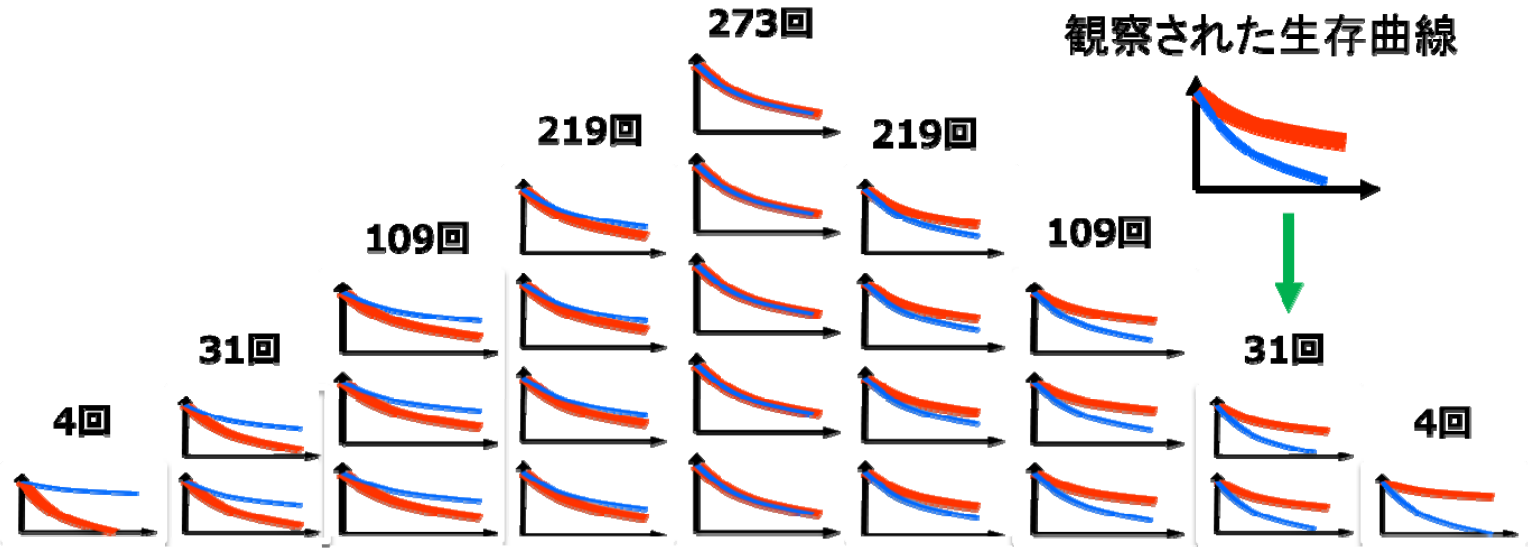
第2回：仮説検定2

生物統計部門　口羽文

生物統計基礎セミナー　2014/1/16（木）

前回の復習： RT群とCRT群の生存曲線の比較

- 帰無仮説「RTとCRTに差がない」の下での分布
- 観察されたRTとCRTの差以上に大きな差になる確率 = P値



前回の復習： P値の定義・解釈

- P値とは，
  帰無仮説「RTとCRTに差がない」の下で
  データよりも極端なことが起きる確率

- P値が大きい
  ・・・当たり前のことが起きた
  ・・・帰無仮説「差がない」を保留

- P値が小さい
  ・・・稀なことが起きた
  ・・・帰無仮説「差がない」を棄却し，「差がある」と判断

今日の目標:
いろいろな検定方法の考え方を理解すること

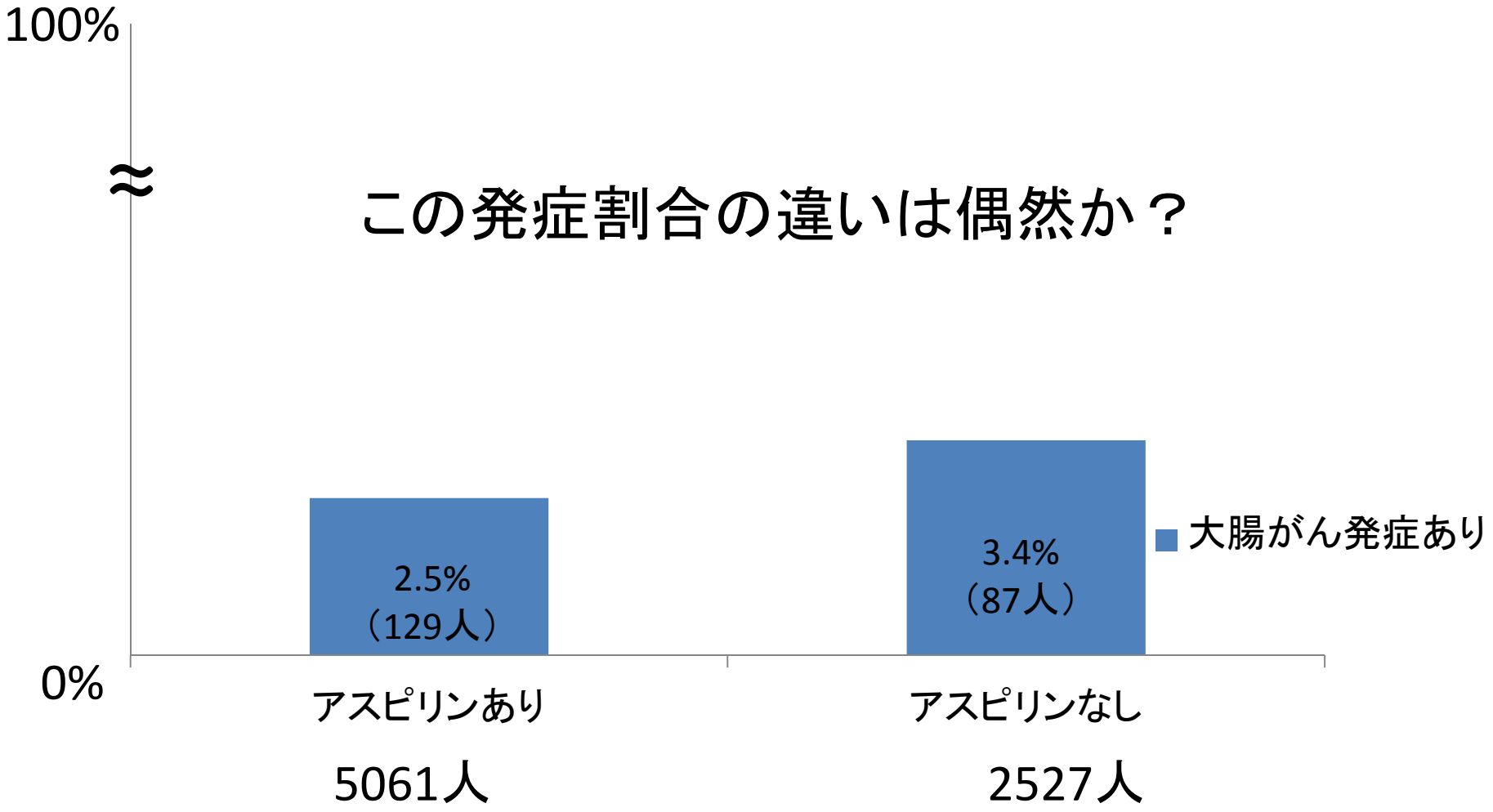
検定にはいろいろな種類

- データの型
  - 発症「あり」・「なし」のような2値
  - 血圧やBMIのような連続量
  - QOLスコア「0，1，2，3」のような離散量
- 研究デザイン
  - どのようにとられたデータか
- 検討したい仮説

流れ

- 2値データ
  - データの要約：2×2分割表
  - Fisherの直接確率検定
  - カイ二乗検定
- 連続量データ
  - データの要約：分布の中心とばらつき
  - t検定
  - Wilcoxon順位和検定
- 離散量データ
- まとめ

2値データの例
研究課題：アスピリンは大腸がんを予防するか



Flossmann et al. (2007) Lancet

2値データの要約：2×2分割表

- 2群×2値のアウトカム

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129人	4932人	5061人
	なし	87人	2440人	2527人
	合計	216人	7372人	7588人

ランダム化比較試験から得られたデータの場合

- 研究課題：アスピリンは大腸がんを予防するか
- ランダム化比較試験

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129	4932	5061
	なし	87	2440	2527
	合計	216	7372	7588

Flossmann et al. (2007) Lancet

研究デザインにより各群の人数は固定できる

- 研究者が決めている

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129	4932	5061
	なし	87	2440	2527
	合計	216	7372	7588

帰無仮説の下では発症あり･なしも固定できる

真にアスピリンと大腸がん発症に関連がなかったら

- 大腸がんを発症した216人は
  アスピリン有無にかかわらず発症する人
- 大腸がんを発症しなかった7372人は
  アスピリン有無にかかわらず発症しない人

		大腸がん発症		
		あり	なし	
アスピリン	あり	129	4932	5061
	なし	87	2440	2527
		216	7372	7588

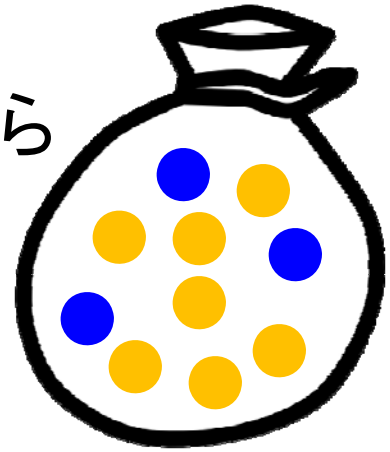
帰無仮説の下では周辺和が固定

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129	4932	5061
	なし	87	2440	2527
	合計	216	7372	7588

- 帰無仮説の下で，ランダム化とは，
  発症する216人と発症しない7372人から構成されている7588人を，ランダムに5061人と2527人に分けていることと同じ
- 超幾何分布に基づく確率計算

超幾何分布の例

- 青玉3個，オレンジ玉7個入っている袋から
  5つ取り出した時に
  青玉が1つ入っている確率は？



	青	オレンジ	合計
袋から取り出された玉	1	4	5
袋に残された玉	2	3	5
合計	3	7	10

このパターンの分割表が得られる確率$= \dfrac{{}_3C_1 \times {}_7C_4}{{}_{10}C_5}$

Fisherの直接確率検定

- 周辺の和を固定した下で，得られる全ての分割表パターンとそれぞれが起きる確率を求め，観察された分割表以上に極端な結果が得られる確率を計算する

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	a	b	5061
	なし	c	d	2527
	合計	216	7372	7588

Fisherの直接確率検定

- すべてのパターンを考える

216	4845	5061
0	2527	2527
216	7372	7588

Pr = 2.4 ×10^{-37}

215	4846	5061
1	2526	2527
216	7372	7588

Pr = 1.3×10^{-35}

214	4847	5061
2	2525	2527
216	7372	7588

Pr = 5.0×10^{-34}

・・・

129	4932	5061
87	2440	2527
216	7372	7588

Pr = 0.0053

観測データと同じ分割表

・・・

0	5061	5061
216	2311	2527
216	7372	7588

Pr = 1.2×10^{-106}

- 片側P値 = 0.0053 + ・・・ + 1.2 ×10^{-106} = 0.0174

（片側2.5%水準で有意！）

ランダム化比較試験では
「ランダム化」に基づいて検定方法が決まる

データの型	検定方法
2値データ	Fisherの直接確率検定
連続量データ	並び替え検定 Wilcoxon順位和検定，など
生存時間データ	ログランク検定，など

ランダム化していないとFisherの直接確率検定を使ってはいけないか？

- そんなことはない
- 多くの非ランダム化研究でも周辺和を固定することが正当化できる
- 例えば，コホート研究デザインによる観察研究
  - アスピリンあり･なしはデザインにより固定できる
  - 帰無仮説の下では，発症のあり･なしも固定と考えられる
  - よって，Fisherの直接確率検定を使うのがよい

流れ

- 2値データ
  - データの要約：2×2分割表
  - Fisherの直接確率検定
  - カイ二乗検定
- 連続量データ
  - データの要約：分布の中心とばらつき
  - t検定
  - Wilcoxon順位和検定
- 離散量データ
- まとめ

母集団からのランダムサンプリングによって得られたデータの場合

- 研究課題：アスピリンは大腸がんを予防するか
- 母集団からランダムサンプリング
  - 普段からアスピリンを飲んでいる集団から5061人，普段アスピリンを飲んでいない集団から2527人をランダムにサンプリングし，発症を追跡

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129	4932	5061
	なし	87	2440	2527
	合計	216	7372	7588

観察研究で，対象者をランダムサンプリングできた場合に等しい

帰無仮説の下で期待される分割表

- もし，アスピリンに効果がなかったら，
  どちらの群の発症割合も216/7588 ≈ 2.8%のはず
- アスピリンに効果がないときに期待される分割表

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	144（2.8%）	4917（97.2%）	5061
	なし	72（2.8%）	2455（97.2%）	2527
	合計	216（2.8%）	7372（97.2%）	7588

- 5061 × 216 / 7588 = 144.06 ≈ 144
- 2527 × 216 / 7588 = 71.93 ≈ 72

期待される分割表と観測された分割表

- 帰無仮説の下で期待される分割表

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	144（2.8%）	4917（97.2%）	5061
	なし	72（2.8%）	2455（97.2%）	2527
	合計	216（2.8%）	7372（97.2%）	7588

- 観察された分割表

		大腸がん発症		
		あり	なし	合計
アスピリン	あり	129（2.5%）	4932（97.5%）	5061
	なし	87（3.4%）	2440（96.6%）	2527
	合計	216	7372	7588

カイ二乗検定

観測		期待	
129（2.5%）	4932（97.5%）	144（2.8%）	4917（97.2%）
87（3.4%）	2440（96.6%）	72（2.8%）	2455（97.2%）

- 観測度数と期待度数の差に基づく検定
  - 観測度数と期待度数の差が大きいほど
    帰無仮説の下で稀な結果が観測されたといえる
- 統計量 $\frac{(129-144)^2}{144}+\frac{(4932-4917)^2}{4917}+\frac{(87-72)^2}{72}+\frac{(2440-2455)^2}{2455}$
- 自由度1のカイ二乗分布に近似的に従う
  - 期待度数が大きいほど近似がよい
- P = 0.0273 （両側5%水準で有意！）

Fisherの直接確率検定とカイ二乗検定

- Fisherの直接確率検定
  - 周辺和の固定
  - 正確な確率分布
  - 両側P = 0.0334
- カイ二乗検定
  - ランダムサンプリングと十分なサンプルサイズ
  - カイ二乗分布への近似
  - 両側P = 0.0273
- 実際にランダムサンプリングできることはほとんどないのでカイ二乗検定の前提は成り立っていない
- 多くの場合，Fisherの直接確率検定を使うのが正しい
- しかし，サンプルサイズが大きければ，結果はほぼ同じ

たいていの場合，どちらを使っても問題にならない
（結果は変わらない）

Tips 1: 2×2分割表の検定によくあるハウツー

		発がん	
		あり	なし
アスピリン	あり	2	6
	なし	4	4

仮想例

- サンプルサイズが小さければFisherの直接確率検定，大きければカイ二乗検定
- Yatesの連続補正カイ二乗検定をしよう

Tips 1: 2×2分割表の検定によくあるハウツー

- 2×2分割表でサンプルサイズが小さかったら
  Fisherの直接確率検定，大きかったらカイ二乗検定？
- 2×2分割表ではYatesの連続補正をしよう？

- ランダム化試験ならFisherの直接確率検定
- 多くの観察研究でもFisherの直接確率検定が良い
  - カイ二乗検定はFisherの直接確率検定の近似と
    考えられるのでOK
- Yatesの連続補正はFisherの直接確率検定の近似
  - Fisherの直接確率検定を行えばよい

流れ

- 2値データ
  - データの要約：2×2分割表
  - Fisherの直接確率検定
  - カイ二乗検定
- 連続量データ
  - データの要約：分布の中心とばらつき
  - t検定
  - Wilcoxon順位和検定
- 離散量データ
- まとめ

平均値の特徴

- 数値例

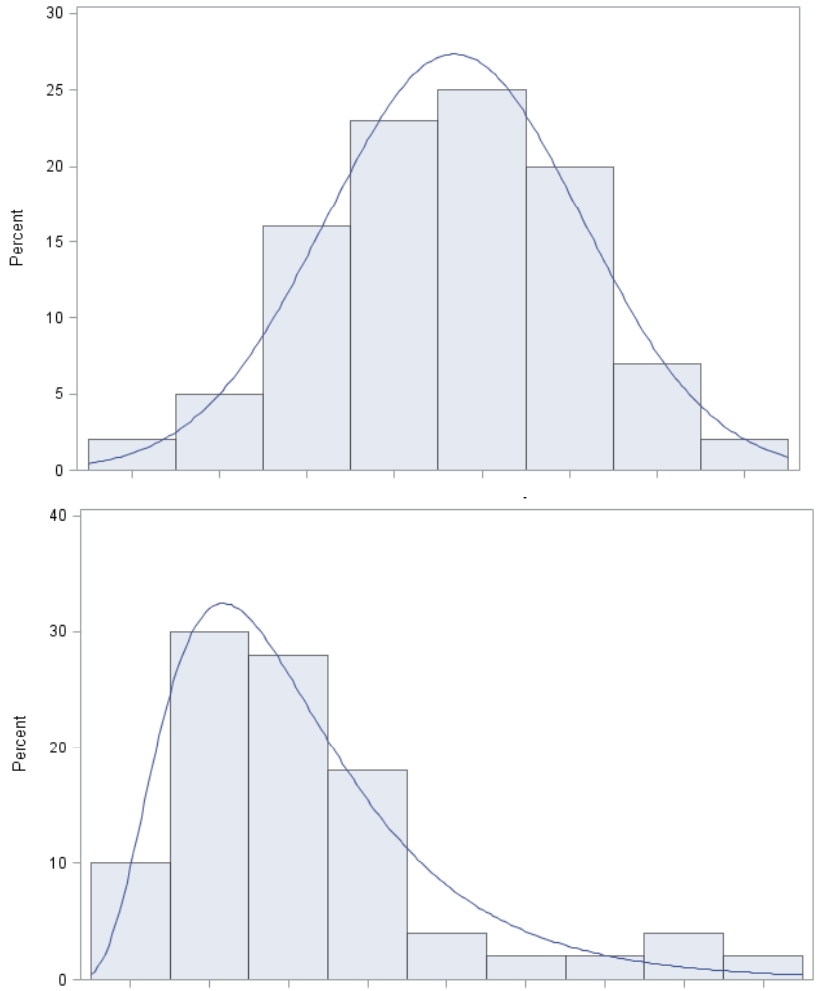
データ	平均値	中央値
1, 3, 5, 7, 9	5	5
1, 3, 5, 7, **100**	23.2	5

  - 平均値：すべてを足して個数で割った値
  - 中央値：データの真ん中の値
- 平均値は中央値と比べて大きな(小さな)値に影響されやすい
  - 大きな(小さな)値 = 外れ値

水澤純基先生のスライドをもとに作成させていただきました

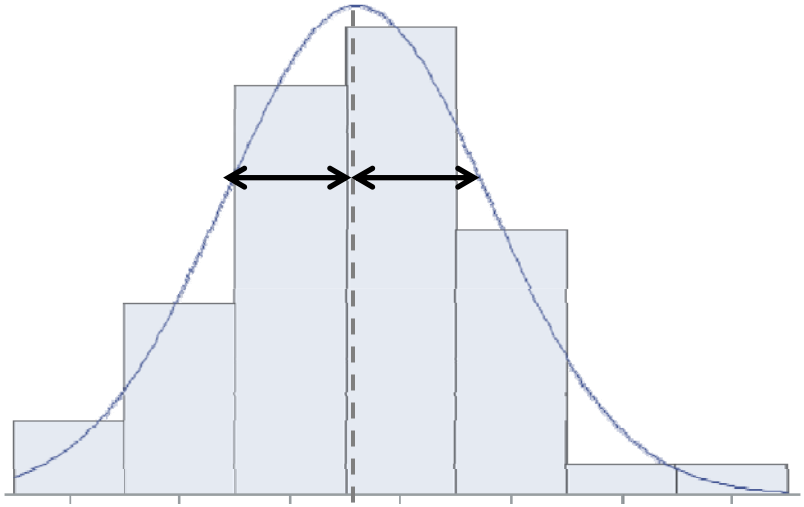
データの要約: 分布の中心

- 平均値
  - ヒストグラムが
    だいたい左右対称
  - 外れ値がない
  - 平均値と中央値は
    だいたい同じ
- 中央値
  - 上記以外
  - 例えば，ヒストグラムが
    歪んでいる



データの要約: 分布のばらつき

- 標準偏差（Standard Deviation; SD）
  - 平均値まわりにデータが
    どの程度ばらついているか
  - 分散=（標準偏差）^{2}
- 最小値，最大値
- 分位点
  - 例えば，25%点と75%点



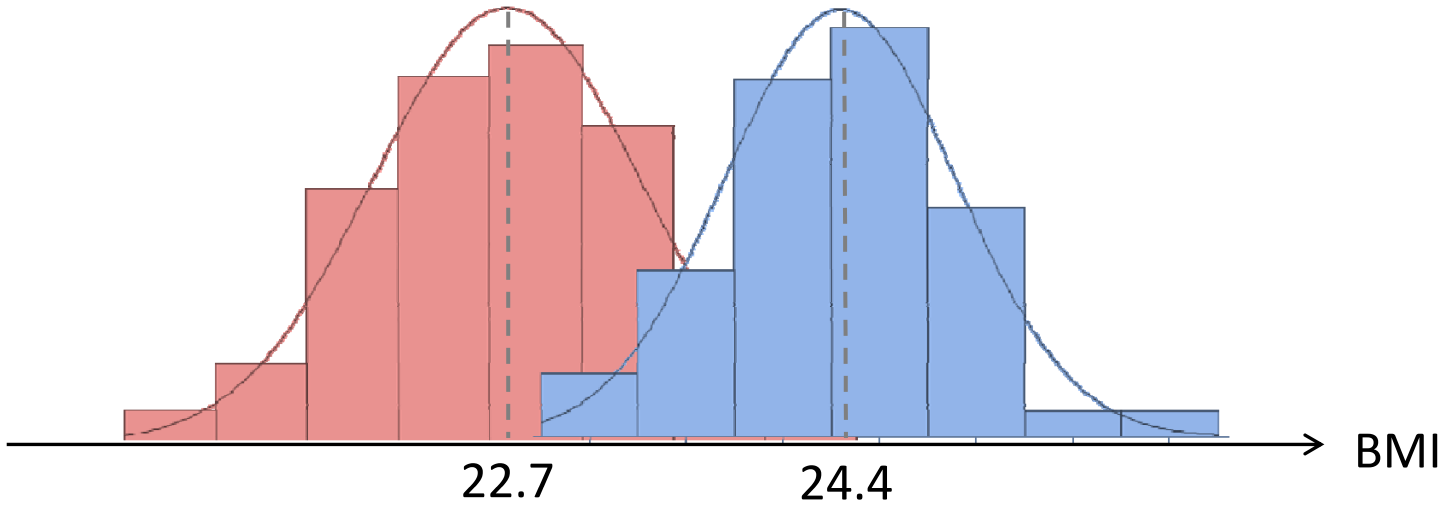
連続量データの比較

- 研究課題：ダイエット法A（A群）とダイエット法B（B群）で，ダイエット後のBMIが異なるか？
- 帰無仮説：A群とB群でダイエット後のBMIが等しい

BMI(Kg/m^2)			
A群 (n=100)	22.0	・・・	22.3
B群 (n=100)	23.3	・・・	23.5

まず，ヒストグラムで分布の形を確認

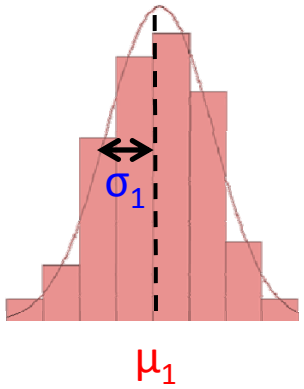
BMI(Kg/m^2)				平均値
A群 (n=100)	22.0	・・・	22.3	22.7
B群 (n=100)	23.3	・・・	23.5	24.4

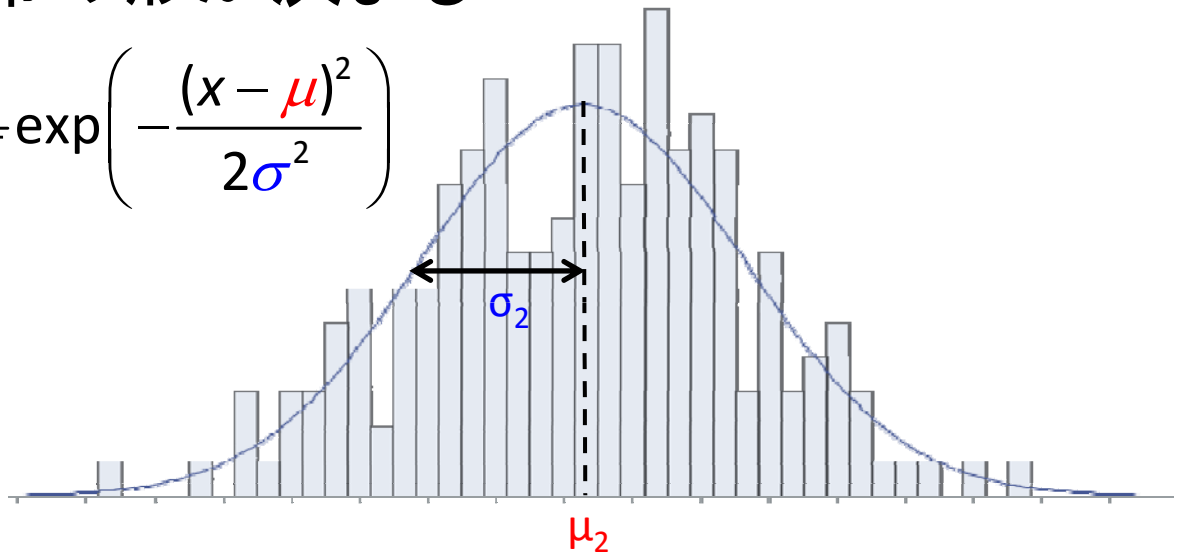


正規分布

- 平均と分散で分布の形が決まる

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$





- データが正規分布に従っている場合，
  2群の分散が等しいことが仮定できれば，
  「平均値の比較」が「分布の比較」

2群の平均値の比較: t検定

- t統計量

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{V\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$\overline{X}_A$: A群での平均値

$\overline{X}_B$: B群での平均値

V : 両群に共通する分散の推定値

n_A : A群のサンプルサイズ

n_B : B群のサンプルサイズ

- データが正規分布に従うとき，t統計量はt分布に従う
- データからt統計量を計算し，そのt統計量とt分布とを比べることによってP値を計算
  - P値に基づいて，ダイエット法の効果に差があると言えるかどうかを判断

Tips 2: 連続量データの検定によるあるハウツー

BMI(Kg/m^2)の比較				
A群 (n=100)	22.0	・・・	・・・	22.3
B群 (n=100)	23.3	・・・	・・・	23.5

- 等分散の検定を行い，群間で分散が異なるときは Welchのt検定
- 正規性の検定を行い，正規分布に従っていないときは t検定をしてはいけない

Welchのt検定

- Welchのt統計量

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{(\frac{V_A}{n_A} + \frac{V_B}{n_B})}}$$

$\overline{X}_A$: A群での平均値

$\overline{X}_B$: B群での平均値

V_A : A群での分散の推定値

V_B : B群での分散の推定値

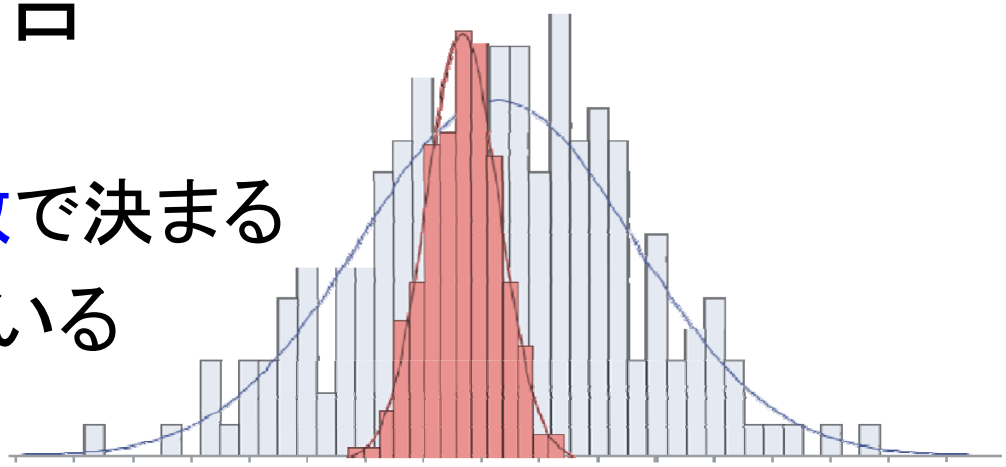
n_A : A群のサンプルサイズ

n_B : B群のサンプルサイズ

- 2群の分散が異なっても，Welchのt統計量は近似的にt分布に従う
- データからt統計量を計算し，P値を求め，2群の平均値が等しいかを判断

検定としては正しいが，これでいいのでしょうか？？

分散が等しくない場合



- 正規分布は平均と分散で決まる
- すでに分布が異なっている
  - これ自体が結果
  - それでも平均値が等しいかどうかを検討したいか，を考える必要がある
- t検定は群間の分散の違いに比較的頑健（ロバスト）
  - 特に，両群のサンプルサイズが等しい場合は気にする必要はない

正規分布に従っていない場合

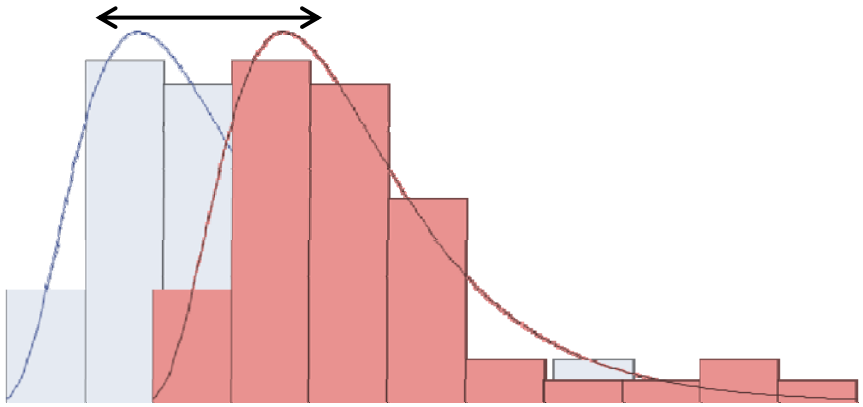
- t検定は正規性の仮定に比較的頑健（ロバスト）
  - αエラーは名義水準以下に抑えられる
  - ただし，あまりにずれている，歪んでいる，外れ値があるとpowerが下がる

連続量データを比較するときのまとめ

- ヒストグラムを描いて分布の形を確認
- 2群間で分布の形がそれほど異なっていない場合
  - t検定を行えばよい
- 2群間で分布の形が大きく異なる場合
  - 平均値を比較することに意味があるかどうかを考える
  - もし比較したい場合，多くはt検定で問題ない

連続量データを比較するもう一つの方法

- 正規分布からあまりにもずれている，外れ値がある
- ノンパラメトリックな方法
  - データに分布を仮定しない
- 同じ形の分布の位置のずれを検定



Wilcoxon順位和検定

	BMI (Kg/m^2)				平均値
A群	22.0 [2]	28.3 [8]	19.4 [1]	22.3 [3]	23.0
B群	23.3 [4]	25.1 [7]	24.6 [6]	23.5 [5]	24.1

- 群を併合して，値が小さい順に[順位]をつける
- 群ごとに順位の和を計算する
  - A群の順位和 = 2+8+1+3 = 14
  - B群の順位和 = 4+7+6+5 = 22
- 「順位」なので外れ値の影響を受けない
  - 外れ値に対してロバスト
  - 28.3 (Kg/m^2) → 40.0 (Kg/m^2):
A群の平均値 → 25.9
A群の順位和 → 14

Wilcoxon順位和検定

- 帰無仮説の下での順位和を考える
  - 2群間に差がなかったら，順位も均等に分布するはず
  - (1+2+…+8)/2 = 18
- 「観測された順位和」と「帰無仮説の下での順位和」を比較する
  - 観測値は14
  - 帰無仮説の下では18
  - 差は4
- 差の理論分布を用いてP値を計算
  - P=0.346 （有意ではない）
  - ダイエット法の効果に差があるとは言えない，と判断

Tips 3: 連続量データの検定によくあるハウツー

BMI(Kg/m^2)の比較				
A群 (n=4)	22.0	28.3	19.4	22.3
B群 (n=4)	23.3	25.1	24.6	23.5

- サンプルサイズが小さければWilcoxon順位和検定
  （サンプルサイズが十分あればt検定）

t検定とWilcoxon順位和検定

- t検定
  - 正規分布を仮定し，平均値を比較
    - 分布の仮定に比較的ロバスト
  - 正規分布の下では，最も検出力が高い
- Wilcoxon順位和検定
  - 分布を仮定せず，順位和を比較
    - 正規分布の下で使ってもよい
    - この場合，検出力も十分高い
  - 外れ値に対してロバスト

流れ

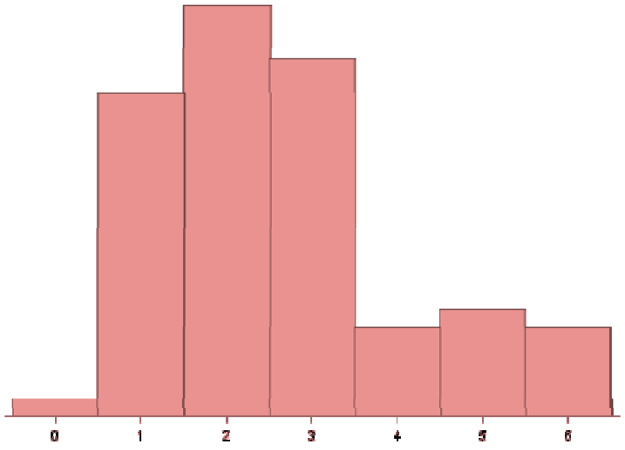
- 2値データ
  - データの要約：2×2分割表
  - Fisherの直接確率検定
  - カイ二乗検定
- 連続量データ
  - データの要約：分布の中心とばらつき
  - t検定
  - Wilcoxon順位和検定
- 離散量データ
- まとめ

離散量データの比較

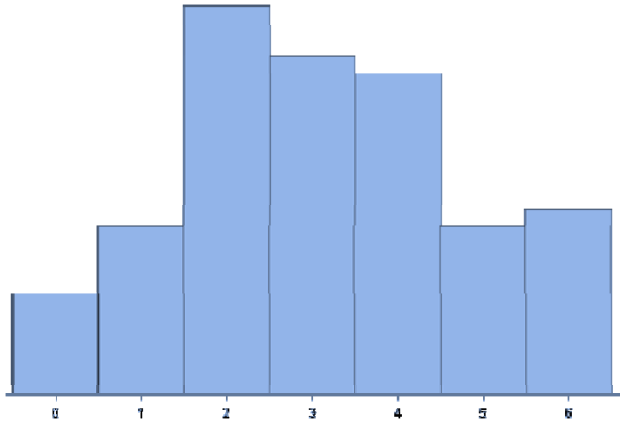
- QOLや満足度調査など

	非常に満足	満足	やや満足	普通	やや不満	不満	非常に不満
スコア	0点	1点	2点	3点	4点	5点	6点

- ダイエット法A群とダイエット法B群とで満足度スコアのヒストグラムを描いてみる



A群



B群

離散量データの分布を比較したい場合

- t検定
  - 正規分布を仮定し，平均値を比較
    - 分布の仮定に比較的ロバスト
  - 正規分布の下では，最も検出力が高い
- Wilcoxon順位和検定
  - 分布を仮定せず，順位和を比較
    - 正規分布の下で使ってもよい
    - この場合，検出力も十分高い
  - 外れ値に対してロバスト

- 離散型のデータでも，t検定やWilcoxon順位和検定を用いることができる

流れ

- 2値データ
  - データの要約：2×2分割表
  - Fisherの直接確率検定
  - カイ二乗検定
- 連続量データ
  - データの要約：分布の中心とばらつき
  - t検定
  - Wilcoxon順位和検定
- 離散量データ
- まとめ

今日の目標:
いろいろな検定方法の考え方を理解すること

検定にはいろいろな種類

- データの型
  - 発症「あり」・「なし」のような2値
  - 血圧やBMIのような連続量
  - QOLスコア「0，1，2，3」のような離散量
- 研究デザイン
  - どのようにとられたデータか
- 検討したい仮説