

確率・統計

第9回 標本調査と実験計画

兵庫県立大学 社会情報科学部

川嶋宏彰

kawashima@sis.u-hyogo.ac.jp

(発展) 繰り返しのない二元配置分散分析

2

- 母集団の平均に差があるか調べたい → 分散分析
 - 複数の要因を考えるときは？

肥料と設定温度によるトマトの重さの違い

	温度A	温度B	温度C	温度D
肥料1	8	5	7	10
肥料2	3	1	3	7
肥料3	6	3	2	5

肥料の違いだけでなく
設定温度の影響も知りたい

→ 二元配置分散分析

左の例は「繰り返しのない」
二元配置分散分析

→ 実際には各セルに複数の観測値
「繰り返しのある」二元配置分散分析

交互作用 (二つの因子が絡み合って
起こす効果) を考えることができる
(例：温度Dにしつつ肥料1を使うと
効果が大きいなど)

本日の講義内容

3

- テキスト
 - 「統計学入門」12.1節, 12.5節, 1.3節 (少しだけ)
- 講義トピック
 - 社会調査と抽出方法
 - 実験と調査観察 (介入があるかないか)
 - 代表的バイアス (選択バイアス, 情報バイアス, 交絡)
 - エビデンスレベル
 - ランダム化比較試験, コホート研究, ケース・コントロール研究
 - フィッシャーの三原則
- 演習

調査の方法

4

- 正しい調査のためにはそれなりの「作法」がある
 - 課題調査の設定
 - 現状把握型
 - 仮説検証型
 - 集計・解析方法
 - 調査対象の設定（標本の取り方）
 - 調査時期
 - 調査方法

代表的な抽出方法を
覚えておこう

調査課題の設定

5

- 調査目的
 - 現状把握型
 - 「兵庫県・大阪府民の食生活」はどうなっているだろうか
 - 仮説検証型
 - 仮説「兵庫県・大阪府民は他県と食文化が異なる」を検証したい
- 調査課題の設定と調査項目
 - 現状把握型
 - 調査課題：「大阪・兵庫県民の食事メニューを調べる」
 - 調査項目：「昼食メニューは次のうちどれですか？」
 - 仮説検証型
 - 調査課題：「大阪・兵庫の人はたこ焼きや明石焼きが好きな人が多い」
 - 調査項目：「実家は大阪・兵庫ですか？」「たこ焼き器を持っていますか？」

調査対象の設定と無作為抽出

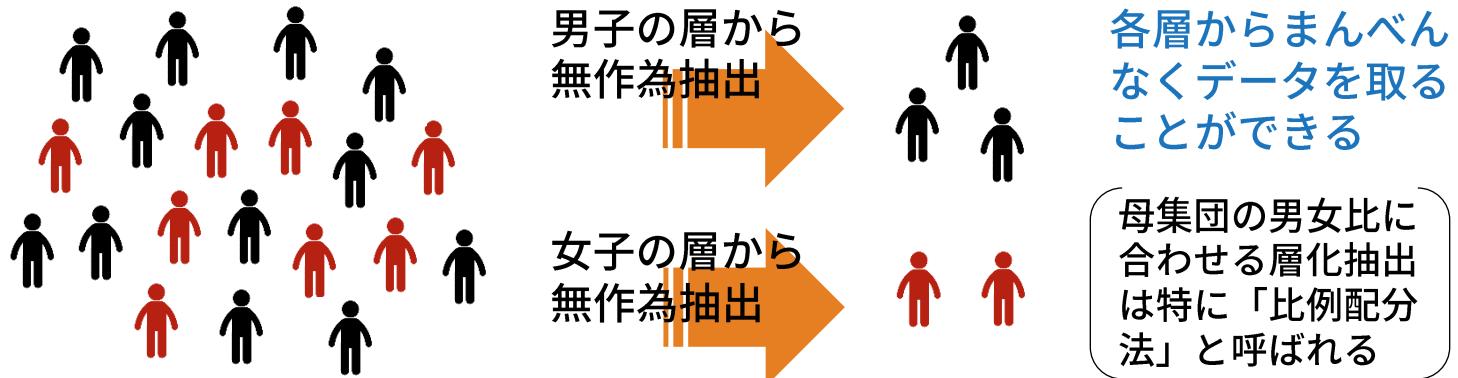
6

- 「何を母集団と考えるか」 「効率・実現可能性」で抽出法は様々
 - **単純無作為抽出**
 - 母集団(電話帳, 住民台帳など)に通し番号をつけ, 亂数で選ぶ
- 社会調査でよく用いる無作為抽出 (下線は特に重要: 組み合わせることも)
 - 多段抽出: 母集団からの抽出を何段階かに分けて行う方法
 - (例) 全国調査時に, 最初に調査する都道府県を無作為に選び, 都道府県の中から市町村を選び, 最後に市町村の中から人を選ぶ
 - 系統抽出: 母集団の名簿に番号をつけ一定間隔で標本を選ぶ
 - 層化抽出 (層別抽出): 既知の母集団の状況(男女比, 年齢, 地域)に合わせて, あらかじめ層(グループ)に分けて置き, 各層から無作為に抽出
 - 集落抽出 (クラスタ抽出): 母集団を小集団(クラスタ)に分けたうえで, まずクラスタを無作為に選び, 選ばれたクラスタ内のメンバ全員を調査

層化抽出と集落抽出

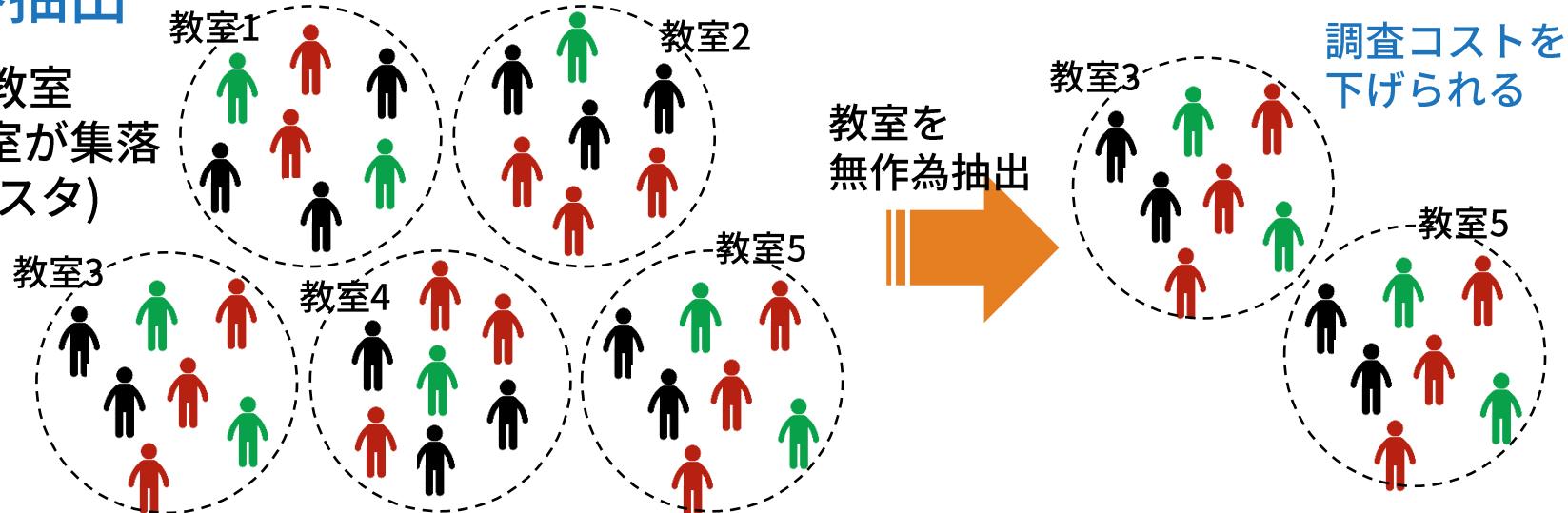
・ 層化抽出

例: 男女2層



・ 集落抽出

例: 5教室
各教室が集落(クラスタ)



(例) e-Stat の家計調査は?

8

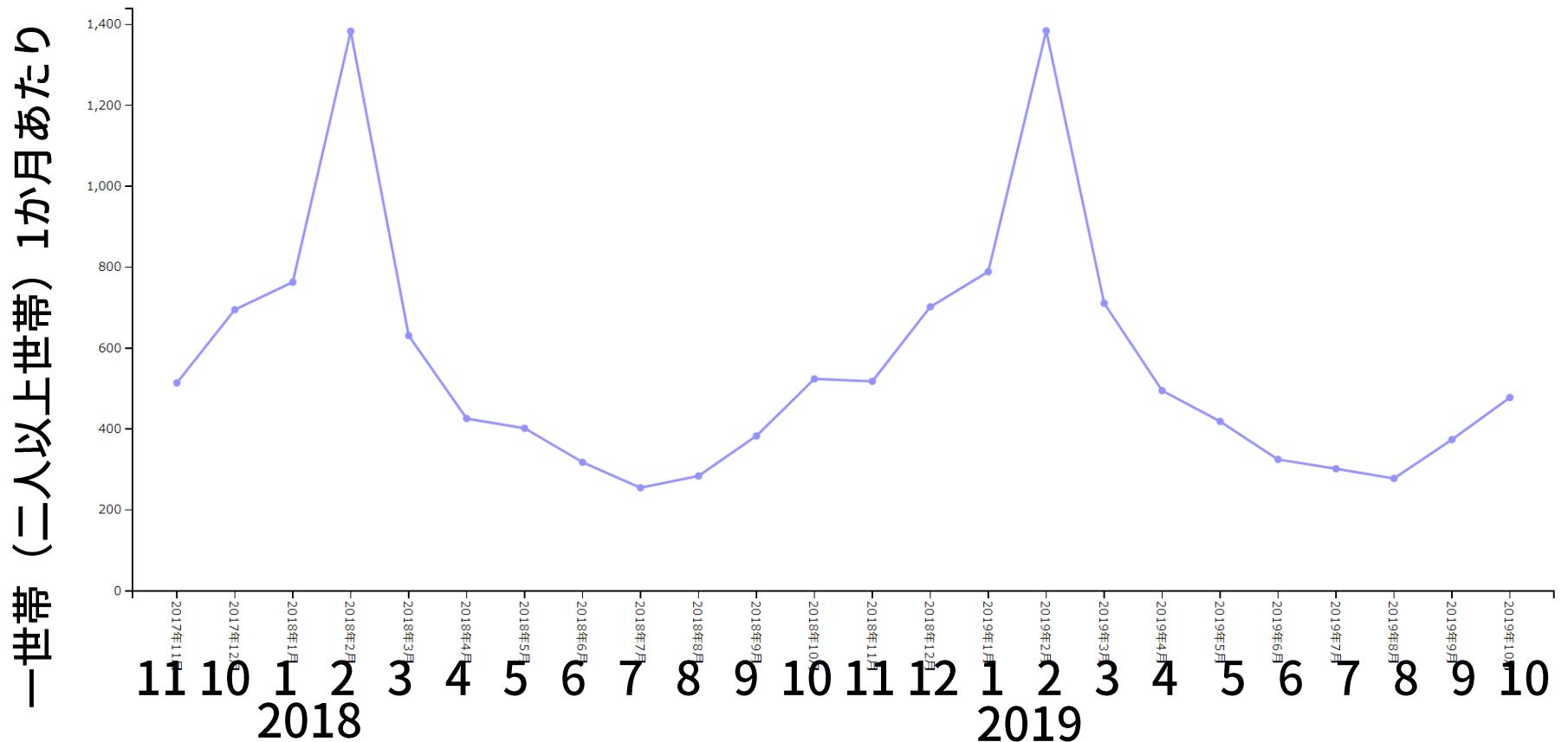
- ・都市別ランキングなどで非常によく用いられる
 - ・e-Stat → 分野から探す → 家計調査 → データベース → 家計収支編-二人以上の世帯-月次 → 品目分類(最新の年のもの)(総数:金額)-DB
 - ・表示項目選択で「品目分類」の数字をクリック → 全解除して必要なものだけ選択 → 表示を更新 → CSVをダウンロード
 - ・グラフを「折れ線グラフ」, 基準軸を「行」とすれば変化も確認しやすい

データベース	件数 更新日	ファイル	件数 更新日
 新着	98件 2019-12-06	 新着	21,268件 2019-12-06

(参考) チヨコの購入金額 (2017.11-2019.10)

9

- ・寒いときの方がよく売れる（2月は特に多い）



全国9000世帯・・・調査方法は？

10

- ・ iマークから詳細ページへ
- ・ <http://www.stat.go.jp/data/kakei/>

家計調査の概要、結果等

調査の概要

- [調査の目的](#)
- [調査事項](#)
- [調査の沿革](#)
- [調査票](#)
- [調査の根拠法令](#)
- [調査の時期](#)
- [調査の対象](#)
- [調査の方法](#)
- [抽出方法](#)
- [標本設計の概要](#)
- [家計調査調査市町村一覧 \(PDF : 272KB\)](#)

調査の結果

家計調査 i

一覧形式で表示

家計調査は、統計理論に基づき選定された全国約9千世帯を対象として、家計の収入・支出、貯蓄・負債などを毎月調査しています。

家計調査の結果は、我が国の景気動向の把握、生活保護基準の検討などの基礎資料として利用のほか、地方公共団体、民間の会社などでも利用されています。

二人以上の世帯の結果は、主に、地域・世帯属性ごとに1世帯当たり1か月間の収支金額にまとめ毎月公表、単身世帯及び総世帯の家計収支に関する結果並びに二人以上の世帯の貯蓄・負債に関する結果を四半期ごとに公表しています。

3 調査世帯の選定

家計調査は標本調査であり、層化3段抽出法（第1段—市町村、第2段—単位区、第3段—世帯）により世帯を選定している。選定にあたっては特定の世帯が統けて調査の対象にならないように配慮している。市町村の抽出の仕方は次のとおりである。都道府県庁所在市及び政令指定都市については各市を1層とし52層に分けた。その他の人口5万以上の市については直近の国勢調査の結果に基づき、地方、都市階級に分けた後、

- (1) 人口集中地区人口比率
- (2) 人口増減率
- (3) 産業的特色
- (4) 世帯主の年齢構成

を考慮して74層に分けた。また、人口5万未満の市及び町村は、地方で分けた後、(1)地理的位置（海沿い、山地等）、(2)世帯主の年齢構成を用いて、計42層に分けた。このようにして分けられた全国計168層の各層から1市町村ずつ抽出した。

調査世帯数の割当て

地域	調査市町村数	二人以上の調査世帯数	単身調査世帯数
全国	168	8,076	673
都道府県庁所在市及び大都市	52	5,472	456
人口5万以上の市（上記の市を除く）	74	2,100	175
人口5万未満の市及び町村	42	504	42

層化3段抽出法 (市町村→単位区→世帯)

県庁所在市・政令指定都市はすべて(52層)
他の市町村は「年齢構成」などいくつかの特色に基づき
人口5万以上の市を74層、5万未満を42層へグループ化
→ (1段目) 各層から市町村を抽出 → (2段目) 単位区を抽出
→ (3段目) さらに各単位区から6世帯を抽出

家計調査における中・小都市の層の例

11

地方	都市階級	層番号	層化基準	二人以上の世帯の調査対象世帯数	調整係数	対象市町村数	層に含まれる市	・下線はH30標本改正時の調査市
近畿 中都市 近畿: 5層	中	1	人口集中地区人口比率92.5%未満	335,109	18.7	4	(27大阪府) (28兵庫県)	202岸和田市 201 姫路市 210加古川市
		2	人口集中地区人口比率92.5%～96.1%未満	335,675	18.7	4	(28兵庫県)	203明石市 217 川西市 204 西宮市 214宝塚市
		3	人口集中地区人口比率96.1%以上、 65歳以上世帯数比率36.3%未満	352,531	19.6	4	(27大阪府) (28兵庫県)	203 豊中市 207伊丹市 205吹田市 211茨木市
		4	人口集中地区人口比率96.1%以上、 65歳以上世帯数比率36.3%以上、 人口増減率-0.9%未満	372,543	20.8	4	(26京都府) (27大阪府)	204宇治市 207高槻市 212八尾市 227 東大阪市
		5	人口集中地区人口比率96.1%以上、 65歳以上世帯数比率36.3%以上、 人口増減率-0.9%以上	313,197	17.5	3	(27大阪府) (28兵庫県)	210枚方市 202尼崎市 215寝屋川市
小都市A (5万人以上) 近畿: 6層 (うち3層→)	小A	1	65歳以上世帯数比率33.1%未満	228,813	19.1	10	(25滋賀県) (26京都府) (27大阪府) (28兵庫県) (29奈良県) (30和歌山県)	206草津市 211湖南市 214 木津川市 206泉大津市 219三田市 210香芝市 209岩出市 207守山市 208栗東市 208貝塚市
		2	65歳以上世帯数比率33.1%～36.2%未満	284,198	23.8	11	(25滋賀県) (26京都府) (27大阪府)	202彦根市 211京田辺市 213泉佐野市 224摂津市 229四條畷市 203長浜市 213東近江市 218大東市 225高石市 226藤井寺市
		3	65歳以上世帯数比率36.2%～40.8%未満、 第2次産業就業者数比率22.5%未満	222,108	18.6	10	(26京都府) (27大阪府) (28兵庫県) (29奈良県)	210八幡市 204池田市 206芦屋市 204天理市 209生駒市 231大阪狭山市 232阪南市 222橋本市 223御影市 222伊丹市

- 調査の時期も影響するので十分に注意が必要
 - 調査結果への影響例
 - 選挙への関心
 - 調査された時期の政局に左右
 - 食品安全問題への関心
 - その前にどんな食品事故が生じたかに左右
 - 鍋物が好きか
 - 8月と1月では結果が変わる可能性
 - 回収率への影響
 - 年末の商店街での街頭調査
 - 試験前の受験生へのアンケート

調査方法

13

- ・伝統的方法はコストが高い
- ・訪問面接法
 - ・調査員が調査の対象者を訪問して、インタビュー形式で質問に答えてもらう方法
 - ・回答の記入は、調査員が行う場合と、回答者が自分で行う場合がある
 - ・質問の意味をその場で回答者に説明できるので、誤解を防げる
 - ・回収率が高く、回答の信頼性が高い
 - ・複数の調査員を使う場合は、調査員全員が、調査票の意味をよく理解し、共通した説明ができるようにしておく必要がある
 - ・費用がかかるが、調査の過程で、調査票の問題点などがわかるので、調査票作成のプレテストとして行うときにはこの方法がよい



調査方法

14

- 郵送調査法

- 調査票を対象者に郵送で配布し、郵送で回収する方法
 - 利点：調査票回収の人手が要らないぶん費用はややおさえられる
 - 欠点：返送されない割合が高い（回収率が低い）



- 留置調査法

- 調査員が対象者を回って調査票を配布し数日後に回収する方法
 - 利点：回答者が回答に時間を見る場合などに有効
 - 欠点：人手がかかる



- 街頭調査法・店頭調査法

- 調査員が街頭や店頭で対象者を見つけてインタビューする方法
 - インタビューに応じる人を見つけるのが難しい
 - 回答者が偏りやすい



- 電話調査法

- 調査員が対象者に電話で質問して答えてもらう方法
 - 電話調査法では質問の数を少なくして、相手に時間を取らせないようにする必要がある。回答してくれる人に出会うのが難しい

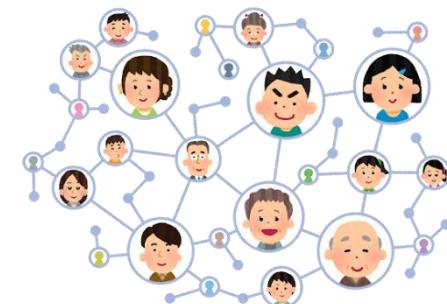
- 集合調査法

- 対象者をある会場に集めてその場で回答してもらう方法
 - 会場で一度に多数の調査票を回収できるという利点があるが、会場の確保や集合のための連絡・準備に人手が必要という欠点がある

調査方法

16

- 最近は比較的コストの低い方法も（選択バイアス注意！）
 - Webサイト上での調査
 - フォームを利用したインタラクティブなアンケートの設計が可能
 - 回答を手作業で入力する必要がない
 - Webページへの誘導が難しく、選択バイアスが生じやすい
 - e-mailによる調査
 - 一斉配信によるプッシュ型のアンケート依頼ができる
 - 回答フォームの設定に難があり、回答の形式が一定しない可能性がある
→ 最近はアンケート用のWebサイトへ誘導することが多い
 - SNSを利用した調査
 - 人的ネットワークを利用できる
 - アンケート用のWebサイトへ誘導することも多い
 - 選択バイアスが避けられない



様々なデータ分析手法

17

- データや分析目的に合わせ適切なデータ分析手法を選ぶ
 - データの種類、サイズ、収集条件などを考慮

		要因・条件（説明変数）	
		量的 	質的 
結果（目的変数）	量的 	(散布図) 回帰分析	(箱ひげ図) 母平均の検定 (Z検定, t検定, 分散分析) 回帰分析+ダミー変数 (数量化I類)
	質的 	ロジスティック回帰分析 など	(クロス表, 分割表) 母比率の検定 カイ二乗検定, フィッシャーの正確検定

様々なデータ分析手法

18

- データや分析目的に合わせ適切なデータ分析手法を選ぶ
 - 要因・条件が「質的」なもの、かつ代表的なものに絞ってみる

		要因・条件 (説明変数)	
		質的	量的
結果 (目的変数)	量的	(箱ひげ図) 母平均の検定 (Z検定, t検定, 分散分析)	
	質的	(クロス表, 分割表) 母比率の検定 カイ二乗検定	

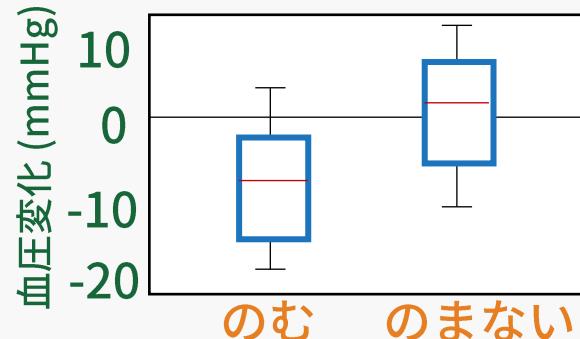
(補足) 上の手法は1群を理論値と比較する際（例：全国平均と等しいか）にも用いられるが、ここでは異なる条件で得られた複数群を比較する場合に注目している

新薬は血圧に対して効果があるか？

要因 : { 新薬のむ or 新薬のまない }



→ 結果 : 血圧が ○mmHg 下がった



		要因・条件 (説明変数)	
		質的	量的
結果 (目的変数)	量的	要因・条件を変えた群の間 (グループ間) の比較 (例) 2群の検定	(箱ひげ図) 母平均の検定 (Z検定, t検定, 分散分析)
	質的		(クロス表, 分割表) 母比率の検定 カイ二乗検定

(補足) 上の手法は1群を理論値と比較する際 (例: 全国平均と等しいか) にも用いられるが、ここでは異なる条件で得られた複数群を比較する場合に注目している

要因・条件の異なる2群の検定

20

- 要因（条件）が2値を取る質的変数ならば2群の検定になる

(例) 新薬を飲む群（治療群*），従来薬を飲む群（対照群）

- 結果が連続変数（例：血中のある物質の濃度が変わった）

- 母平均の検定 → t 検定，ウェルチの t 検定
- 母分散の検定 → F 検定

- 結果も質的変数（例：回復した，回復しなかった）

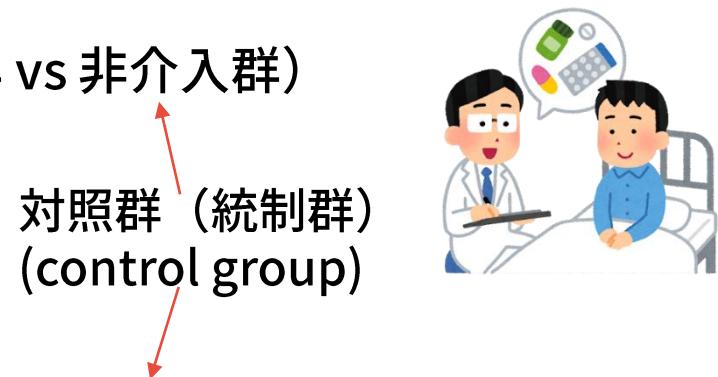
- 母比率の検定 → 次回（2群なら χ^2 検定と一致）

(*) この特定の条件を与えた群のことを実験群 (experiment group) や
処置群・処理群 (treatment group)，介入群 (intervention group) とも呼ぶ。
すでに何らかの特性を持った群を作る場合は，曝露群 (exposed group) (例: 喫煙群)
などと呼ぶ。（これらの言葉の使い方は研究や分野による）

実験と調査観察

21

- 統計の研究 (research/study) には実験研究と調査観察研究がある
- 実験研究 (experimental study)
 - 直接条件を操作する (介入する) (介入群 vs 非介入群)
 - 実験計画法
- 調査観察研究 (observational study)
 - 特定の特性を持つ群の経過を調べる (曝露群 vs 非曝露群)
 - 特定の疾患を持つ群の過去の経歴を調べる (症例群 vs 対照群)
 - 地域ごとの特徴を調べる → 社会調査法



- (*) 日本語の「調査」は文脈によって意味が変わるので注意 (英語も参考に)
- あるときは上の分類 (条件操作できるかどうか) の observation ,
あるときは具体的手段の survey, あるときは研究 study の意味・・

バイアスの問題

22

- ・バイアス（偏り， bias）によって誤った関連性が生じる
 - ・ある要因(条件)の効果やリスクを調べるには少なくとも2群を比較
 - ・調べたい条件以外はグループ間でそろえる必要がある
 - ・しかし注意深く設計しないとすぐにバイアスが生じる…
 - ・選択バイアス
 - ・標本が母集団を代表していないことで生じるバイアス（対象者の選び方の問題）
 - ・交絡
 - ・第三の要因（疑似相関）が存在する
 - ・情報バイアス
 - ・観察方法や測定方法で生じるバイアス（例）先入観，過少申告
 - ・対処：盲検法など

選択バイアス

・選択バイアス (selection bias)

- 標本が母集団を代表していないこと (対象者の選び方の偏り) で生じる

(例) 職場で健康調査し「この職場の人は一般人より健康」と結論
→ そもそも調査時に職場に来ているなら一定以上健康

(例) 実技と筆記試験の関連を合格者で調べ「一般に実技よくない人は筆記試験はよいなあ」と結論

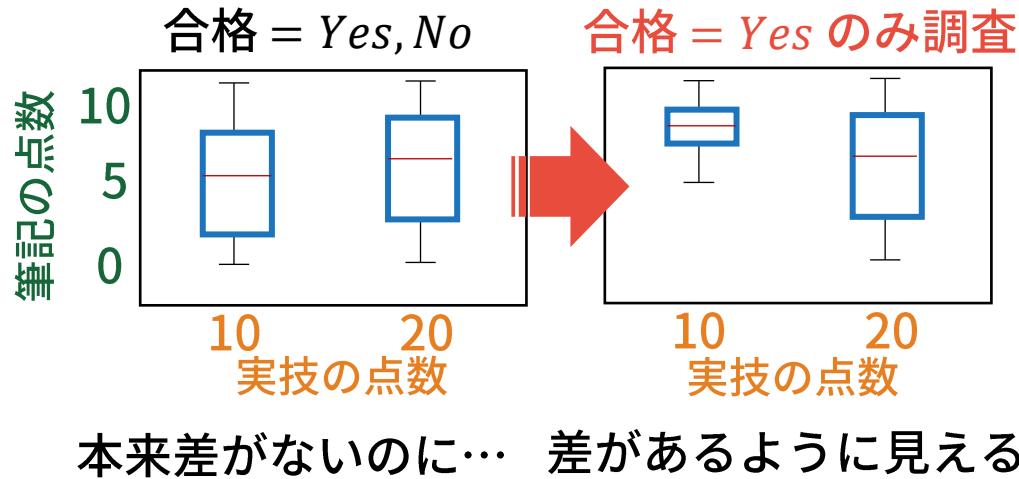


合格 $\in \{\text{Yes}, \text{No}\}$

合計15点以上合格

実技の点数
配点: 10 or 20

筆記の点数
配点: 0~10



交絡

24

・ 交絡 (交絡バイアス) (confounding)

- 第三の要因の影響で偽りの因果関係 (疑似相関, 偽相関) が生じる
 - この要因を**交絡因子**という

(例) アイスクリームがよく売れると水難事故が増える

(例) プラセボ効果 (飲むこと自体の心理的効果で回復)

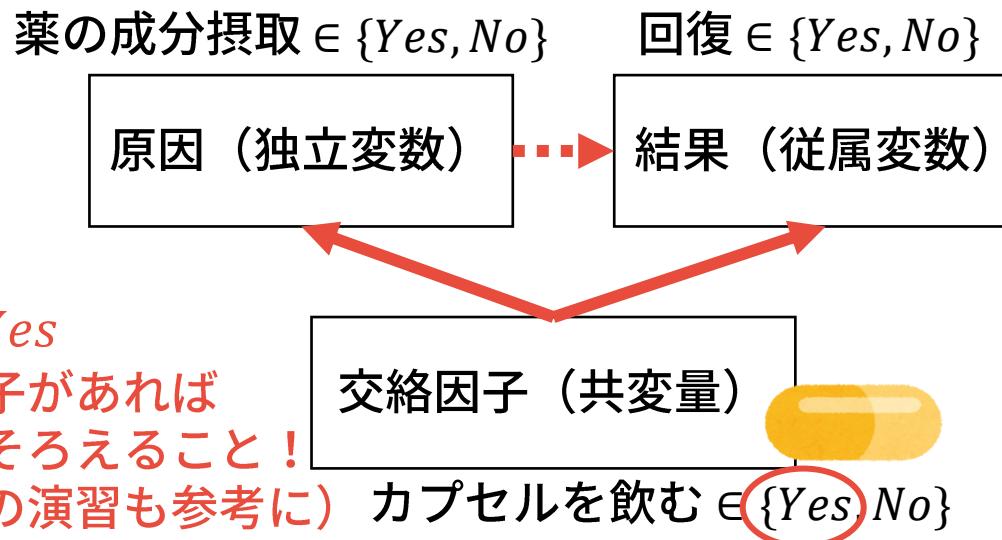
投与群

- カプセルを飲む = Yes
- 薬の成分摂取 = Yes

非投与群

- カプセルを飲む = ~~No~~ = Yes
- 薬の成分摂取 = No

交絡因子があれば
条件をそろえること！
(今日の演習も参考に) カプセルを飲む ∈ {Yes, No}



本日の講義内容

25

- テキスト
 - 「統計学入門」12.1節, 12.5節, 1.3節 (少しだけ)
- 講義トピック
 - 社会調査と抽出方法
 - 実験と調査観察 (介入があるかないか)
 - 代表的バイアス (選択バイアス, 情報バイアス, 交絡)
 - エビデンスレベル
 - ランダム化比較試験, コホート研究, ケース・コントロール研究
 - フィッシャーの三原則
- 演習

- 実験・調査方法によってエビデンスレベルが異なる
 - エビデンス (evidence) : 根拠, 証拠
(例) EBM (Evidence Based Medicine) : 根拠に基づく医療
 - 医療, 疫学で発展したが他の分野でも重要！ (よく出てきます)
- エビデンスレベル (信頼性の高い順に)
 - (0. メタアナリシス - ランダムに研究自体を複数抽出)
 - 1. ランダム化比較試験 (RCT) 実験研究 (介入)
 - 2. コホート研究
 - 3. ケース・コントロール研究 (症例対照研究) }
 - (4. 記述的研究 (症例報告))

・コホート研究 (cohort study)

- ・時間軸方向に解析する
- ・原因 (要因) → 結果

(例) 喫煙者と非喫煙者を追跡して将来の心疾患の発生を比較する

- ・コストや時間がかかる

「現在 → 未来」ではなく「過去の時点A → Aより後の過去」
を見る「後ろ向きコホート」もある

・ケース・コントロール研究 (case-control study, 症例対照研究)

- ・時間軸と逆方向に解析する
- ・結果 (症例) → 原因 (要因)

(例) 心疾患の発症者と非発症者の過去の喫煙状況を調べて比較する

- ・選択バイアスが入りやすい (マッチングなどで対処)
- ・稀な疾病でも使える

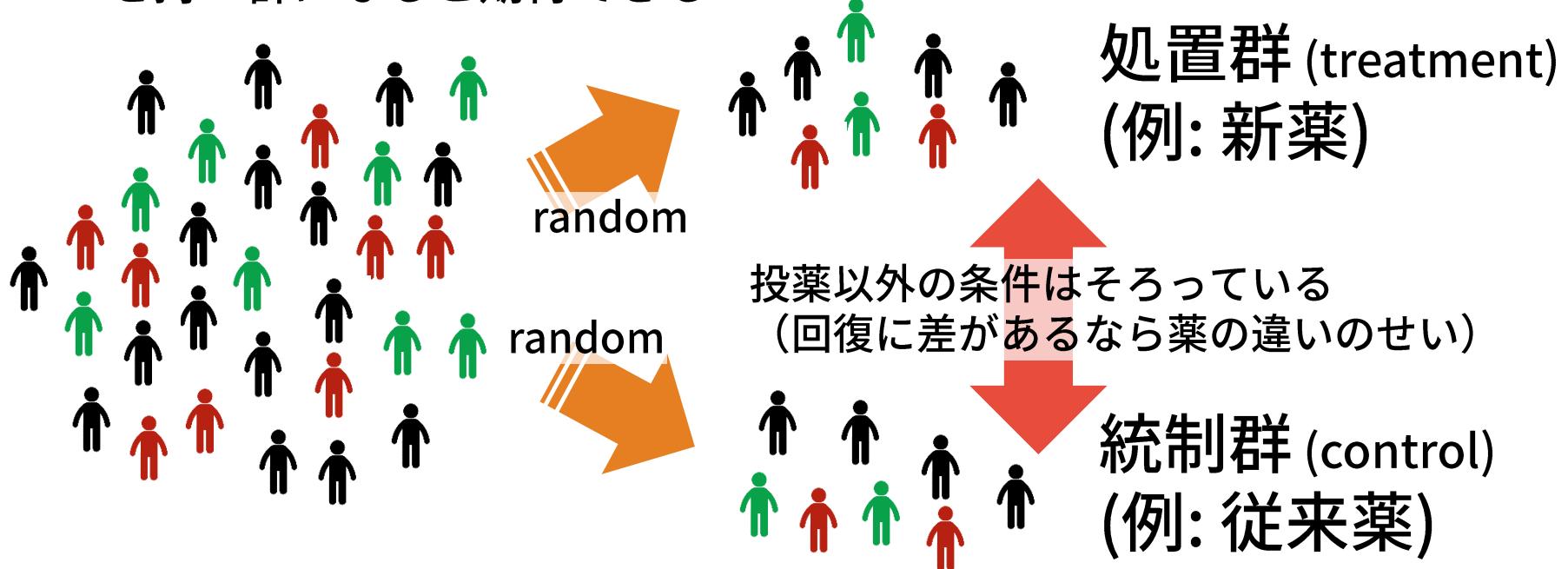
ケース：症例群 (ある時点で既に発生)

コントロール：対照群 (発生していない)

ランダム化比較試験

28

- ・ランダム化比較試験 (Randomized Controlled Trial, RCT)
 - ・処置群, 統制群の2水準へランダムに割り付ける (3水準以上も同様)
 - ・母集団がさまざまな要因を持っていても, 割り付けた各群は同じ特徴を持つ群になると期待できる



フィッシャーの三原則と割り付け

29

- フィッシャーの三原則
 - 反復 (replication)
 - 各条件(水準)において (異なる個体・被験者で) 繰り返し実験を行う
 - 無作為化 (randomization)
 - 処理群をランダムに割り付け → 注目の要因以外の影響減らす
 - 局所管理 (local control)
 - 実験を行う場所や時間を区切って (ブロックと呼ぶ) , 各ブロックで全条件(水準)の実験を行う → 注目の要因以外の影響減らす
(例) 農場試験で, 区画間の差があっても区画内は環境一定に管理しやすい
- 乱塊法 (randomized block design)
 - 上の三原則を満たした割り付け方法 → 分散分析利用 (水準 × ブロック)

乱塊法 (randomized block design)

30

(例) 3水準(条件A, B, C)の実験を各3回, 3日かけて行う

- ・ ブロックは「日にち」
- ・ 同日であれば条件A, B, C以外の状況 (温度, 湿度など) は近いはず

無作為: No, 局所管理: No

- 1日目: A, A, A よくない
(日付による状況の
違いが交絡しうる)
- 2日目: B, B, B
- 3日目: C, C, C

無作為: Yes, 局所管理: No

- 1日目: B, B, C
- 2日目: C, A, B まずまず
(でも同条件が
同日に集まることも)
- 3日目: A, C, A

全体で無作為化

(乱塊法) 無作為: Yes, 局所管理: Yes

- 1日目: A, B, C
- 2日目: B, A, C
- 3日目: C, A, B

よさそう
(さらに, 1日の中での実験順序も
変えるのが一般的)



水準 × 日 (3×3)
の2元配置分散分析

日ごとに
無作為化

(発展) 二元配置分散分析

31

- 母集団の平均に差があるか調べたい → 分散分析
- 複数の要因を考えるときは? → 二元配置分散分析

鳥人間コンテスト用の飛行機の機体作成で

翼の形状を3通りA, B, C変えて

3日間テストしたときの飛行距離 (km)



「日付」の影響と「翼形状」の影響
のうち、「翼形状」の違いで飛行距離が
有意に変わることを示したい

	形状A	形状B	形状C
1日目	10	3	4
2日目	6	1	3
3日目	8	3	2

全体変動

= 形状による変動 + 日付による変動 + 誤差変動

$$F_{wing} = \frac{\text{形状による変動}/\phi_{wing}}{\text{誤差変動}/\phi_e}$$

$$F_{day} = \frac{\text{日付による変動}/\phi_{day}}{\text{誤差変動}/\phi_e}$$

本日の講義内容

32

- テキスト
 - 「統計学入門」12.1節, 12.5節, 1.3節 (少しだけ)
- 講義トピック
 - 社会調査と抽出方法
 - 実験と調査観察 (介入があるかないか)
 - 代表的バイアス (選択バイアス, 情報バイアス, 交絡)
 - エビデンスレベル
 - ランダム化比較試験, コホート研究, ケース・コントロール研究
 - フィッシャーの三原則
- 演習

(復習) 対応のない2群の標本に対する検定

33

- 母集団の平均に差があるかどうかを検定したい
 - 例：AクラスとBクラスの成績に差があるか？

Aクラス	Bクラス
69	49
52	40
68	52
46	37
72	55
40	38
45	45
62	
53	



	Aクラス	Bクラス
標本サイズ n	$n_1 = 9$	$n_2 = 7$
標本平均 \bar{X}	$\bar{X}_1 = 56.33$	$\bar{X}_2 = 45.14$
標準偏差 s	$s_1 = 11.76$	$s_2 = 7.105$
不偏分散 s^2	$s_1^2 = 138.3$	$s_2^2 = 50.48$
母平均	?	?
母分散	?	?

差があるか？

- データの横持ち → 縦持ち

横持ち

Aクラス	Bクラス
69	49
52	40
68	52
46	37
72	55
40	38
45	45
62	
53	



縦持ち (Rcmdr はこちらを使う)

クラス	点数
A	69
A	52
A	68
A	46
A	72
A	40
A	45
A	62
A	53
B	49
B	40
B	52
B	37
B	55
B	38
B	45

2回生前期（データ分析演習）で学ぶ

- Python: pandasでstack()
- Excel: Power Query で「列のピボット解除」

[Rcmdr] 演習1： 2群の平均の差の検定

35

授業ページのリンクより， testscore.csv をダウンロード

- (1) 各クラスの標本の箱ひげ図をプロットせよ.
 - (2) 2クラスの点数の等分散性を仮定し， 2群の両側 t 検定を行え.
 - (3) 等分散性を仮定せずに(2)を行え.
 - (4) 2群に対して一元配置分散分析を行うと， (2)の結果と一致することを確認せよ.
 - (5) (発展：余力あれば) (4)で等分散性を仮定しない場合についても確認してみよう
-
- (2), (3) では， 各群の標本平均， 帰無仮説， 対立仮説を述べること.
 - (2), (3) それぞれで， 用いた t 分布の自由度， 統計量 t の値， p 値を記録し， 有意水準 5% での検定結果を述べること。
(例) 自由度 ○○ の t 分布で検定を行った結果， $t = \dots, p = \dots < 0.05$ より …
 - (4) は， 統計量 F の値および p 値を示しながら， (2) (発展では(3)) の p 値と比較せよ.
 - (5) の等分散性を仮定しない分散分析は， 内容については講義では扱っていないが実行してみよう.
 - t 検定や分散分析は， 実行方法は次ページを確認すること. 分散分析が何か，よく復習しよう.

[Rcmdr] 2群の t 検定，分散分析

36

- (復習) csvの読み込み
 - データ > データのインポート > テキストファイルまたは…
 - データセット名: 適当 (例: Score), フィールドの区切り記号: カンマ
- (復習) 箱ひげ図
 - グラフ > 箱ひげ図 > 変数: 点数, 層別: クラス
- 2群の t 検定
 - 統計量 > 平均 > 独立サンプルt検定 > グループ: クラス, 目的変数: 点数
 - オプションで, 等分散性や対立仮説 (両側か片側か) を設定可能
 - 結果はRコンソールに表示される (信頼区間も表示される)
- 分散分析 → 2群では t 検定と結果が一致
 - 統計量 > 平均 > 1元配置分散分析 > グループ: クラス, 目的変数: 点数
 - 「Welchの等分散性を仮定しないF検定」
 - チェックなし: 等分散性を仮定, チェックあり: 等分散性仮定しない

Rコマンダーの「信頼水準」は
区間推定の「信頼係数」, 検定では「1 - 有意水準」のこと

[Rcmdr] 演習2: 入浴時間帯と疲れの関係?

37

授業ページのリンクより, bathdummy.csv をダウンロード

- ・サンプルサイズ $n = 101$ (それぞれ以下の3つのデータを持つ)
 - ・ 入浴時間帯 (Early: 21時よりもまえに入浴, Late: 21時以降に入浴)
 - ・ バイト (Y: あり, N: なし)
 - ・ 疲れ指数 (0-10の連続値)

入浴時間帯は疲れに影響するか, 分析結果を示しながら論じなさい.

- (1) 入浴時間帯に対する疲れ指数を箱ひげ図でプロットしてみる.
- (2) 入浴時間帯の違いによる疲れ指数の差を検定してみる. vs
- (3) バイトの有無はどう関係するか?
 - (3a) バイトの有無と入浴時間帯の関係 (統計量 > 分割表 > 2元表) vs
 - (3b) バイトの有無と疲れ指数の関係 vs

(注意: 本データ bathdummy.csv は実データではなく, 演習用に乱数で生成したデータです)

- 実験計画段階
 - 可能ならランダム化比較試験 (RCT) を行う
 - 倫理的に許されない場合もある
 - 完全なランダム割り付けができない場合も、マッチングなど可能な限り選択バイアスを減らす
 - 盲検法も実施
 - 交絡因子になりそうなものをそろえる、もしくは一緒に測っておく(性別、年齢、喫煙の有無、etc.)
- 分析段階
 - 交絡因子で層別して、交絡の影響を減らす

[Rcmdr] 交絡因子で層別化

39

- 交絡因子の値で層別化（グループ化）して交絡の影響をなくす
 - 交絡因子による層別化：「入浴時間」が「疲れ」に与える影響を調べたいときに、「バイトの有無」の影響をなくすには、「バイト有」（もしくは「バイト無」）のグループだけにして（交絡因子をそろえて），あらためて「入浴時間」と「疲れ」の関係を，箱ひげ図やt検定などで調べればよい
- [Rcmdr] バイト有だけのグループにする
 - データ>アクティブセット>アクティブデータセットの部分集合を抽出
 - 「すべての変数を含む」にチェック，「部分集合の表現」を「バイト == "Y"'」，「新しいデータセットの名前」を「BathJobY」など別名とする。
- [Rcmdr] バイト無だけのグループにする
 - データ>アクティブデータセット>アクティブデータセットの選択
 - csvを読み込んだ時の元のデータセット名を選ぶ（アイコンからも選択可）
 - 「アクティブデータセット」とは現在の分析対象データセット
 - バイト有を取り出したときと同様に「部分集合の表現」を「バイト == "N"'」，「新しいデータセットの名前」を「BathJobN」など別名とする。



「バイト」以外は半角文字
「=」は2個

宿題（レポート予告）

40

- 今日の演習1, 演習2
 - 来週ぐらいにレポートを出題しますが, その一部になります.
 - 来週も課題があるので, 今週中にやっておいてください.

(参考) t 検定・ F 検定の実行例 (Excel)

41

Book1 - Microsoft Excel

片側確率にするには2で割る

	A	B	C	D	E	G
1		69		49		0.118
2		52		40		
3		68		52		
4		46		37		
5		72		55		
6		40		38		
7		45		45		
8		62				
9		53				
10	標本サイズ	9	=COUNTA(B1:B9)	7	=COUNTA(D1:D7)	
11	標本平均	56.33	=AVERAGE(B1:B9)	45.14	=AVERAGE(D1:D7)	
12	標準偏差	11.76	=STDEV.S(B1:B9)	7.105	=TDEV.S(D1:D7)	
13	標本分散	138.3	=VAR.S(B1:B9)	50.48	=VAR.S(D1:D7)	

F検定
 $=F.TEST(B1:B9, D1:D7) / 2$

t検定 (等分散が仮定できる場合)
 $=T.TEST(B1:B9, D1:D7, 1, 2)$

t検定 (等分散が仮定できない場合)
 $=T.TEST(B1:B9, D1:D7, 1, 3)$

1: 片側 2: 両側

1: 対応のある場合
2: 対応のない&等分散の場合
3: 対応のない&異分散の場合

(参考) t 検定・ F 検定の実行例 (Excel)

42

ここから簡単に分析することも可能

	変数 1	変数 2
平均	56.3333333	45.1428571
分散	138.25	50.4761905
観測数	9	7
自由度	8	6
観測された分散比	2.73891509	
$P(F \leq f)$ 片側	0.11811957	
F 境界値 片側	4.14680416	

	変数 1	変数 2
平均	56.3333333	45.1428571
分散	138.25	50.4761905
観測数	9	7
プールされた分散	100.632653	
仮説平均との差異	0	
自由度	14	
t	2.21355027	
$P(T \leq t)$ 片側	0.02198415	
t 境界値 片側	1.76131014	
$P(T \leq t)$ 兩側	0.0439683	
t 境界値 兩側	2.14478669	

	変数 1	変数 2
平均	56.3333333	45.1428571
分散	138.25	50.4761905
観測数	9	7
仮説平均との差異	0	
自由度	13	
t	2.35539414	
$P(T \leq t)$ 片側	0.0174368	
t 境界値 片側	1.7709334	
$P(T \leq t)$ 兩側	0.03487359	
t 境界値 兩側	2.16036866	