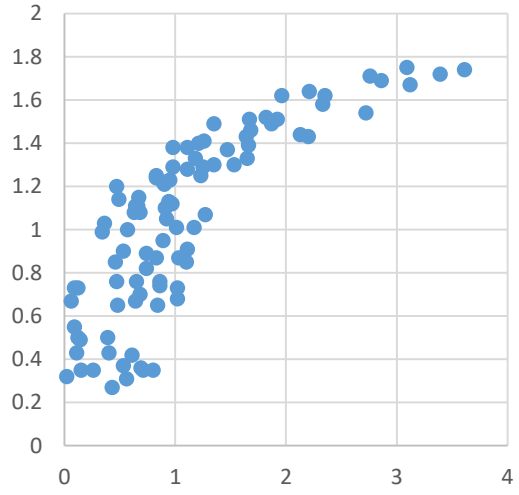
$$\Gamma(y) = ax + b$$

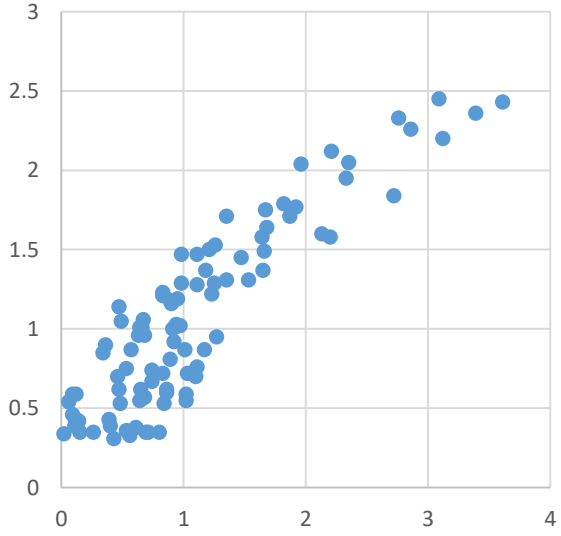
$$\Gamma(y) = e^{ax+b}$$

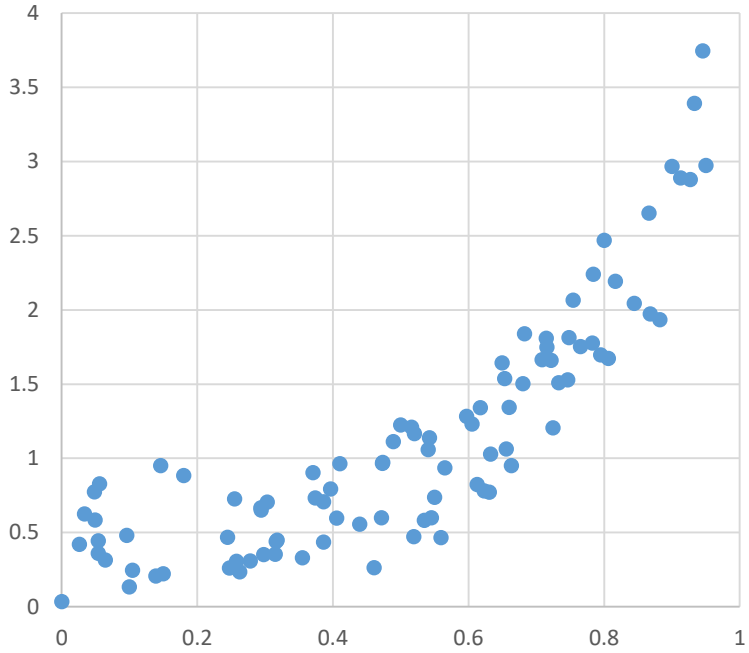
$$\Gamma(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{-\alpha} e^{-\frac{y}{\beta}}$$

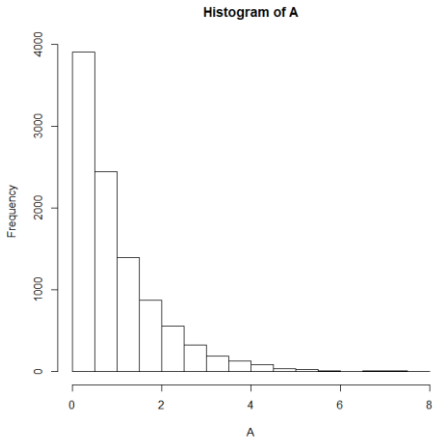
		95% 信頼区間		有意確率	AIC
		下限	上限		
(切片)	-0.10	-0.30	0.10	0.33	127.65
X	2.39	2.04	2.75	0.00	
(切片)	0.27	0.18	0.36	0.00	132.21
X	1.55	1.25	1.86	0.00	
(切片)	-1.17	-1.38	-0.96	0.00	114.64
X	2.17	1.80	2.54	0.00	



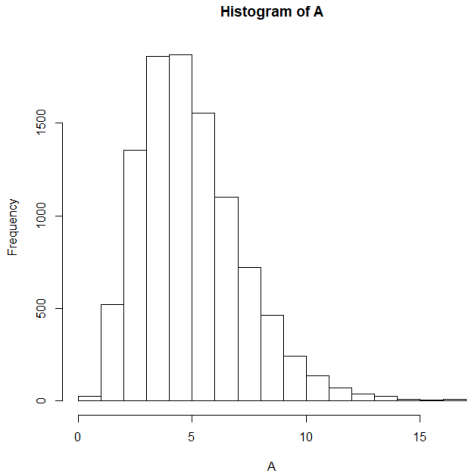




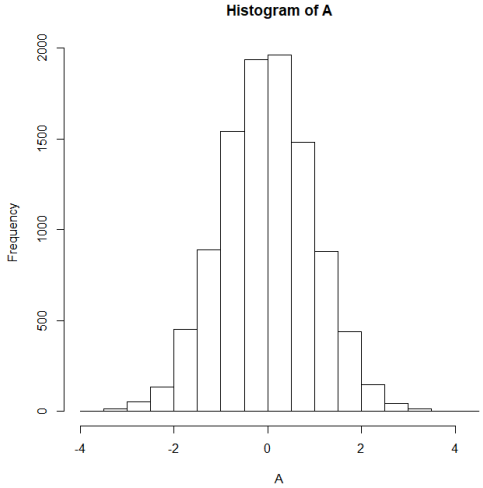




$\alpha=1$
$\beta=1$



$\alpha=5$
$\beta=1$

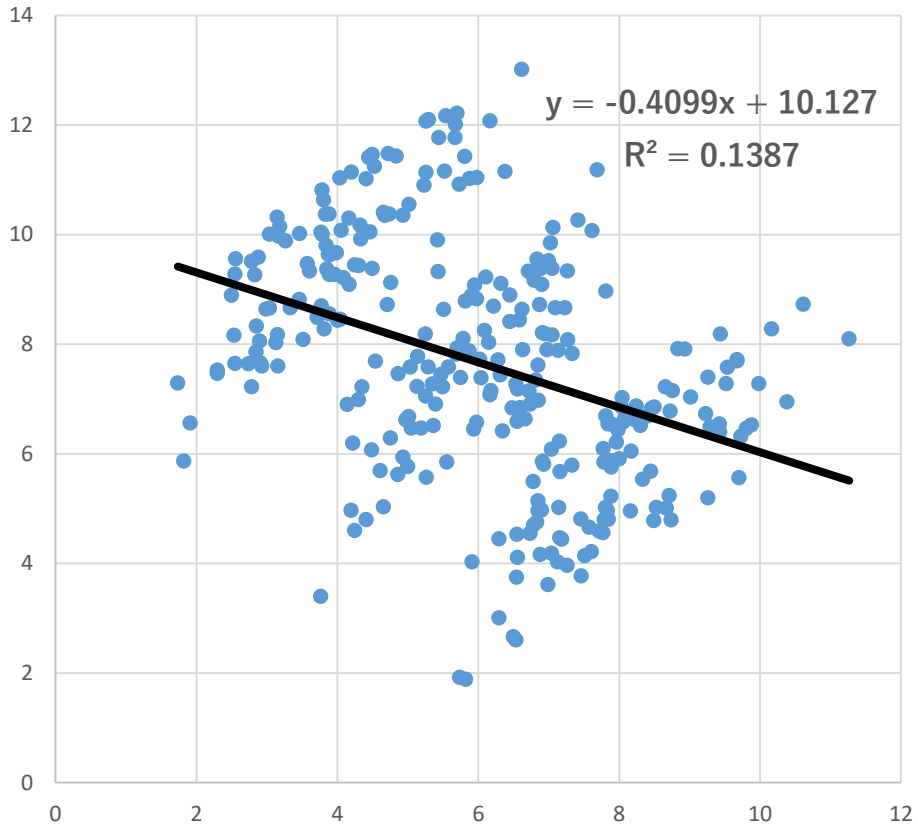


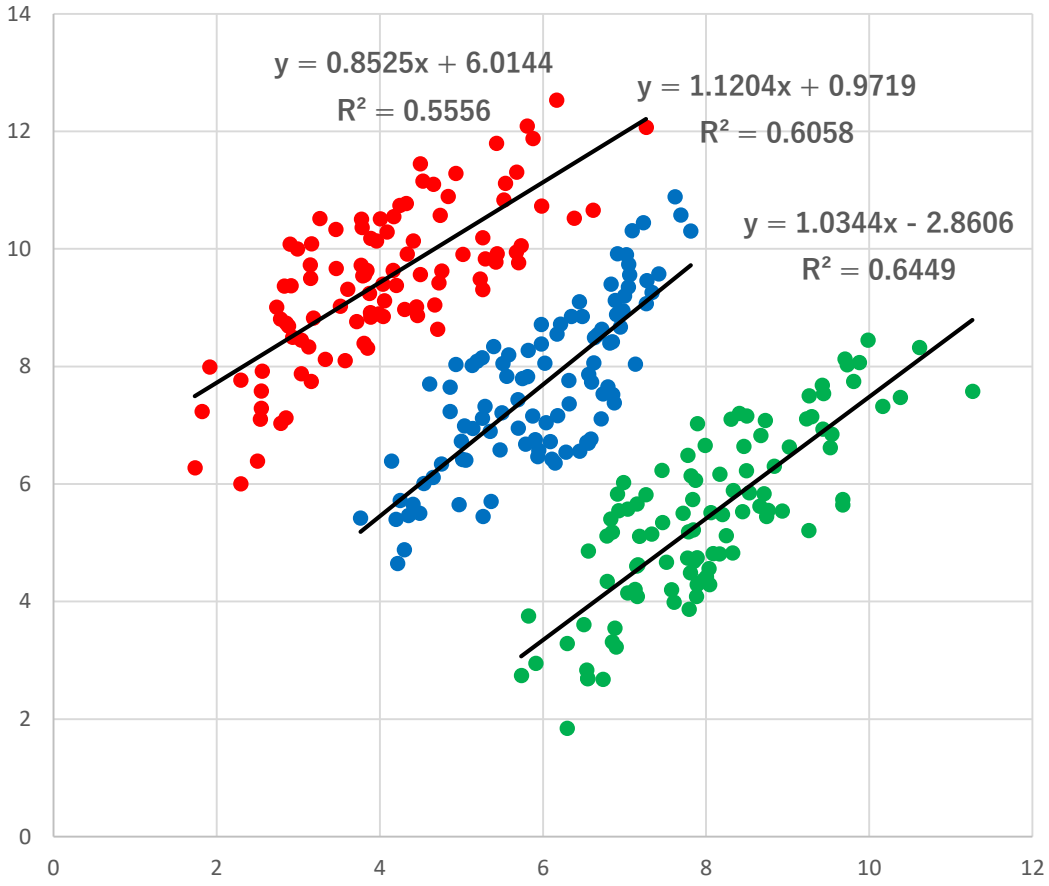
Contents

Part I　Basics
- 統計の基本的な考え方
- 統計学的検定の基本

Part II　Advanced
- 回帰分析の注意点
  - 確率分布
  - 一般化線形モデル
  - 線形混合モデル
  - アンバランスなデータ
  - 0が多いモデル

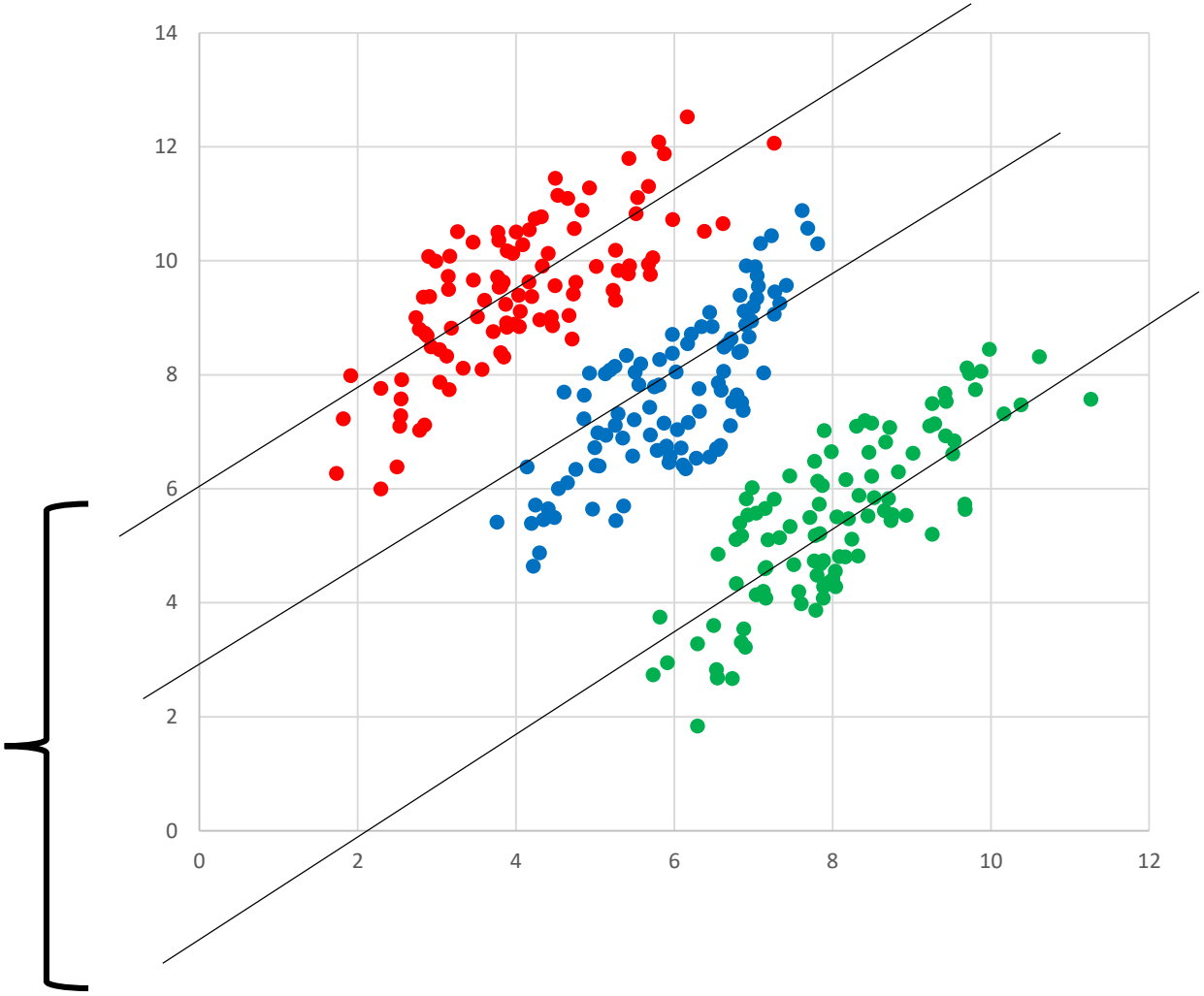
負の相関が見られる




y = 0.8525x + 6.0144
R² = 0.5556
y = 1.1204x + 0.9719
R² = 0.6058
y = 1.0344x - 2.8606
R² = 0.6449
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12

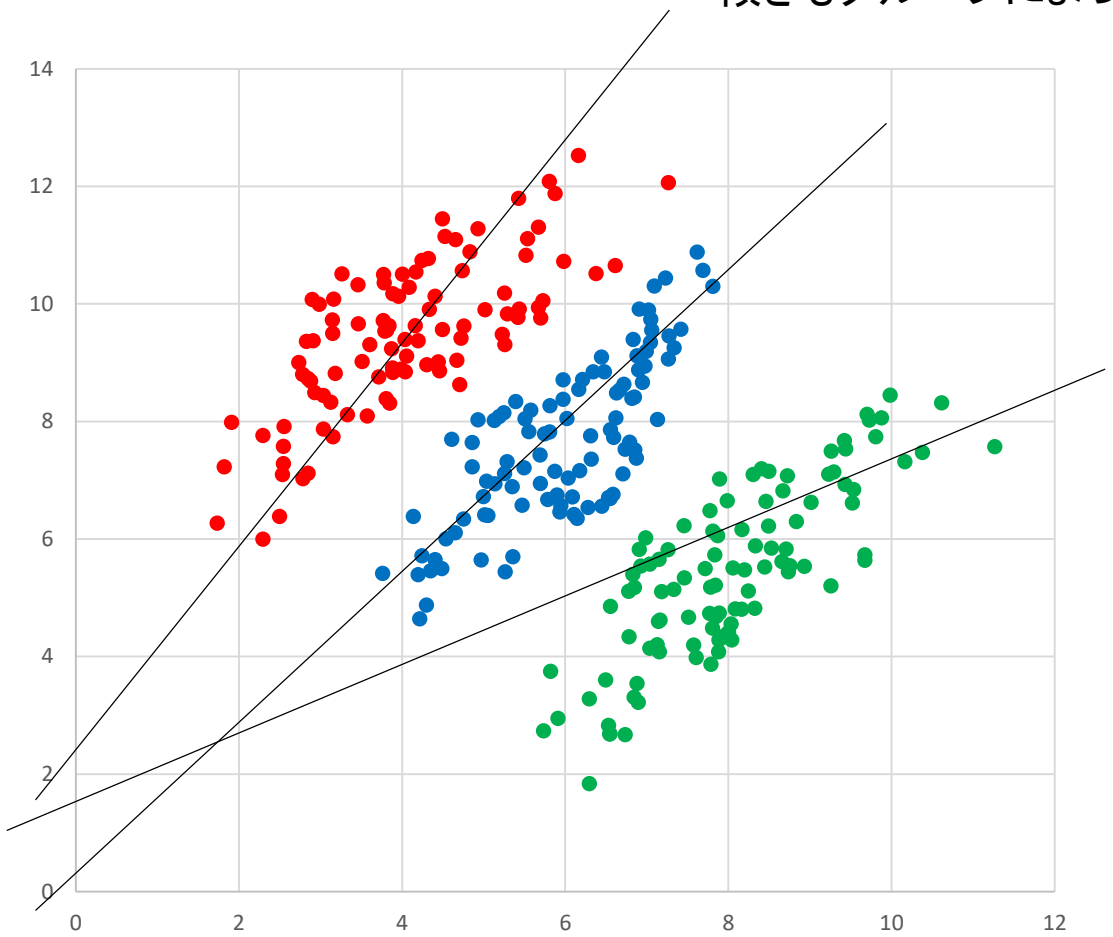
Random Intercept Model

切片がグループによって異なる



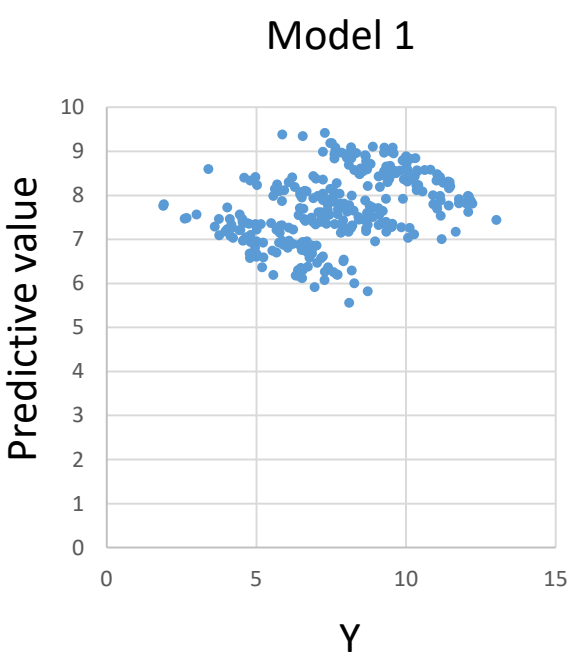
Random Slope and Random Intercept Model

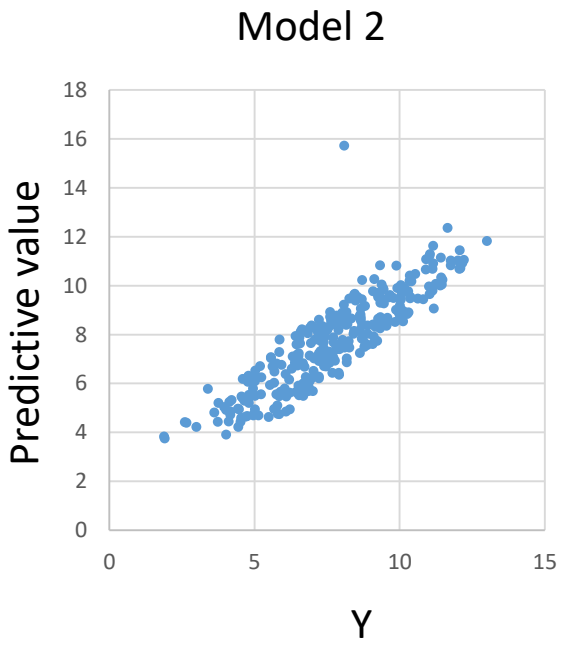
傾きもグループによって異なる

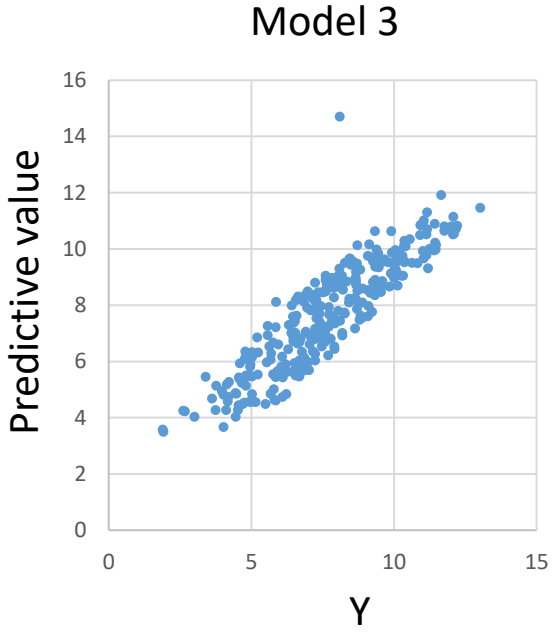


切片がグループによって異なる

		パラメータ	推定値	95% 信頼区間		P-value	AIC
				下限	上限		
Model 0	Regresion Model	(定数)	10.11	9.37	10.86	<0.001	
		X	-0.40	-0.52	-0.29	<0.001	
Model 1	Fixed Effect Model	切片	10.11	9.37	10.86	<0.001	1279.43
		X	-0.40	-0.52	-0.29	<0.001	
Model 2	Random Effect Model	切片	2.60	-6.27	11.48	0.344	876.09
		X	0.84	0.74	0.94	<0.001	
Model 3	Random slope and random intercept model	切片	2.21	-8.93	13.35	0.484	870.85
		X	0.88	0.40	1.35	0.015	







線形混合モデルの特徴

欠損値に強い
反復測定に対応可能
個人の変動を変量効果として処理できる

データシート作成が非常に大変

数理がかなり難しい。

共分散行列の選択にかなりの知識が必要

In what follows, fixed effects (resp. random effects or error terms) are denoted by Greek letters (resp. Alphabet).

- Model 1

  Site, Tooth, and Patient are indexed by i, j, k, respectively.

  - L1: $(\Delta\text{CAL})_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{CAL at Baseline})_{ijk} + \sum_{m=2}^{20} \pi_{2jk}^{(m)}(\text{Tooth surface indexed by } m)_{ijk} + e_{ijk}$
  - L2: $\pi_{0jk} = \beta_{00k} + \sum_{n=2}^{3} \beta_{01k}^{(n)}(\text{Tooth mobility indexed by } n)_{jk} + r_{0jk}$
  - L2: $\pi_{1jk} = \beta_{10k}$
  - L2: $\pi_{2jk}^{(m)} = \beta_{20k}^{(m)}$
  - L3: $\beta_{00k} = \gamma_{000} + \gamma_{001}^{(2)}(\text{Salivary levels of } A.a)_k + \gamma_{002}^{(2)}(\text{Salivary levels of } P.g)_k + u_{00k}$
  - L3: $\beta_{01k}^{(n)} = \gamma_{010}^{(n)}$
  - L3: $\beta_{10k} = \gamma_{100}$
  - L3: $\beta_{20k}^{(m)} = \gamma_{200}^{(m)}$,

  where $e_{ijk} \sim \mathcal{N}(0, \sigma_e^2)$, $r_{0jk} \sim \mathcal{N}(0, \sigma_r^2)$, and $u_{00k} \sim \mathcal{N}(0, \sigma_u^2)$.

Table 1. Multilevel random intercept model for changes in CAL between the baseline and after 24 months (Model 1).

			Coefficient	95% CI		P-value
				Lower	Upper	
Intercept			0.965	0.856	1.074	<0.001
Subject-level explanatory variable						
Salivary levels of *A. a*		<0.00006%	Reference			
		0.00006%<	0.264	0.008	0.519	0.043
Salivary levels of *P. g*		<0.0067%	Reference			
		0.0067%<	0.174	0.026	0.321	0.021
Tooth-level explanatory variable						
Tooth mobility		0	Reference			
		1	0.367	0.285	0.449	<0.001
		2–3	0.840	0.592	1.088	<0.001
Site-level explanatory variable						
CAL at Baseline			-0.436	-0.448	-0.423	<0.001
Mandibular	Anterior	Lingual	Reference			
		Labial	0.122	0.042	0.201	0.003
		Approximal	0.171	0.108	0.234	<0.001
	Premolar	Lingual	0.098	-0.009	0.204	0.073
		Buccal	0.214	0.107	0.321	<0.001
		Approximal	0.251	0.163	0.338	<0.001
	Molar	Lingual	0.428	0.317	0.539	<0.001
		Buccal	0.354	0.243	0.465	<0.001
		Approximal	0.433	0.340	0.527	<0.001
		Distal	0.450	0.342	0.558	<0.001
Maxillary	Anterior	Paratal	-0.117	-0.211	-0.022	0.016
		Labial	-0.020	-0.115	0.075	0.677
		Approximal	0.146	0.066	0.227	0.000
	Premolar	Palatal	0.116	0.009	0.223	0.034
		Buccal	0.241	0.134	0.348	<0.001
		Approximal	0.365	0.277	0.453	<0.001
	Molar	Palatal	0.634	0.520	0.747	<0.001
		Buccal	0.760	0.646	0.873	<0.001
		Approximal	0.701	0.605	0.797	<0.001
		Distal	0.642	0.533	0.751	<0.001

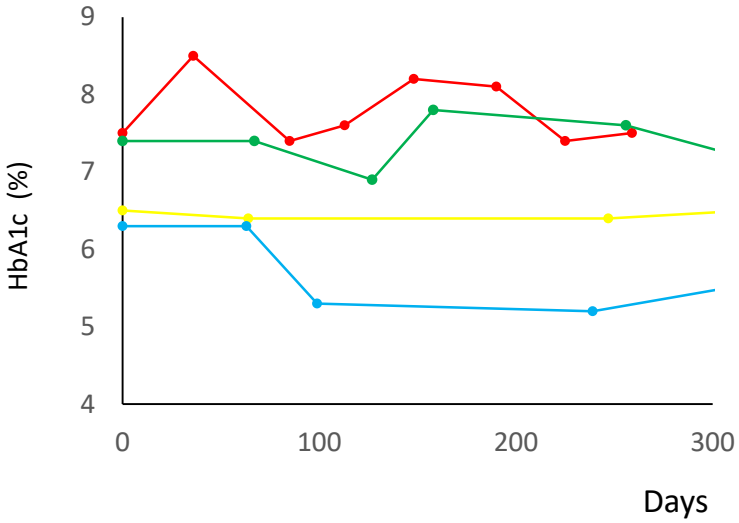
CAL: clinical attachment level; *A. a*: *Aggregatibacter actinomycetemcomitans*; *P. g*: *Porphyromonas gingivalis*

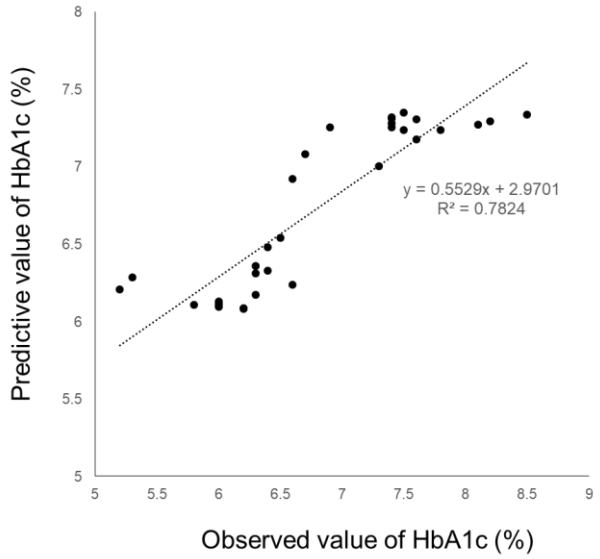
https://doi.org/10.1371/journal.pone.0188670.t001

歯周病菌が有意

動揺度が有意

上顎大臼歯が進行しやすい（特に頬側面）





Model

Subjects, tooth and time are indexed by *ijk*.

$$-\text{ L1: } P_o(\text{HbA1c})_{ijk} = P_o(\mu_{ijk}) = \pi_{0ijk} + \pi_{1jk}(\text{subject})_{ijk} + \varepsilon_{ijk}$$

$$-\text{ L2: } \pi_{0jk}$$

$$= \beta_{00k}^{(m)} + \sum_{m=2}^{3} \beta_{001}^{(m)}(\text{days indexed by } m)_{jk} + \beta_{002}^{(m)}(\text{HbA1c at baseline indexed by } m)_{jk} r_{0jk}$$

$$-\text{ L2: } \pi_{1jk} = \beta_{10k}$$

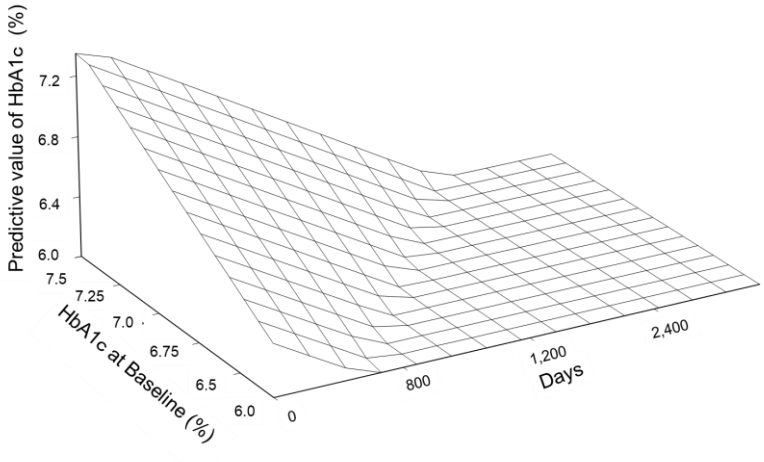
$$\text{where HbA1c} \sim \gamma(\alpha_{ijk}, \tau_{ijk}) \text{ with } \mu_{ijk} = \alpha_{ijk}/\tau_{ijk}$$

$$e_{ijk} \sim N(0, \delta_e^2),\ r_{0jk} \sim N(0, \delta_r^2), \text{ and } u_{01k} \sim N(0, \delta_{u(m)}^2)$$

Fixed effect: days, HbA1c at baseline

Random effect: days

Covariance Type: AR1



Model 2 and 3

Subjects, tooth and tooth surface are indexed by *ijk*.

– L1: $P_o(\text{ICDAS socre})_{ijk} = P_o(\mu_{ijk}) = \pi_{0ijk} + \pi_{1jk}(\text{time})_{ijk} + \varepsilon_{ijk}$

– L2: $\pi_{0jk} = \beta_{00k}^{(m)} + \sum_{m=2}^{3} \beta_{001}^{(m)}(\text{tooth surface indexed by } m)_{jk} \times (\text{Bacterial level})_i + r_{0jk}$

– L2: $\pi_{1jk} = \beta_{10k}$

– L3: $\beta_{00k} = \gamma_{000}$

– L3: $\beta_{01k}^{(m)} = \gamma_{010}^{(m)} + u_{01k,}^{(m)}$

where ICDAS score $\sim P_o(\alpha_{ijk}, \tau_{ijk})$ with $\mu_{ijk} = \alpha_{ijk}/\tau_{ijk}$

$e_{ijk} \sim N(0, \delta_e^2)$, $r_{0jk} \sim N(0, \delta_r^2)$, and $u_{01k} \sim N(0, \delta_{u^{(m)}}^2)$

Model 4

Subjects, tooth and tooth surface are indexed by *ijk*.

– L1: $P_o(\text{ICDAS socre})_{ijk} = P_o(\mu_{ijk}) = \pi_{0ijk} + \pi_{1jk}(\text{time})_{ijk} + \varepsilon_{ijk}$

– L2: $\pi_{0jk} = \beta_{00k}^{(m)} + \sum_{m=2}^{3} \beta_{001}^{(m)}(\text{tooth surface indexed by } m)_{jk} \times \beta_{001}^{(m)}(\text{fissure sealant indexed by } m)_{jk} + r_{0jk}$

– L2: $\pi_{1jk} = \beta_{10k}$

– L3: $\beta_{00k} = \gamma_{000}$

– L3: $\beta_{01k}^{(m)} = \gamma_{010}^{(m)} + u_{01k,}^{(m)}$

where ICDAS score $\sim P_o(\alpha_{ijk}, \tau_{ijk})$ with $\mu_{ijk} = \alpha_{ijk}/\tau_{ijk}$

$e_{ijk} \sim N(0, \delta_e^2)$, $r_{0jk} \sim N(0, \delta_r^2)$, and $u_{01k} \sim N(0, \delta_{u^{(m)}}^2)$

		S. mutans			LB	
		Coefficient	P-value		Coefficient	P-value
Mandibular Maxilla						
Intercept		0.013	0.002		0.054	<0.001
Mandibular	*S. mutans* -	Reference		LB -	Reference	
	S. mutans +	0.095	<0.001	LB +	0.081	<0.001
Maxilla	*S. mutans* -	0.022	0.006	LB -	0.015	0.071
	S. mutans +	0.122	<0.001	LB +	0.12	<0.001
Tooth Type						
Intercept		0.027	<0.001		0.043	<0.001
Anterior	*S. mutans* -	Reference		LB -	Reference	
	S. mutans +	0.055	<0.001	LB+	0.06	<0.001
Premolar	*S. mutans* -	-0.011	0.237	LB -	-0.004	0.645
	S. mutans +	0.061	<0.001	LB +	0.091	<0.001
Molar	*S. mutans* -	-0.002	0.876	LB -	0.069	<0.001
	S. mutans +	0.194	<0.001	LB +	0.216	<0.001
Tooth Surface						
Intercept		0.001	0.767		0.019	<0.001
Lingual	*S. mutans* -	Reference		LB -	Reference	
	S. mutans +	0.054	<0.001	LB +	0.056	<0.001
Buccal	*S. mutans* -	0.06	<0.001	LB -	0.071	<0.001
	S. mutans +	0.204	<0.001	LB +	0.271	<0.001
Approximal	*S. mutans* -	0.005	0.069	LB -	0.001	0.95
	S. mutans +	0.05	<0.001	LB +	0.048	<0.001
Occlusal	*S. mutans* -	0.054	0.002	LB -	0.222	<0.001
	S. mutans +	0.418	<0.001	LB +	0.46	<0.001

		Coefficient	P-value
Mandibular Maxilla			
Intercept		0.366	<0.001
Mandibular	Sealant -	Reference	
	Sealant +	-0.098	0.062
Maxilla	Sealant -	-0.033	0.782
	Sealant +	-0.345	<0.001
Tooth Type			
Intercept		0.125	<0.001
Premolar	Sealant -	Reference	
	Sealant +	-0.071	0.399
Molar	Sealant -	0.233	<0.001
	Sealant +	0.366	0.096

Contents

Part I　Basics
- 統計の基本的な考え方
- 統計学的検定の基本

Part II　Advanced
- 回帰分析の注意点
  - 確率分布
  - 一般化線形モデル
  - 線形混合モデル
  - アンバランスなデータ
  - 0が多いモデル

Unbalance なデータ

0： 10

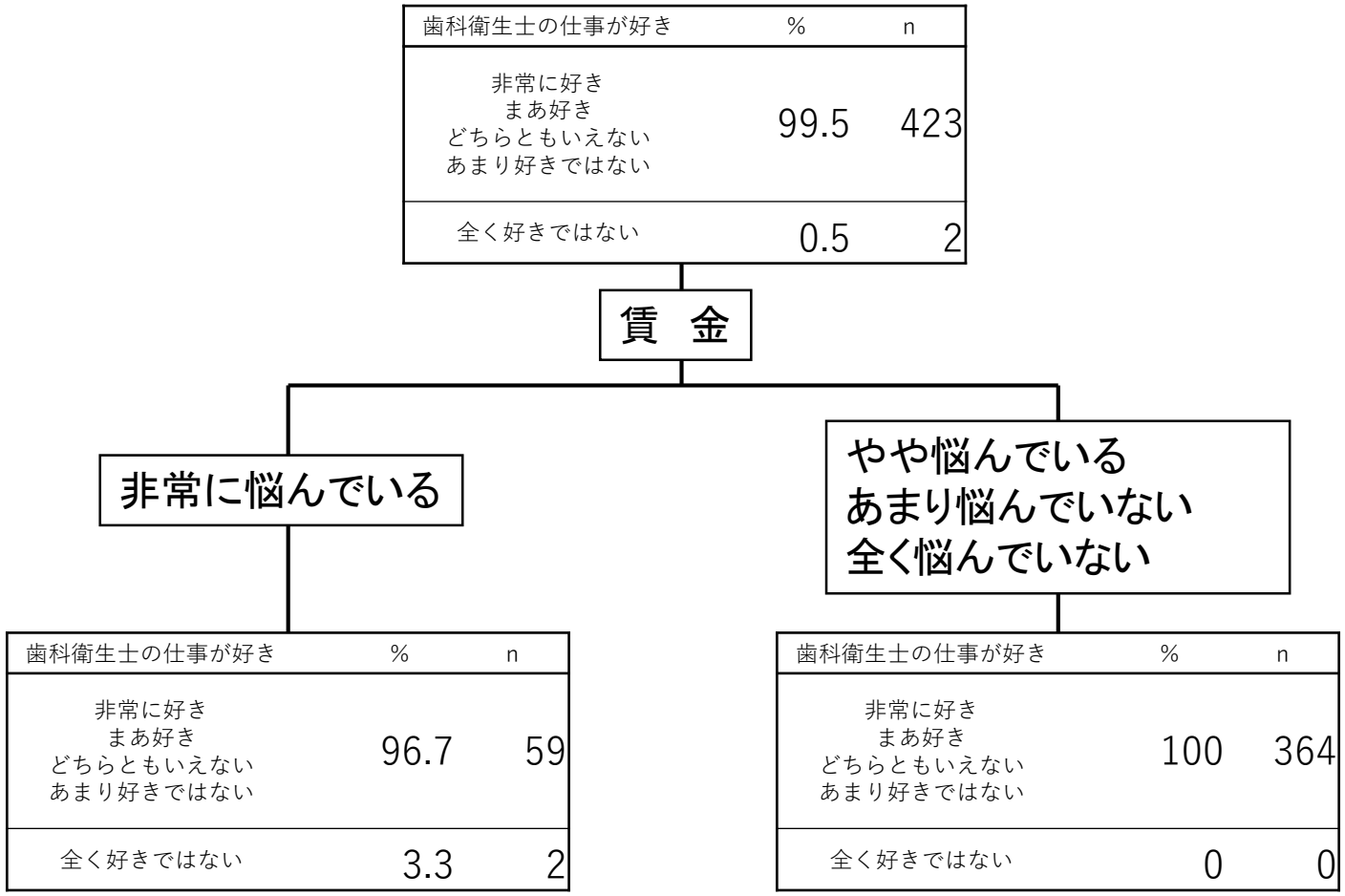
1： 84

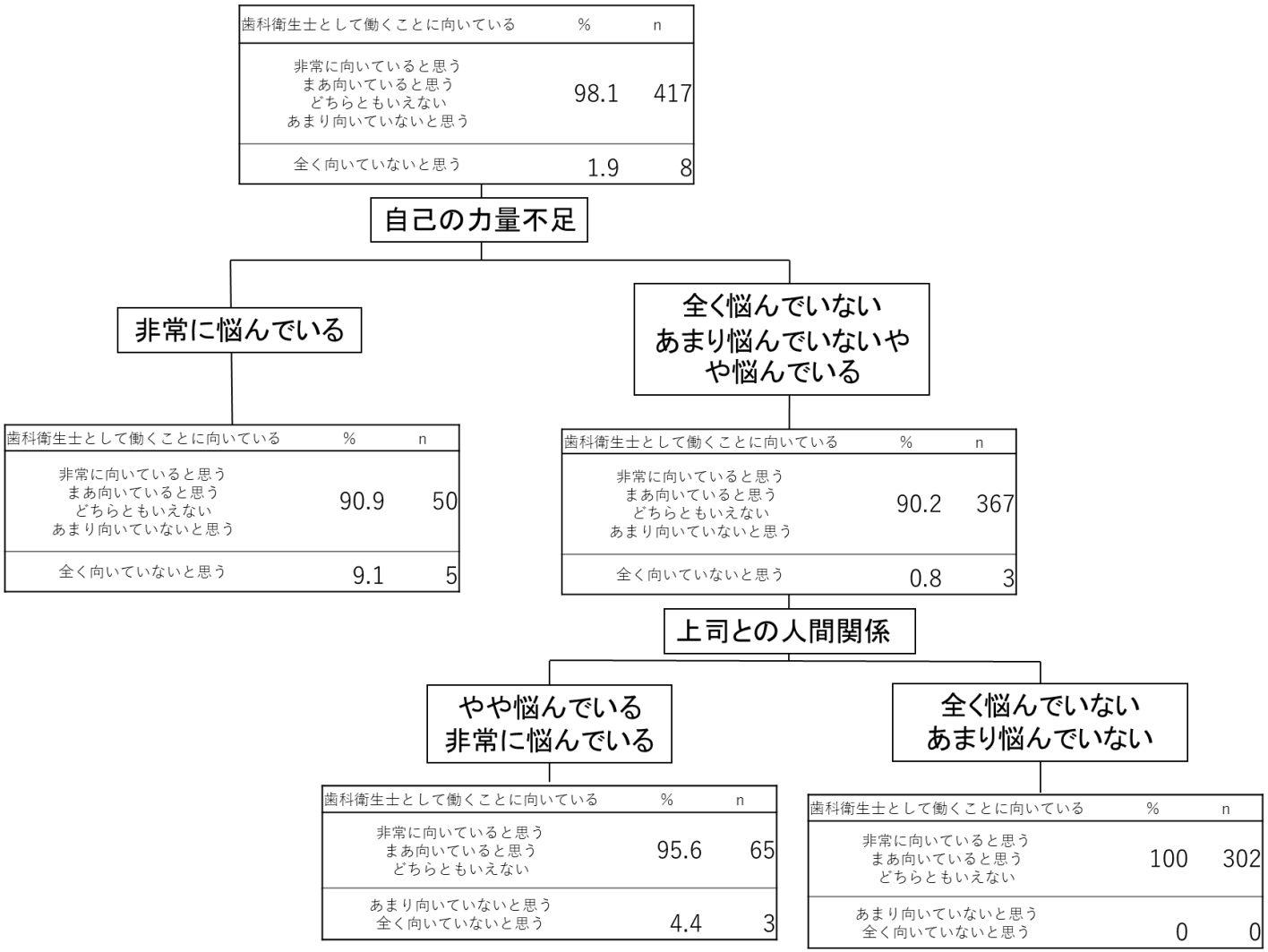
$$ln\left(\frac{pi}{1-pi}\right) = \alpha + \beta_1 x_{1,i} + \cdots\cdots$$

	Exp(B)	95% 信頼区間		有意確率
		下限	上限	
Y	13.107	1.922	89.394	.009
定数	1.280			.697

		予測値	
		0	1
観測値	0	0	10
	1	0	84

観測値0をすべて1と誤った予測をしても89.3%の確率で正しい予測


歯科衛生士の仕事が好き
%
n
非常に好き
まあ好き
どちらともいえない
あまり好きではない
99.5
423
全く好きではない
0.5
2
賃　金
非常に悩んでいる
やや悩んでいる
あまり悩んでいない
全く悩んでいない
歯科衛生士の仕事が好き
%
n
非常に好き
まあ好き
どちらともいえない
あまり好きではない
96.7
59
全く好きではない
3.3
2
歯科衛生士の仕事が好き
%
n
非常に好き
まあ好き
どちらともいえない
あまり好きではない
100
364
全く好きではない
0
0


歯科衛生士として働くことに向いている % n
非常に向いていると思う
まあ向いていると思う
どちらともいえない
あまり向いていないと思う
98.1 417
全く向いていないと思う
1.9 8
自己の力量不足
非常に悩んでいる
全く悩んでいない
あまり悩んでいないや
や悩んでいる
歯科衛生士として働くことに向いている % n
非常に向いていると思う
まあ向いていると思う
どちらともいえない
あまり向いていないと思う
90.9 50
全く向いていないと思う
9.1 5
歯科衛生士として働くことに向いている % n
非常に向いていると思う
まあ向いていると思う
どちらともいえない
あまり向いていないと思う
90.2 367
全く向いていないと思う
0.8 3
上司との人間関係
やや悩んでいる
非常に悩んでいる
全く悩んでいない
あまり悩んでいない
歯科衛生士として働くことに向いている % n
非常に向いていると思う
まあ向いていると思う
どちらともいえない
95.6 65
あまり向いていないと思う
全く向いていないと思う
4.4 3
歯科衛生士として働くことに向いている % n
非常に向いていると思う
まあ向いていると思う
どちらともいえない
100 302
あまり向いていないと思う
全く向いていないと思う
0 0

Unbalanced dataへの対応

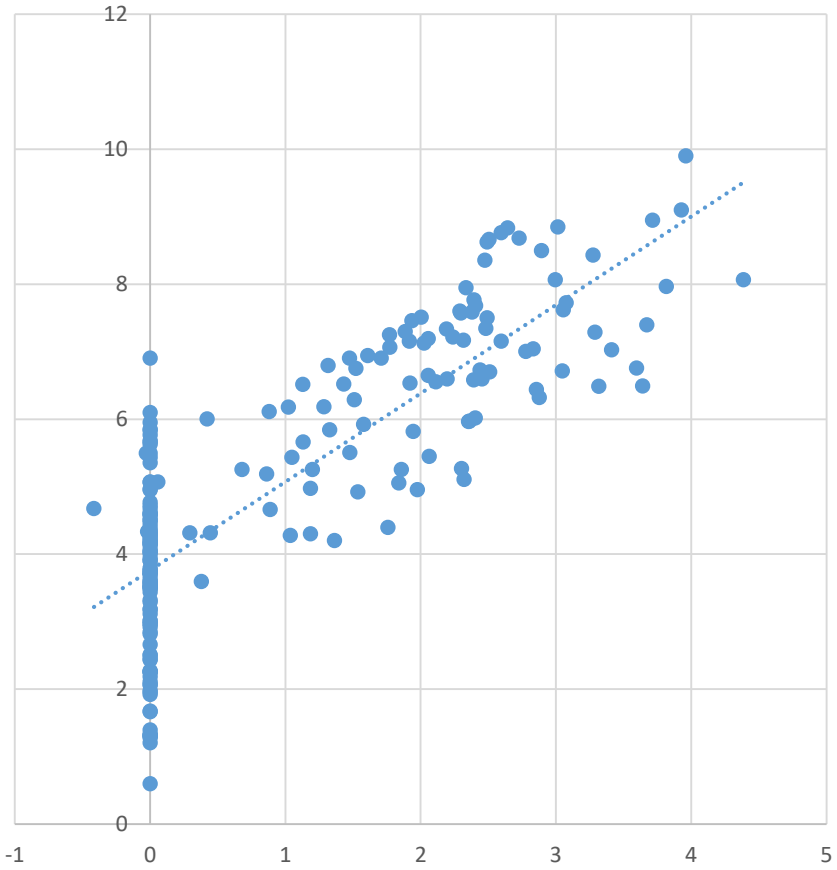
- サンプル数が多いとき　比率1:1になるようにリサンプリングを行いモデル作成
- （データマイニングの考え方）

- 通常は決定木分析により検出したい特性に至るルールを見つける。

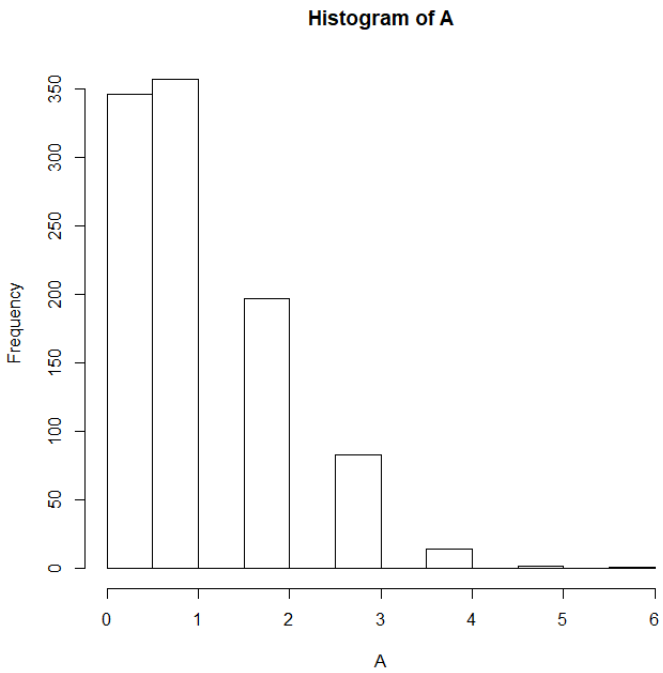
Contents

Part I　Basics
- 統計の基本的な考え方
- 統計学的検定の基本

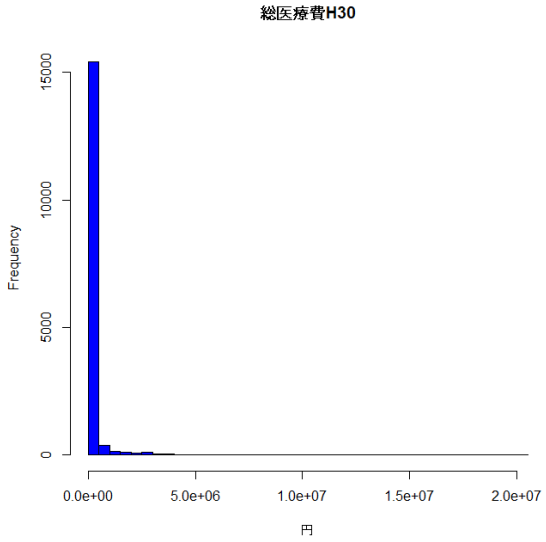
Part II　Advanced
- 回帰分析の注意点
  - 確率分布
  - 一般化線形モデル
  - 線形混合モデル
  - アンバランスなデータ
  - 0が多いモデル

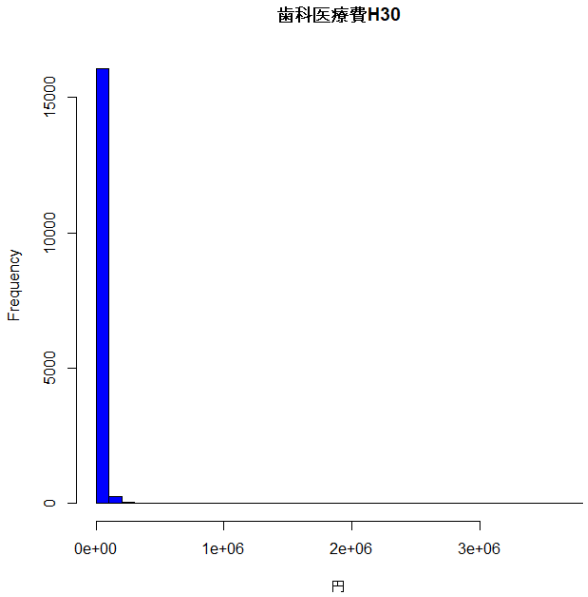


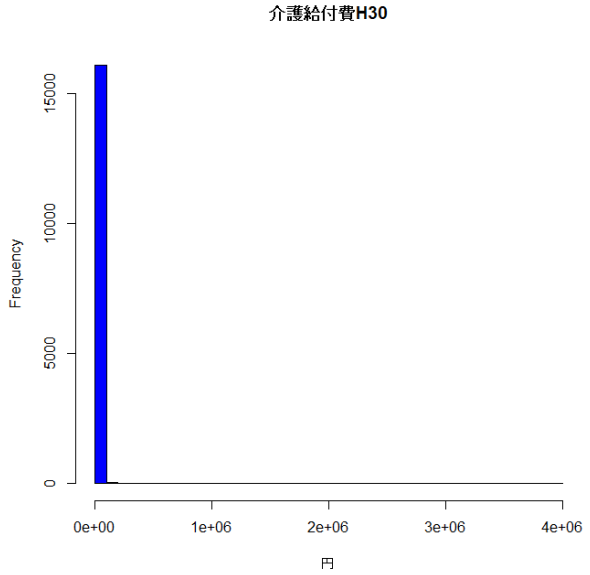
この回帰直線は意味があるでしょうか？

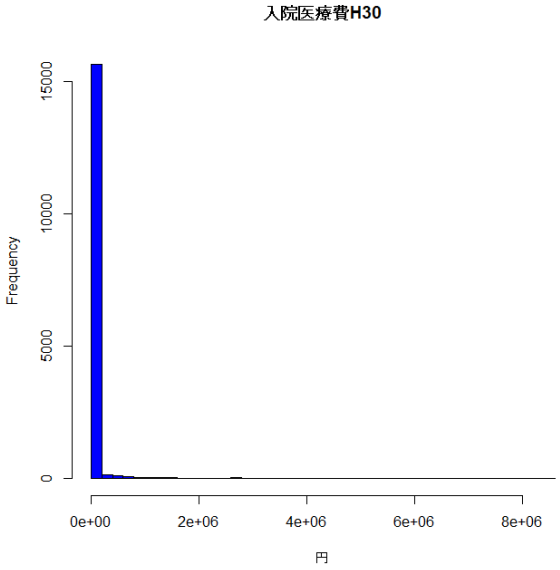


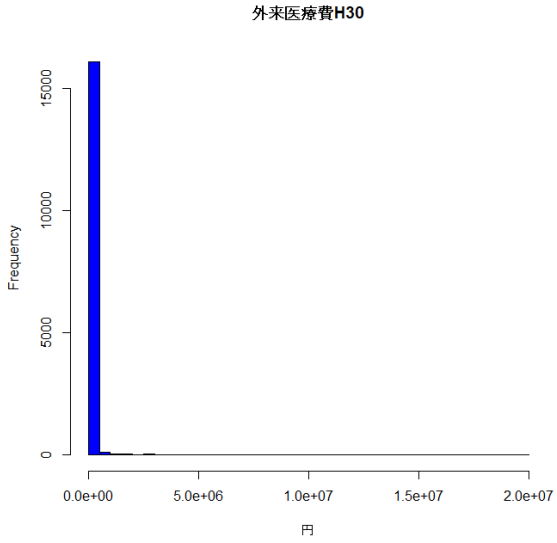
この程度の0であれば大丈夫
（ポアソン分布から発生させた乱数）

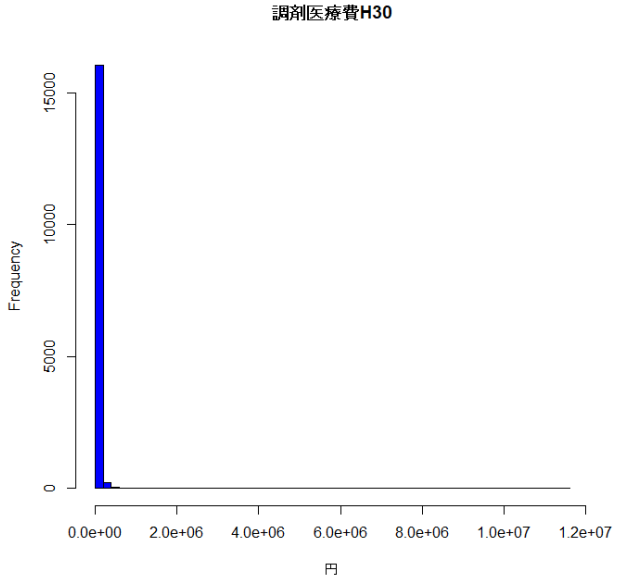

総医療費H30
Frequency
0
5000
10000
15000
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
円


歯科医療費H30
Frequency
0
5000
10000
15000
0e+00
1e+06
2e+06
3e+06
円

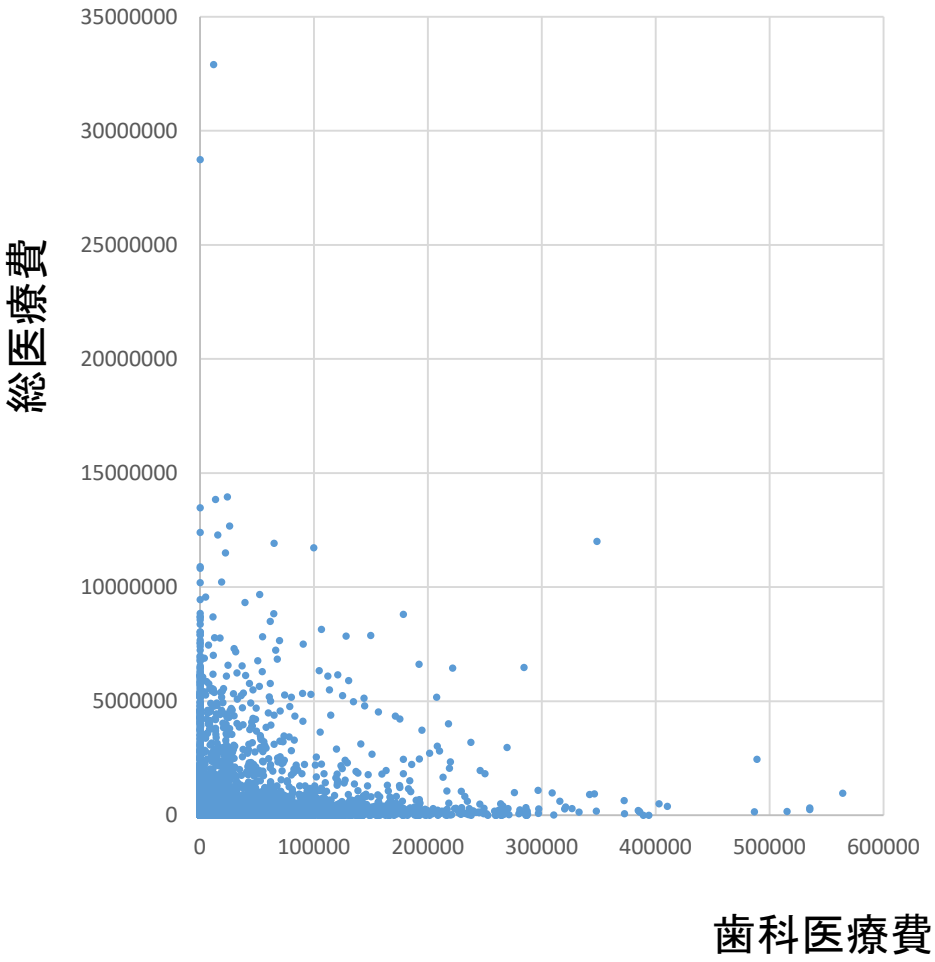

介護給付費H30
Frequency
0
5000
10000
15000
0e+00
1e+06
2e+06
3e+06
4e+06
円


入院医療費H30
Frequency
0
5000
10000
15000
0e+00
2e+06
4e+06
6e+06
8e+06
円

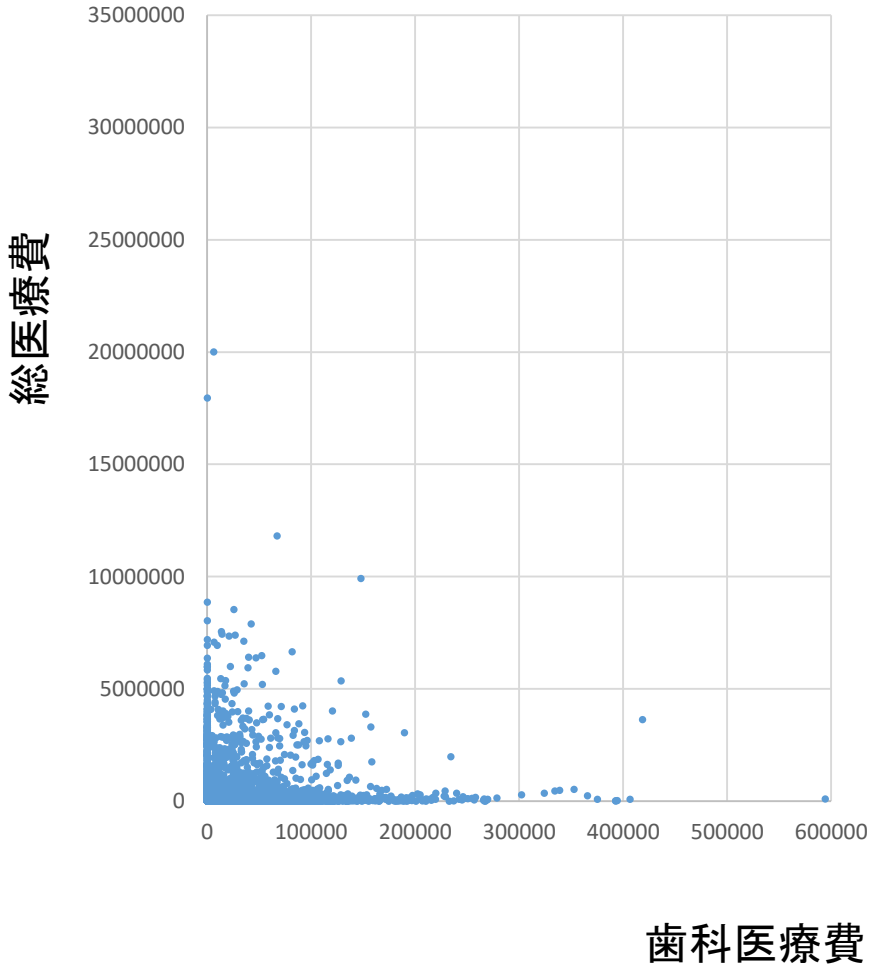

外来医療費H30
Frequency
0
5000
10000
15000
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
円


調剤医療費H30
Frequency
0
5000
10000
15000
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
円

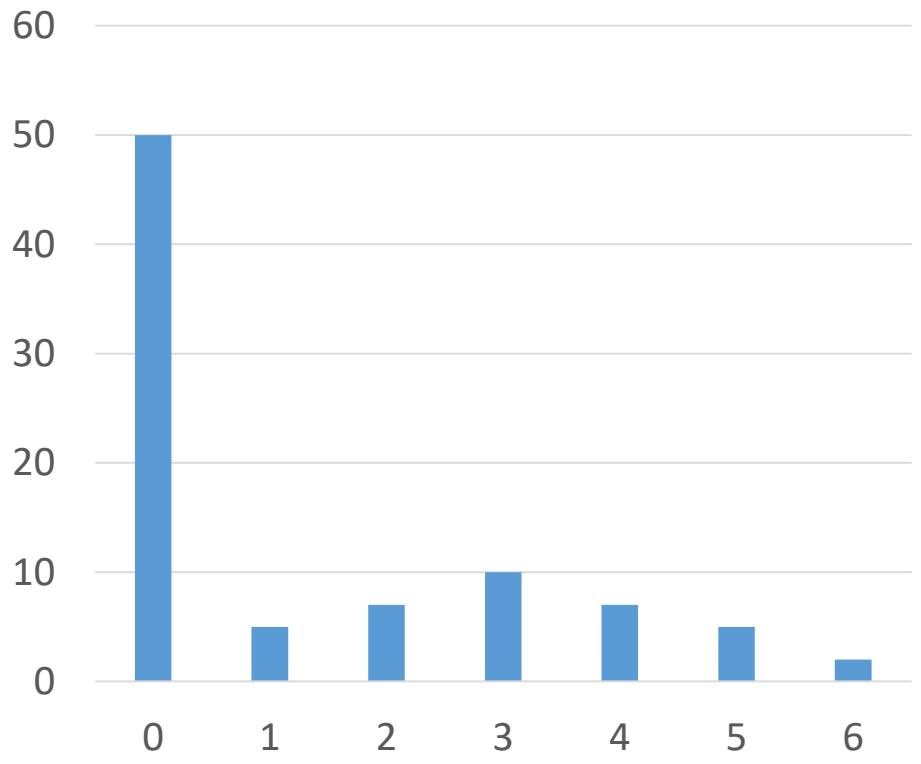
H29


総医療費
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
0
100000
200000
300000
400000
500000
600000
歯科医療費

H30


総医療費
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
0
100000
200000
300000
400000
500000
600000
歯科医療費

$$\mathrm{P}(y_i|\theta,\lambda) = \begin{cases} (1-\theta) + P_o(0|\lambda) \\ \theta Poisson(y_i|\lambda) \end{cases}$$



Hurdle Model, Zero-infrated Modelの利用
H29のデータによるH30の予測分析

		Gebralized linear model				Hurdle model				Zero-inflation model	
		Contenuious varable				Count model				Count model	
						Truncated poisson		Truncated negative binominal		Poisson	Negative binominal
		Poisson	Negative binomnal		Link	log				log	
Intercept	Estimate	10.260	10.070	Intercept	Estimate	0.629	0.629	0.386	3.955	4.354	0.248
	P-value	<0.001	<0.001		P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
歯科医療費	Estimate	3.05E-06	3.66E-06	歯科医療費	Estimate	0.0002	0.000	-0.002	0.094	0.003	0.043
	P-value	<0.001	<0.001		P-value	0.987	0.987	0.932	0.932	0.798	0.021
年齢	Estimate	0.031	0.033	年齢	Estimate	0.007	0.007	0.008	0.025	0.007	0.011
	P-value	<0.001	<0.001		P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
性別	Estimate	-0.290	-0.233	性別	Estimate	-0.164	-0.164	-0.185	-0.361	-0.161	-0.137
	P-value	<0.001	<0.001		P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
						Hurdle model				Zero hurdle model	
						Binomial	Geometric	Binomial	Censored negative binominal	Binomial	
					Link	logit link	log link	logit link	log link	logit link	logit link
				Intercept	Estimate	-1.093	-1.093	-1.093	-1.061	0.916	0.449
					P-value	0.009	0.009	<0.001	<0.001	<0.001	<0.001
				歯科医療費	Estimate	1.061	1.061	1.061	0.565	-1.440	-12.886
					P-value	<0.001	<0.001	<0.001	<0.001	<0.001	0.844
				年齢	Estimate	0.028	0.028	0.028	0.018	-0.027	-0.028
					P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
				性別	Estimate	0.202	0.202	0.202	0.090	-0.437	-0.703
					P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
AIC		7421761777	292391	AIC		57099	7264328	51732	51669	57110	51962
Theta			0.1305	Theta				0.4272			2.5124

数理統計学

実験計画法と検定の多重性

調査票の設計と分析

サンプルサイズの設計

生存分析

項目反応理論

まとめ

- 統計でできることは　有意な関連、有意な差
- 統計は標本から母集団を推測するもの
- 変数の型で統計方法は決まる
- バイアスのないサンプリングは統計解析以上に重要
- 回帰分析は変数の分布が重要。これを間違えると誤った結論を導きやすい