

# 38さまざまなデータモデル /統計的探究プロセス

# この時間の目標

- 統計的探究プロセス
- S 統計的探究プロセスについてよく理解でき、探究で活用しようと思った
- A 統計的探究プロセスについてよく理解できた
- B 統計的探究プロセスについて理解できた
- C 統計的探究プロセスについて理解できなかった

# 38さまざまなデータモデル

# データモデル

- データモデル: データを扱う枠組み
  - 階層型データモデル
  - ネットワーク型データモデル
- 関係データベース (Relational Database: RDB)
  - 関係モデルによるデータベース
  - すべてのデータを表の形で表現

The diagram shows a table with the following structure:

- テーブル** (Table): 書籍表 (Books Table)
- フィールド** (Field): 書名 (Book Title)
- レコード** (Record): 著者名 (Author Name)
- 属性名** (Attribute Name): 分類記号 (Classification Code)

図書番号	書名	著者名	出版社	分類記号
100070408	パスワード忍びの里	松原秀行	A社	913
100020411	福音の少年	あさのあつこ	B社	913
100108216	そのときは彼によろしく	市川拓司	C社	913
100018233	晩夏のプレイボール	あさのあつこ	D社	913
100039247	星のアカバール	オグ・マンディーノ	E社	933

# データベース管理システム

- データベース管理システム
  - Database Management System:DBMS
  - データベースの作成運用管理をおこなう
  - ソフトウェアの1種(規模によっては専用ハードウェア)
- DBMSの機能
  - 共有データを同時に操作しても矛盾が生じない
  - データの重複や不正なデータの登録・更新を防ぐ
  - アクセス権の設定, ユーザIDとパスワードによる認証
  - データとプログラムを分けて管理し, 障害からの復旧を行う

# 構造化データと非構造化データ

- 構造化データ
  - データを関係(Relation)により構造化して蓄積
  - 集計や分析がしやすい
- 非構造化データ
  - 構造化できないデータ
  - 膨大かつ多種多様な文書や画像, 音声, 動画など
  - コンピュータの性能の向上やインターネットの普及に伴い急増
- 有効活用のために
  - データ形式や関係性を把握
  - その関係をコンピュータに認識させる必要
  - 実現するための技術開発

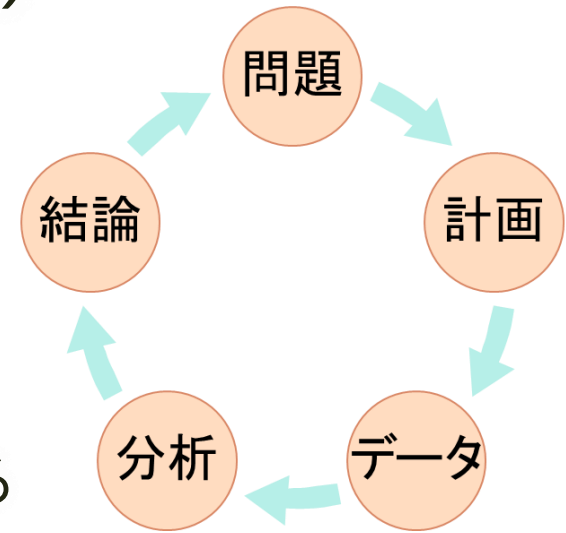
# 統計的探究プロセス

仮説検定・相関係数・標準偏差

# 統計的探究プロセス (数学 I 201ページ)

## PPDACサイクル

- P(problem、問題)
  - 解決すべき事柄を把握し、統計で扱える問題を設定する
- P(plan、計画)
  - 設定した問題に対して、集めるべきデータとその集め方を考える
- D(data、データ収集)
  - 計画に従ってデータを集め、表などに整理する
- A(analysis、分析)
  - 目的やデータの種類に応じてグラフにまとめたり、データに関する数値を求めたりして、特徴や傾向を把握する
- C(conclusion、結論)
  - 見出した特徴や傾向から結論をまとめて表現したり、さらなる課題や改善点を見出したりする





# 仮説検定の考え方(数学 I 202ページ)

- ボールペンの新製品B
  - 30人に聞いてみた
  - BがAより書きやすい 21人
  - Bが書きやすいと言えるか
- 差がないと考えると
  - Aを選ぶ・Bを選ぶは0.5の確率
- コイン投げ30回で1セットとしてシミュレーション
  - 200セット繰り返す！
  - 6000回投げて正確に記録！

# 仮説検定の考え方 (数学 I 202ページ)

- コイン投げ30回で1セットとしてシミュレーション
  - 乱数で0か1を生成
  - 1の数を加算すれば表の回数わかる

```
import random
x=0
for i in range(30):
    x=x+random.randint(0, 1)
print(x)
```

# 仮説検定の考え方(数学 I 202ページ)

- コイン投げ30回で1セットとしてシミュレーション
  - 200セット繰り返す！

数えるの  
面倒！

```
import random
def coin30():
    x=0
    for i in range(30):
        x=x+random.randint(0, 1)
    return x
a=[]
for j in range(200):
    a.append(coin30())
print(a)
```

コイン投げ30回の表の数

コイン投げ30回を200セット

[17, 13, 20, 20, 15, 19, 14, 17, 14, 20, 20, 16, 12, 16, 10, 11, 14, 19, 19, 13, 15, 19, 12, 17, 19, 17, 15, 15, 17, 14, 15, 18, 14, 14, 13, 12, 16, 20, 11, 14, 14, 14, 16, 13, 14, 13, 12, 11, 18, 20, 12, 16, 14, 14, 17, 16, 18, 13, 15, 11, 12, 14, 16, 17, 22, 14, 11, 18, 15, 18, 17, 12, 11, 16, 17, 15, 13, 15, 14, 16, 13, 16, 13, 14, 16, 19, 12, 16, 19, 14, 12, 13, 12, 17, 17, 17, 15, 14, 12, 16, 11, 14, 20, 17, 9, 14, 16, 15, 12, 10, 13, 12, 18, 19, 16, 9, 13, 15, 11, 13, 13, 12, 15, 12, 12, 16, 11, 17, 14, 16, 14, 14, 16, 17, 14, 20, 7, 13, 13, 13, 15, 16, 20, 17, 12, 13, 19, 17, 18, 13, 11, 14, 16, 12, 13, 19, 13, 14, 15, 11, 12, 15, 20, 14, 18, 14, 15, 18, 15, 16, 12, 23, 14, 12, 17, 13, 18, 15, 14, 15, 12, 17, 20, 17, 18, 9, 15, 12, 17, 15, 15, 15, 9, 14, 12, 21, 12, 13, 12, 18]

# 仮説検定の考え方(数学 I 202ページ)

- コイン投げ30回で1セットとしてシミュレーション
  - 200セット繰り返す！

0回: 0	7回: 1
1回: 0	8回: 0
2回: 0	9回: 4
3回: 0	10回: 2
4回: 0	11回: 11
5回: 0	12回: 26
6回: 0	13回: 23
7回: 1	14回: 31
8回: 0	15回: 24
9回: 4	16回: 21
10回: 2	17回: 21
11回: 11	18回: 21
12回: 26	19回: 12
13回: 23	20回: 10
14回: 31	21回: 11
15回: 24	22回: 1
16回: 21	23回: 1
17回: 21	24回: 1
18回: 12	25回: 0
19回: 10	26回: 0
20回: 11	27回: 0
21回: 1	28回: 0
22回: 1	29回: 0
23回: 1	30回: 0
24回: 0	
25回: 0	
26回: 0	
27回: 0	
28回: 0	
29回: 0	
30回: 0	

```
import random
```

```
def coin30():
```

```
    x=0
```

```
    for i in range(30):
```

```
        x=x+random.randint(0, 1)
```

```
    return x
```

コイン投げ30回の表の数

```
a=[]
```

```
for j in range(200):
```

```
    a.append(coin30())
```

コイン投げ30回を200セット

```
for k in range(31):
```

```
    print(k, '回:', a.count(k))
```

回数ごと集計して表示

# 仮説検定の考え方(数学 I 202ページ)

- コイン投げ30回で1セットとしてシミュレーション
  - 200セット繰り返す!
- シミュレーション結果
  - 21回以上は3回
  - 相対度数は  $\frac{3}{200} = 0.015$
  - 1.5%しか起きないことが起きた
    - 偶然起きるとは言えない
    - Bが書きやすいと考えられる

7回: 1  
8回: 0  
9回: 4  
10回: 2  
11回: 11  
12回: 26  
13回: 23  
14回: 31  
15回: 24  
16回: 21  
17回: 21  
18回: 12  
19回: 10  
20回: 11  
21回: 1  
22回: 1  
23回: 1

# 仮説検定の考え方 (数学 I 202ページ)

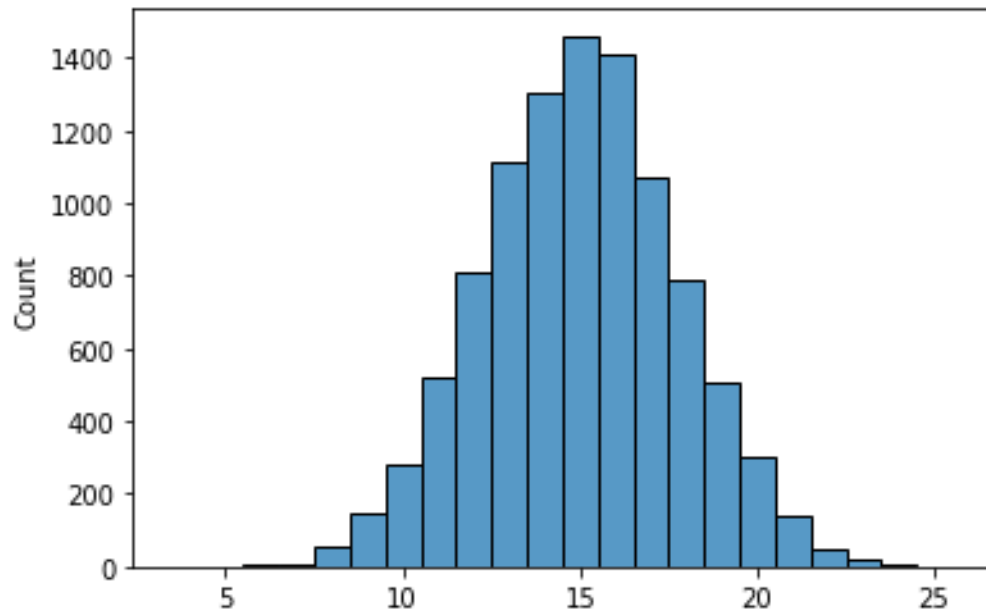
- シミュレーション結果
  - 21回以上は3回
  - 相対度数は  $\frac{3}{200} = 0.015$
  - 1.5%しか起きないことが起きた
- 滅多にないことが起きた
  - 一般的に5%を基準とする

7回: 1  
8回: 0  
9回: 4  
10回: 2  
11回: 11  
12回: 26  
13回: 23  
14回: 31  
15回: 24  
16回: 21  
17回: 21  
18回: 12  
19回: 10  
20回: 11  
21回: 1  
22回: 1  
23回: 1

# 仮説検定の考え方(情報なら)

- コイン投げ30回で1セットとしてシミュレーション
  - 10000回繰り返し  
ヒストグラムにする

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f60c29c2760>



```
import random
import seaborn
```

```
def coin30():
    x=0
    for i in range(30):
        x=x+random.randint(0, 1)
    return x
```

コイン投げ30回の表の数

```
a=[]
```

```
for j in range(10000):
    a.append(coin30())
```

コイン投げ30回を10000セット

```
seaborn.histplot(a, discrete=True)
```

ヒストグラムを表示

# 仮説検定 計算による解法



# おいしいと言えるのか

- ある食品についてアンケート

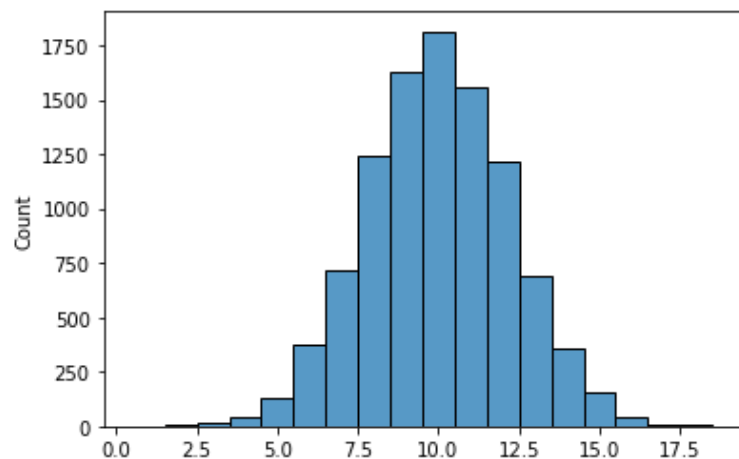
Aがおいしい	Bがおいしい
5	15

- Bがおいしいと結論づけてよいか
- 20人中15人以上が偶然にBを選ぶ確率を求める

# シミュレーションなら

- Pythonでシミュレーション
  - 20回コインを投げて表の数を数える
  - 10000回繰り返す
  - ヒストグラムにする

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f60c24cb520>



- 滅多に起きないようだけど

```
import random
import seaborn
def coin20():
    x=0
    for i in range(20):
        x=x+random.randint(0, 1)
    return x
a=[]
for j in range(10000):
    a.append(coin20())
seaborn.histplot(a, discrete=True)
```

# 確率は計算できる(数学 I 205ページ)

- 反復試行の確率

- 1回の試行で事象Aの起こる確率を $p$ とする。  
この試行を $n$ 回繰り返すとき、  
事象Aがちょうど $r$ 回起こる確率

$${}_n C_r p^r (1 - p)^{n-r}$$

- コイン投げで事象Aがちょうど $r$ 回起こる確率

$${}_n C_r \times \left(\frac{1}{2}\right)^n$$

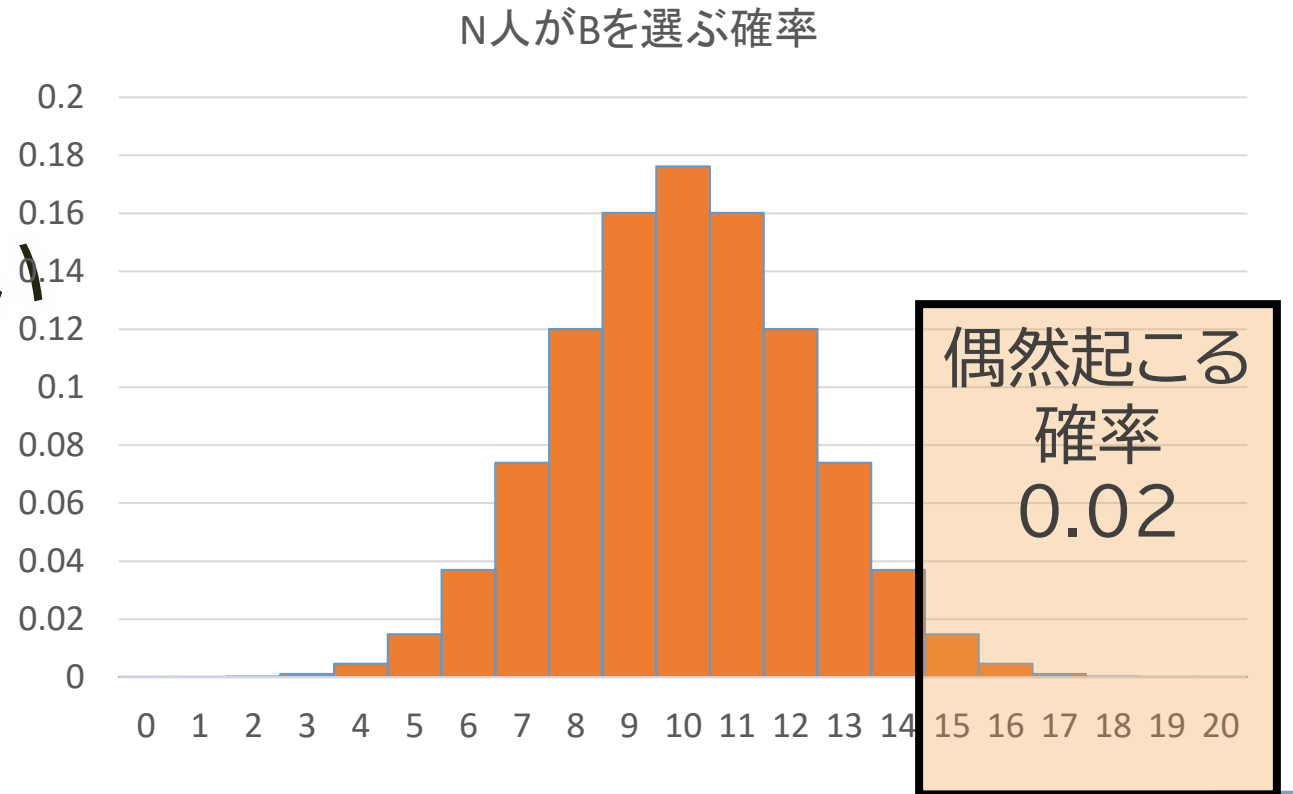
# 確率を計算すると

- Bを20人選ぶ  $\left(\frac{1}{2}\right)^{20} = 0.0000009537$
- Bを19人選ぶ  $\left(\frac{1}{2}\right)^{20} \times {}_{20}C_{19} = 0.0000190735$
- Bを18人選ぶ  $\left(\frac{1}{2}\right)^{20} \times {}_{20}C_{18} = 0.0000181198$
- Bを17人選ぶ  $\left(\frac{1}{2}\right)^{20} \times {}_{20}C_{17} = 0.0010871887$
- Bを16人選ぶ  $\left(\frac{1}{2}\right)^{20} \times {}_{20}C_{16} = 0.0046205520$
- Bを15人選ぶ  $\left(\frac{1}{2}\right)^{20} \times {}_{20}C_{15} = 0.0147857666$

# 確率の合計は

- $0.0000009537 + 0.0000190735 + 0.000181198$   
+  $0.0010871887 + 0.0046205520 + 0.0147857666$   
=  $0.0206947325$

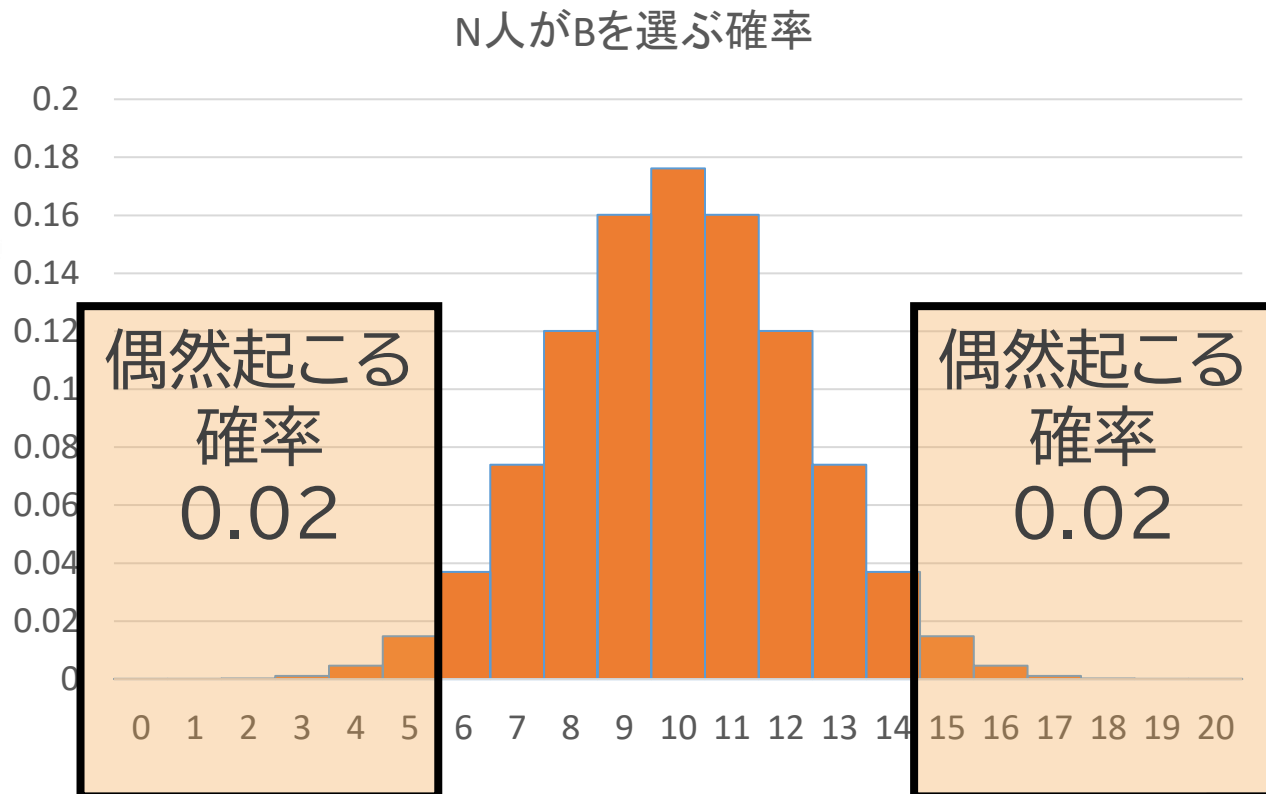
- 2%だから滅多に起きない  
→ Bがおいしいと言える



# 一般的に

- 5%未満だと「滅多に起きない」と考える
  - 厳密な1% 緩やかな10%もある
  - 要求される程度で決める
- 片側と両側がある
  - 2%と見るか4%と見るか
  - 予想がある→片側
  - 予想がない→両側

おいしいかな？→予想あり→片側  
差があるかな？→予想なし→両側  
詳細は数学で学んで！



# 仮説検定

- 仮説検定は重要
  - データが有効かどうか
  - 結論を導けるかどうか
- 仮説検定のためには
  - 毎回シミュレーションしないと
  - 確率を求めないと

# js-STARを使ってみよう

- Web上のツール js-STAR
  - 情報 I のページにリンクがある(ローカル版)
  - js-starで検索(Web版)
- 統計データの分析がサクッとできる



# 1×2表(正確二項検定)

Aがおいしい	Bがおいしい
5	15

# 1×2表(正確二項検定)

- 左上 [1×2表(正確二項検定)]をクリック
- 先ほどのデータを入力

Aがおいしい	Bがおいしい
5	15

# 結果の見方

- 片側確率 0.0207  
→ 2%

	両側検定	片側検定
$\alpha = .01$	ns (p < .005)	ns (p < .01)
$\alpha = .05$	* (p < .025)	* (p < .05)

- 両側検定・片側検定とも  
5% ( $\alpha = .05$ ) で有意(\*)

- 滅多にないこと  
→ Bがおいしい

観測値 1 観測値 2

5	15
---	----

N = 20

Rオプション

区間推定の信頼水準: 0.95

ベイズファクタ: 【パッケージ BayesFactor が必要】

計算!  結果を追加

**結果**

保存 コピー 消去 タブ変換 伸▼ ▲縮

[直接確率計算 1 × 2]

観測値 1	観測値 2
5 (0.2500)	15 (0.7500)

p=0.0207 (片側確率)

	両側検定	片側検定
$\alpha = .01$	ns (p < .005)	ns (p < .01)
$\alpha = .05$	* (p < .025)	* (p < .05)

効果量: g=0.2500

# 2×2表 (Fisher's exact test)

	みかん	いちご
女性	0	5
男性	3	2

# 2×2表(Fisher's exact test) 例題

- みかんといちごのどちらが好きかアンケート
  - 10人の中から3人をランダムに選んだら、女性が0名の確率

$$\bullet \frac{{}_5C_0 \times {}_5C_3}{{}_{10}C_3} = \frac{1}{12} = 0.0833$$

- 計算できるけど

	みかん	いちご	合計
女性	0	5	5
男性	3	2	5
合計	3	7	10

# 2×2表(Fisher's exact test)

## 例題

- [2×2表(Fisher's exact test)]をクリック
- 表のデータを入力
- 結果を見てみよう

	みかん	いちご	合計
女性	0	5	5
男性	3	2	5
合計	3	7	10

# 結果の見方

- 片側検定で 0.0833  
→ 8%

両側検定 :  $p=0.1667$  ns ( $.10 < p$ )  
片側検定 :  $p=0.0833$  + ( $.05 < p < .10$ )

- 片側検定で傾向あり(+)
- 両側検定だと有意でない(ns)

- 断言はできないが  
その傾向にある

	観測値 1	観測値 2
群 1	0	5
群 2	3	2

N = 10

● Rオプション ●

区間推定の信頼水準: 0.95

ベイズファクタ: 【パッケージ BayesFactor が必要】

計算!  結果を追加

**結果**

保存 コピー 消去 タブ変換 伸▼ ▲縮

[直接確率計算 2 × 2]

	観測値 1	観測値 2
群 1	0 (0.0000)	5 (1.0000)
群 2	3 (0.6000)	2 (0.4000)

両側検定 :  $p=0.1667$  ns ( $.10 < p$ )  
片側検定 :  $p=0.0833$  + ( $.05 < p < .10$ )

連関係数 :  $\Phi=0.436$  (イエーツの補正適用)  
効果量 :  $h=-1.7722$   
(大=0.8, 中=0.5, 小=0.2)

# 2×2表(Fisher's exact test) 問題

- 調査を続け、もう少しデータを増やしてみる
  - 30人の中から10人をランダムに選んだら、女性が2名以下の確率

$$\bullet \frac{{}_{15}C_2 \times {}_{15}C_8}{{}_{30}C_{10}} + \frac{{}_{15}C_1 \times {}_{15}C_9}{{}_{30}C_{10}} + \frac{{}_{15}C_0 \times {}_{15}C_{10}}{{}_{30}C_{10}} = 0.02508745627186407$$

- 計算できるけど

	みかん	いちご	合計
女性	2	13	15
男性	8	7	15
合計	10	20	30



# 2×2表(Fisher's exact test) 問題

- 調査を続け、もう少しデータを増やしてみる
- 表のデータを入力
- 結果を見てみよう

	みかん	いちご	合計
女性	2	13	15
男性	8	7	15
合計	10	20	30

# 結果の見方

- 片側検定で 0.0251  
→ 2.5%

両側検定 :  $p=0.0502$  + ( $.05 < p < .10$ )  
片側検定 :  $p=0.0251$  \* ( $p < .05$ )

- 片側検定で有意(\*)
- 両側検定で傾向あり(+)

- 女性はいちごが好きだと考えられる

	観測値 1	観測値 2
群 1	2	13
群 2	8	7

N = 30

● Rオプション ●

区間推定の信頼水準: 0.95

ベイズファクタ: 【パッケージ BayesFactor が必要】

計算!  結果を追加

**結果**

保存 コピー 消去 タブ変換 伸▼ ▲縮

[直接確率計算 2×2]

	観測値 1	観測値 2
群 1	2 (0.1333)	13 (0.8667)
群 2	8 (0.5333)	7 (0.4667)

両側検定 :  $p=0.0502$  + ( $.05 < p < .10$ )  
片側検定 :  $p=0.0251$  \* ( $p < .05$ )

連関係数:  $Fht=0.354$  (イエーツの補正適用)  
効果量:  $h=-0.8900$   
(大=0.8, 中=0.5, 小=0.2)

# 相関係数

# 相関係数(数学 I 194・196ページ)

- 相関関係

- 2つの変量のデータにおいて、一方が増えると他方も増える傾向が認められるとき2つの変量の間には正の相関関係があるという

- 相関係数

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}\{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})^2\}}}$$

# 相関係数計算 例題

- [相関係数の計算と検定]をクリック
- 表のデータを入力
  - Excelに入力して貼付する方法もある

生徒	数学	英語
1	74	81
2	65	66
3	81	74
4	42	59
5	90	88
6	68	45
7	87	65
8	73	76
9	35	47
10	59	71

# 結果の見方

- どこを見るか
- 相関係数 $r=0.657$

生徒	数学	英語
1	74	81
2	65	66
3	81	74
4	42	59
5	90	88
6	68	45
7	87	65
8	73	76
9	35	47
10	59	71

== Means & SDs (SDは標本標準偏差) ==

N= 10

Var.	Mean	S. D.	Min.
	Max.		
1	67.400	17.107	35.000
2	67.200	13.174	45.000

- Item Selection -

Var.	Mean+SD	Mean-SD
1	84.507	50.293
2	80.374	54.026

Correlation Matrix

df= 1 & 8

Var1 Var2

# 結果の見方

- 相関係数  $r=0.657$ 
  - 正の相関
- 相関係数は有意
  - 相関があると考えられる
    - +  $P<.10$
    - \*  $P<.05$
    - \*\*  $P<.01$

Correlation Matrix  
df= 1 & 8

	Var1	Var2
Var1	-	0.657 *
Var2		-

Test of Correlation

Var.	r	F	Test
Var1xVar2	0.657	6.07	*

項目1と項目2には中程度の

/// Analyzed by js-S

相関係数 r	相関の強さ
$0.7 \leq r \leq 1.0$	強い正の相関
$0.4 \leq r \leq 0.7$	正の相関
$0.2 \leq r \leq 0.4$	弱い正の相関
$-0.2 \leq r \leq 0.2$	ほとんど相関がない
$-0.4 \leq r \leq -0.2$	弱い負の相関
$-0.7 \leq r \leq -0.4$	負の相関
$-1.0 \leq r \leq -0.7$	強い負の相関

# 結果の見方

- 散布図もできてる



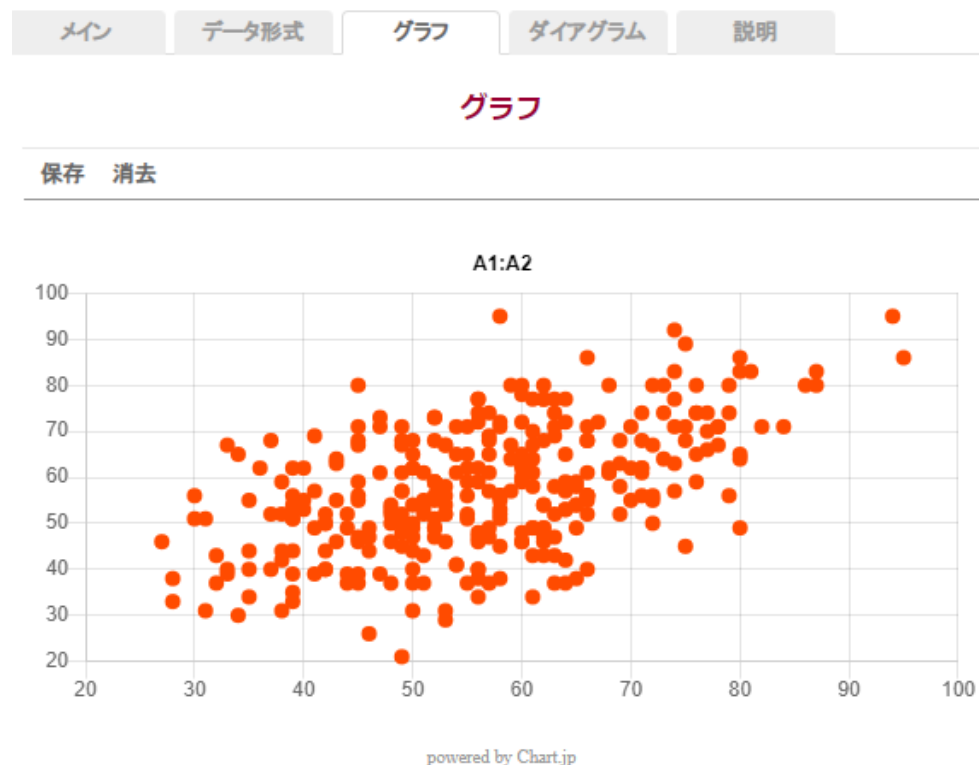


# 相関係数計算 問題

- 情報 I のページ第6回から[ダミーデータ]ダウンロードして開く
- データをコピーしてjs-STARに貼り付ける
- 相関があるか調べてみよう

# 結果の見方

- 316人分の相関係数
- 相関係数  $r=0.524$
- 正の相関



Correlation Matrix  
df= 1 & 314

	Var1	Var2
Var1	-	0.524 **
Var2		-

Test of Correlation

Var.	r	F	Test
Var1xVar2	0.524	118.92	**

項目1と項目2には中程度の関連があります。

/// Analyzed by js-STAR ///

# 情報 + 数学

- データがあれば情報 + 数学の手法
  - データの有効性を示すことができる
    - 相関係数・標準偏差・仮説検定
  - データをわかりやすく伝える事ができる
    - ヒストグラム・箱ひげ図・散布図
    - 円グラフはやめよう
- 探究をPPDACで
  - データに基づき事実を語る

