2021年度

数理統計学

工学部 2 年生向け（水曜日 3 講時）

授業概要

授業の概要

目的　確率と統計の基礎を学び、データ解析のリテラシを養う

内容
（１）確率分布・確率変数・平均値・分散などの確率論の概念
（２）二項分布や正規分布などの基本的な確率分布
（３）母数の点推定・区間推定などの統計的推定
（４）仮説検定の考え方と基本的な形式

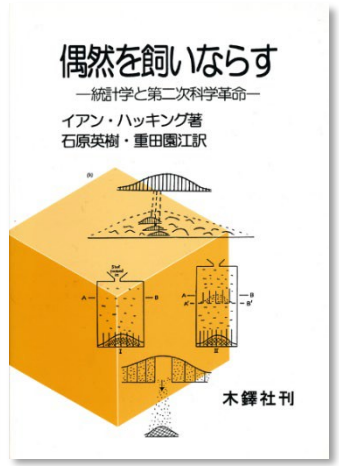
計算機
（１）四則演算・平方根の計算（関数電卓・PCなど）
（２）Excel くらいが使えるとよい
（３）統計ソフト（Rなど）やプログラミング（Python など）は不要

授業の形式

- コロナ感染症対策のため, 教室の授業は流動的
  - 講義時間の前半45分：講義（座学）＋質疑応答
    - この様子は Google Meet によるリアルタイム配信（うまくゆけば）
    - 講義はオンデマンド化を予定（うまくゆけば）
  - 講義時間の後半45分：問題演習など自由学習
- 問題演習
  - 問題集配布(pdf)：やや高度な問題(**)も含めて取り組む
  - 解説ビデオ(mp4)：順次公開する予定
- 成績評価
  - 筆記試験またはレポート試験を予定している（詳細未定）

参考書

[0] 尾畑伸明：「数理統計学の基礎」共立出版, 2014.
[1] P. G. ホーエル（浅井・村上訳）：「入門数理統計学」培風館, 1978.
[2] 東京大学教養学部統計学教室編：「基礎統計学I 統計学入門」東京大学出版会, 1991.
[3] 宮川公男：「基本統計学 第4版」有斐閣, 2015.
[4] 鈴木武・山田作太郎：「数理統計学―基礎から学ぶデータ解析」内田老鶴圃, 1996.



少し易しいテキスト

[5] P. G. ホーエル（浅井・村上訳）：「初等統計学」培風館, 1981
[6] 鈴木義一郎：「初めて学ぶ基本統計学」森北出版, 2005.
[7] 栗原伸一：「入門統計学－検定から多変量解析・実験計画法まで」オーム社, 2011.

演習書

[8] 白砂堤津耶：「例題で学ぶ初歩からの統計学」日本評論社, 2015..
[9] 藤田岳彦：「弱点克服大学生の確率・統計」 東京図書, 2010.

読み物

[10] 神永正博「ウソを見破る統計学」講談社ブルーバックス, 2011.
[11] イアン・ハッキング（石原・重田訳）「偶然を飼いならす」木鐸社, 1999.
[12] 西内啓「統計学が最強の学問である」ダイヤモンド, 2013.
[13] キース・デブリン（原 啓介訳）：「世界を変えた手紙」岩波書店, 2010.

授業予定

週	Lecture 題目と問題番号	教科書の該当部分
1	0　授業概要と序論	第 1 章記述統計 1.1 母集団と標本
2	1　1変量データ 【2.1】	1.2 1変量データの記述
3	2　2変量データ 【10.1～10.4】	1.3 2変量データの記述
4	3　確率の基本 【3.1～4.7】	第 2 章初等確率論 第 3 章 確率変数と確率分布 3.1 確率変数の素朴な導入
5	4　離散型確率分布 【5.1～5.10】	2.　確率変数の分布 3.確率変数の平均と分散 第 5 章 基本的な確率分布 5.1 離散分布
6	5　連続型確率分布 【6.1～6.12】	5.2 連続分布
7	中間まとめ（課題演習等）	
8	6　標本抽出と正規分布 【7.1～7.7】	第 4 章確率変数列 第 6 章大数の法則と中心極限定理 第 7 章母数の推定 7.1 標本抽出の確率モデル
9	7　点推定 【7.1～7.7】	7.2 点推定
10	8　区間推定 【8.1～8.10】	7.3 区間推定
11	中間まとめ（課題演習等）	
12	9　母平均の検定 【9.1～9.9】	第 8 章仮説検定 1.　母数の検定 2.　母平均の検定
13	10 母集団の比較 【9.10～9.12】	2.　母平均の検定（続） 3.　2種類の過誤 4.　等分散の検定
14	11 カイ2乗検定 【9.13～9.15】	8.5 カイ2乗検定
15	まとめ（期末試験を含む）	

数理統計学

一緒に学んでゆきましょう

担 当：尾 畑 伸 明（情報科学研究科）

Lecture 0

序 論

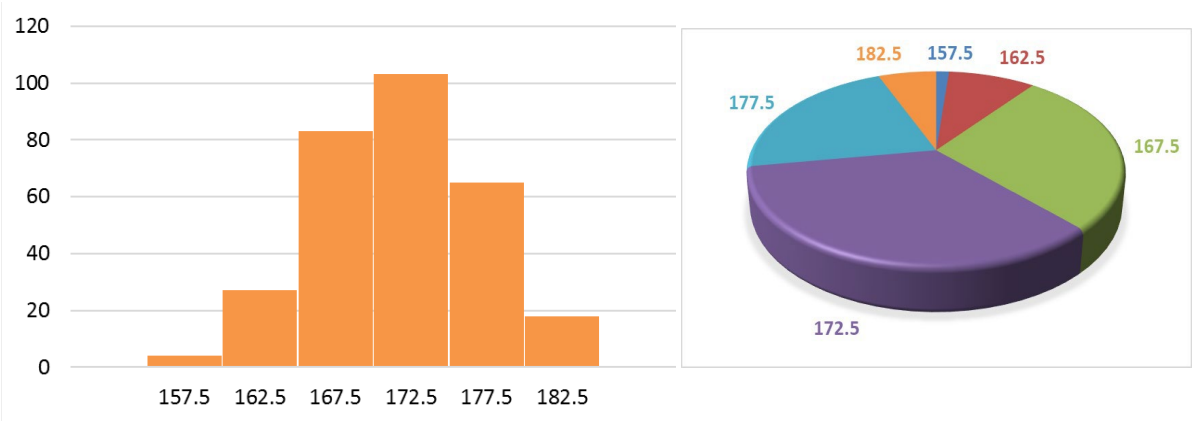
記述統計

調査対象 → 数値データ → 表やグラフ

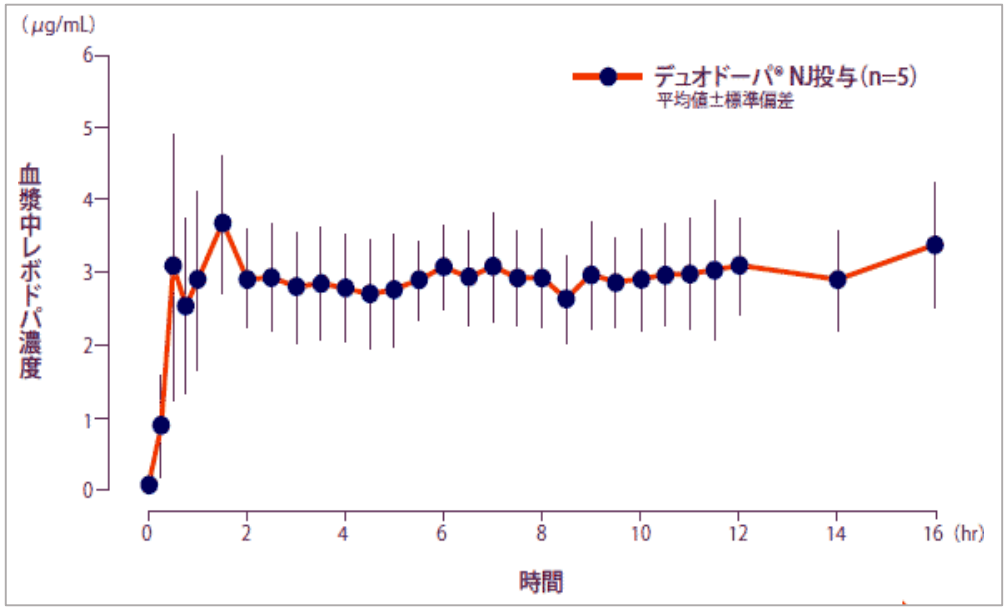
- 国勢調査
- 世論調査
- 視聴率調査
- 様々な実験観察

身長調べ

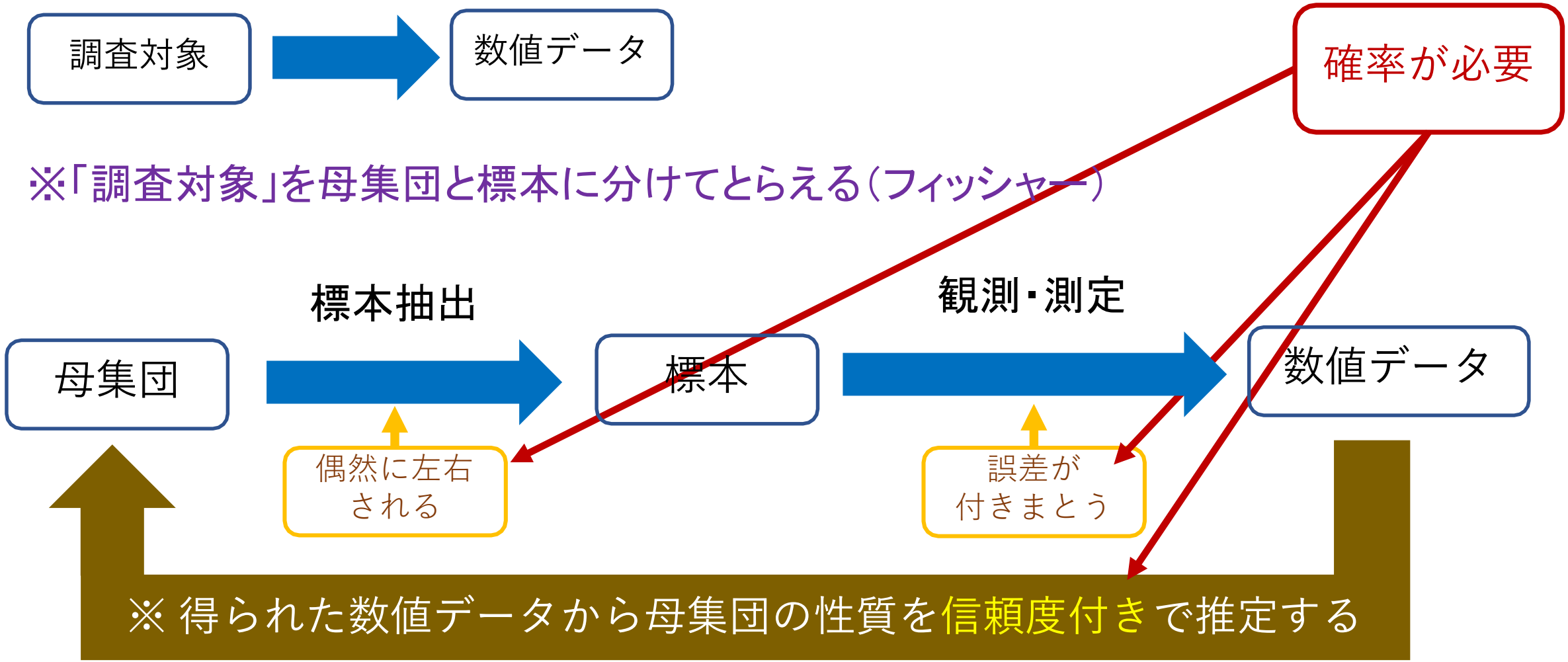
階級	155-160	160-165	165-170	170-175	175-180	180-185	合計
階級値	157.5	162.5	167.5	172.5	177.5	182.5	
度数	4	27	83	103	65	18	300
相対度数	0.013	0.090	0.277	0.343	0.217	0.060	1.000



時間変化（時系列）



推測統計


調査対象
数値データ
確率が必要
※「調査対象」を母集団と標本に分けてとらえる（フィッシャー）
標本抽出
観測・測定
母集団
標本
数値データ
偶然に左右
される
誤差が
付きまとう
※ 得られた数値データから母集団の性質を信頼度付きで推定する

歴史を少し

社会統計 - 人口統計が古い（古代ローマ）

国家運営・各種政策の基礎として重要
現在人口の確認、将来人口の推測、国内の労働力の把握、国家予算、税金・年金、等々

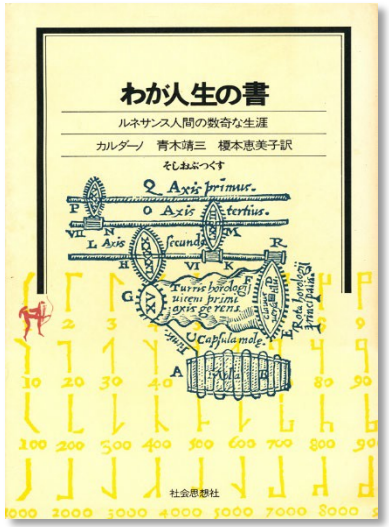
大規模な国勢調査

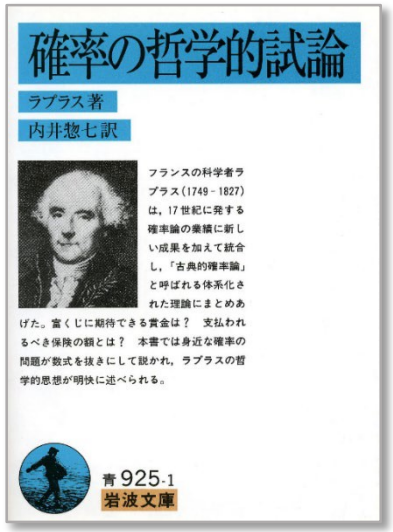
1790 アメリカ国勢調査（Census）
1795 オランダ国勢調査
1801 統計局設置（フランス）, イギリス国勢調査

確率論の発展と応用

サイコロ賭博の話（16〜17世紀）
ド・モアブル：死亡率の計算
　年金理論の始まり（18世紀前半）
ダニエル・ベルヌーイ：天然痘死亡率の寿命への影響
　数理疫学の始まり（18世紀前半）
ラプラス『確率の解析的理論』（1812）
　確率論に高度な微積分を導入、大発展

確率論の創始者カルダーノの自伝





ラプラスによる確率論の啓蒙書

近代統計学

Lambert Adolphe Jacques Quetelet (1796-1874)

- 社会統計を科学的に分析するために確率論を導入

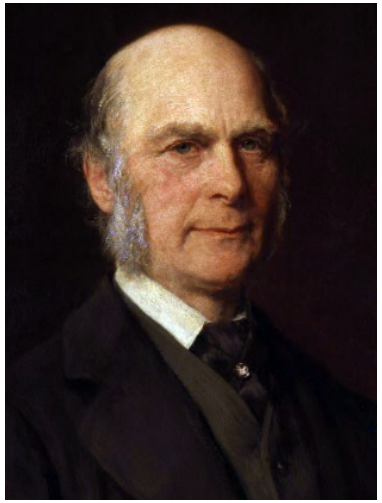
  『人間とその能力の発展について－社会物理学の試み』(1835)

  犯罪率・結婚率・自殺率などの統計法則を他の社会的要因の変数から説明する。

  平均人（社会の重心）の発見

  自由意志で動く人々 ⇒ 社会全体では法則に従っている

- ロンドン統計学会（現王立統計学会）の設立（1834）
- ケトレー指数（Body Mass Index ww/h^2）



Sir Francis Galton (1822-1911)

- C. R. Darwin (1809-1882) のいとこ
- Eugenics「優生学」(1883年造語)
- 近代統計学の父、相関係数の導入、回帰分析の始まり

大標本主義

Karl Pearson (1857-1936)

- 記述統計学の集大成
  データの傾向、関連、規則性の解明
- 現代統計学の基礎概念の確立
  線形回帰、ピアソンの積率相関係数、カイ二乗検定、histogram (1895年に造語)

現代確率論と小標本主義

Sir Ronald Aylmer Fisher (1890-1962)

少数のデータを丁寧に扱って
できるだけ精密な推測を行う

- 推測統計学の創始
  「研究者のための統計学的方法」(1925)
- 実験計画法、分散分析、仮説的無限母集団と無作為標本、最尤法、統計学的十分性、フィッシャーの線形判別関数、フィッシャー情報行列
- 集団遺伝学の創始者（ネオダーウィニズム）

データ駆動型社会の到来
大規模データの利活用

データの構造解析
モデリング理論

高性能計算(HPC)
機械学習などのAI技術

Lecture 1

1変量データの整理

生データ/ ローデータ (raw data)

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180

- わかりやすく整理
- 可視化
- 特徴を抽出

度数分布表 (frequency table)

いくつかの 階級（クラス）に
分類して表に整理する

① データの範囲

最大値 (max)　max = 4180

最小値 (min)　min = 2270

範囲 (range)　R = max - min = 1910

② 階級幅と階級数

決め方には自由度がある

目的や見やすさに応じて決める

生データ/ ローデータ (raw data)

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180

- わかりやすく整理
- 可視化
- 特徴を抽出

度数分布表 (frequency table)

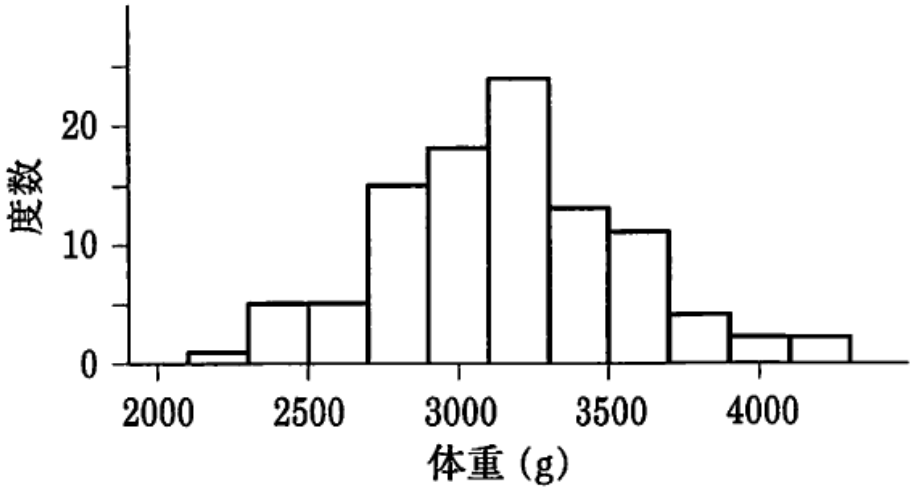
階級番号	階級	階級値	度数
1	2100~2300	2200	1
2	2300~2500	2400	5
3	2500~2700	2600	5
4	2700~2900	2800	15
5	2900~3100	3000	18
6	3100~3300	3200	24
7	3300~3500	3400	13
8	3500~3700	3600	11
9	3700~3900	3800	4
10	3900~4100	4000	2
11	4100~4300	4200	2
計	ー	ー	100

度数分布表 (frequency table)

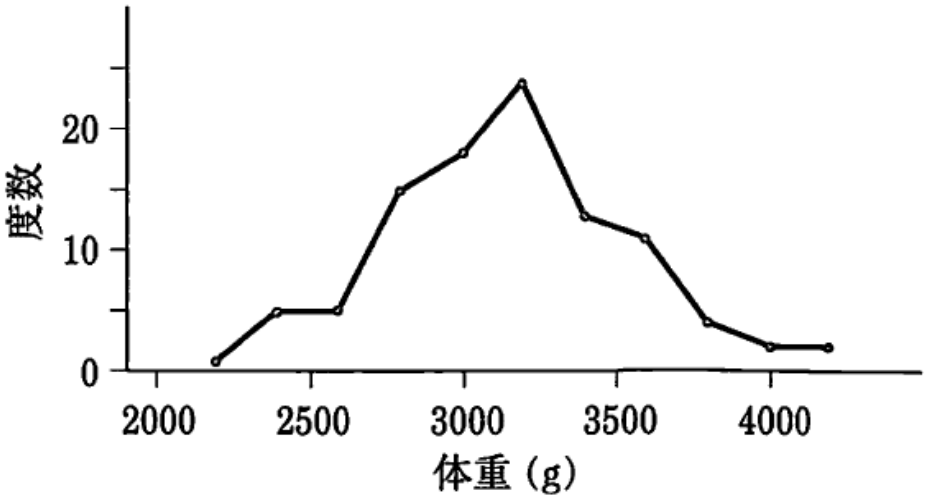
階級番号	階級	階級値	度数
1	2100~2300	2200	1
2	2300~2500	2400	5
3	2500~2700	2600	5
4	2700~2900	2800	15
5	2900~3100	3000	18
6	3100~3300	3200	24
7	3300~3500	3400	13
8	3500~3700	3600	11
9	3700~3900	3800	4
10	3900~4100	4000	2
11	4100~4300	4200	2
計	ー	ー	100

可視化

ヒストグラム



度数多角形



データから特徴を取り出す：データの代表値

生データ

$x_1, x_2, \cdots, x_n$　　これは単なる（長大な）数列

度数データ

度数分布表

階級値	c_1	c_2	$\cdots$	c_j	$\cdots$	c_k	計
度数	f_1	f_2	$\cdots$	f_j	$\cdots$	f_k	n
相対度数	p_1	p_2	$\cdots$	p_j	$\cdots$	p_k	1

相対度数：　$p_j = \dfrac{f_j}{n}$

データの代表値

平均値 (mean/average)

明らかなときは
i の範囲を省略

生データの場合：　$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum x_i$

度数データの場合：

階級値	c_1	c_2	…	c_j	…	c_k	計
度数	f_1	f_2	…	f_j	…	f_k	n
相対度数	p_1	p_2	…	p_j	…	p_k	1

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} c_j f_j = \frac{1}{n}\sum c_j f_j = \sum c_j p_j$$

中央値 (median)

生データの場合：データを小さいほうから並べて順位が中央の値

度数データの場合：多少の計算を要す（教科書参照）

最頻値 (mode)

度数データの場合のみ：最大度数をとる階級値

例　題　　次の10個のデータの平均値と中央値を求めよ.

2　5　6　7　8　3　1　2　4　24

小さいほうから並べると　　1　2　2　3　4　5　6　7　8　24

$$\bar{x} = \frac{1}{10}(1 + \cdots + 24) = 6.2$$

$$Me = \frac{1}{2}(4 + 5) = 4.5$$

データのばらつき・広がり

注意（後出）不偏分散

分散 (variance)　$s^2 = s_x^2$

生データの場合：　$s^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$

度数データの場合：

階級値	c_1	c_2	…	c_j	…	c_k	計
度数	f_1	f_2	…	f_j	…	f_k	n
相対度数	p_1	p_2	…	p_j	…	p_k	1

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} (c_j - \bar{x})^2 f_j = \sum_{j=1}^{k} (c_j - \bar{x})^2 p_j$$

標準偏差 (standard variation)　$s = \sqrt{s^2}$

基本的な性質（１）

偏差の和はゼロである：

偏差

$$\sum_{i=1}^{n} \overbrace{(x_i - \bar{x})} = 0$$

証 明　平均値の定義　$\bar{x} = \frac{1}{n}\sum x_i$　から　$\sum x_i = n\bar{x}$

そうすると，

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

基本的な性質（２）分散公式

分散は2乗の平均値から

平均値の2乗を引いたもの

$$s^2 = \overline{x^2} - \bar{x}^2$$

生データの場合：　$s^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2$

度数データの場合：　$s^2 = \frac{1}{n}\sum_{j=1}^{k} {c_j}^2 f_j - \bar{x}^2$

証 明　生データの場合（度数データの場合は自分で確認）

平均値 $\bar{x}$　　$\bar{x}^2$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) = \frac{1}{n}\sum x_i^2 - 2\bar{x}\frac{1}{n}\sum x_i + \frac{1}{n}\sum \bar{x}^2$$

$$= \frac{1}{n}\sum {x_i}^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

例　題　　次の度数分布表から平均, 分散, 標準偏差を求めよ.

階級値 (c_j)	22	24	26	28	30	32	34	36	38	40	42	計
度 数 (f_j)	1	5	5	15	18	24	13	11	4	2	2	100

例　題　　次の度数分布表から平均, 分散, 標準偏差を求めよ.

階級値 (c_j)	22	24	26	28	30	32	34	36	38	40	42	計
度 数 (f_j)	1	5	5	15	18	24	13	11	4	2	2	100
$c_j f_j$	22	120	130	420	540	768	442	396	152	80	84	3154
c_j^2	484	576	676	784	900	1024	1156	1296	1444	1600	1764	
$c_j^2 f_j$	484	2880	3380	11760	16200	24576	15028	14256	5776	3200	3528	101068

$$\bar{x} = \frac{1}{n}\sum c_j f_j = \frac{1}{100} \times 3154 = 31.54$$

$$\overline{x^2} = \frac{1}{n}\sum c_j^2 f_j = \frac{1}{100} \times 101068 = 1010.68$$

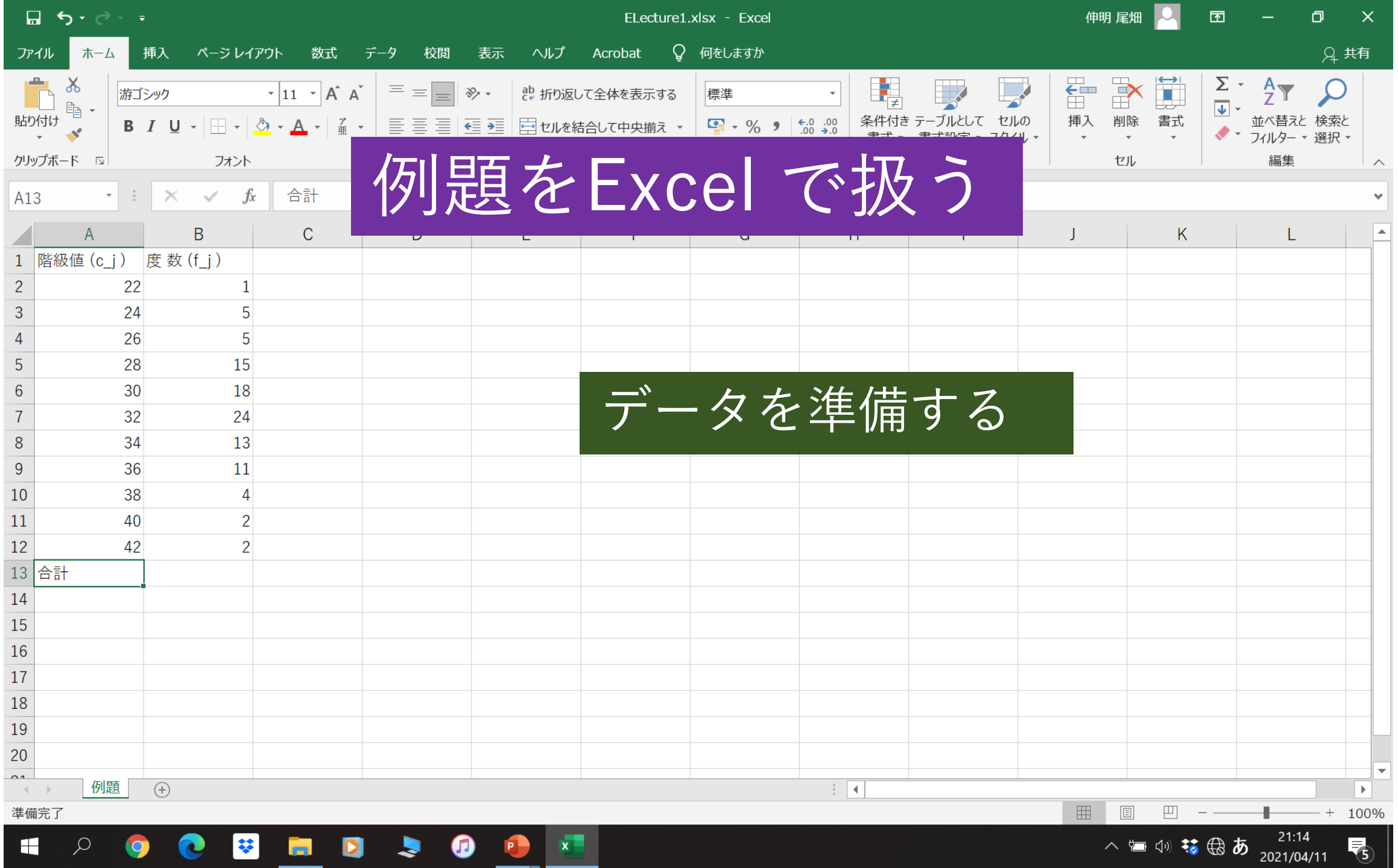
$$s^2 = \overline{x^2} - \bar{x}^2$$

$$= 1010.68 - 31.54^2 = 15.91$$

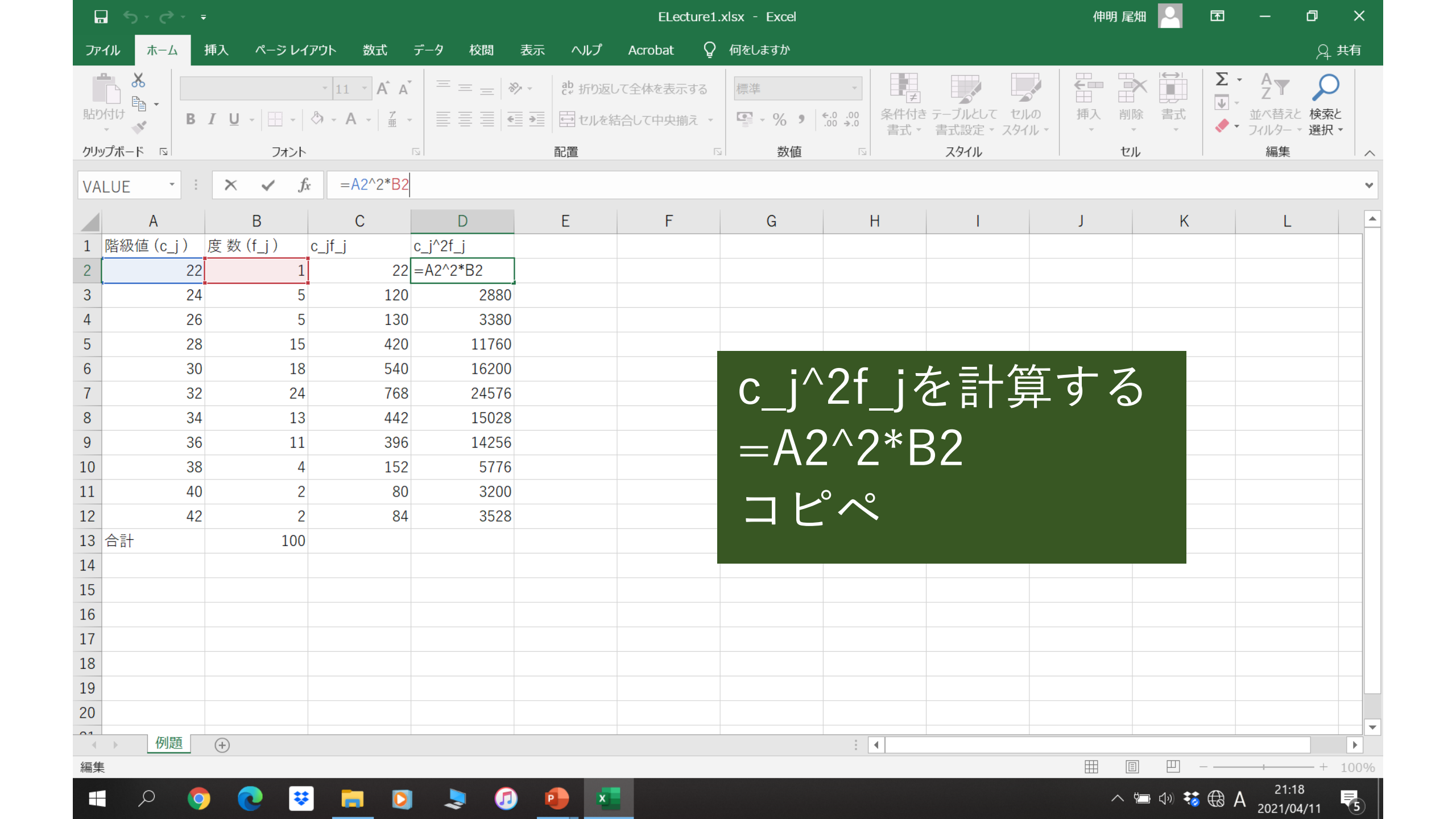
$$s = \sqrt{15.91} = 3.99$$

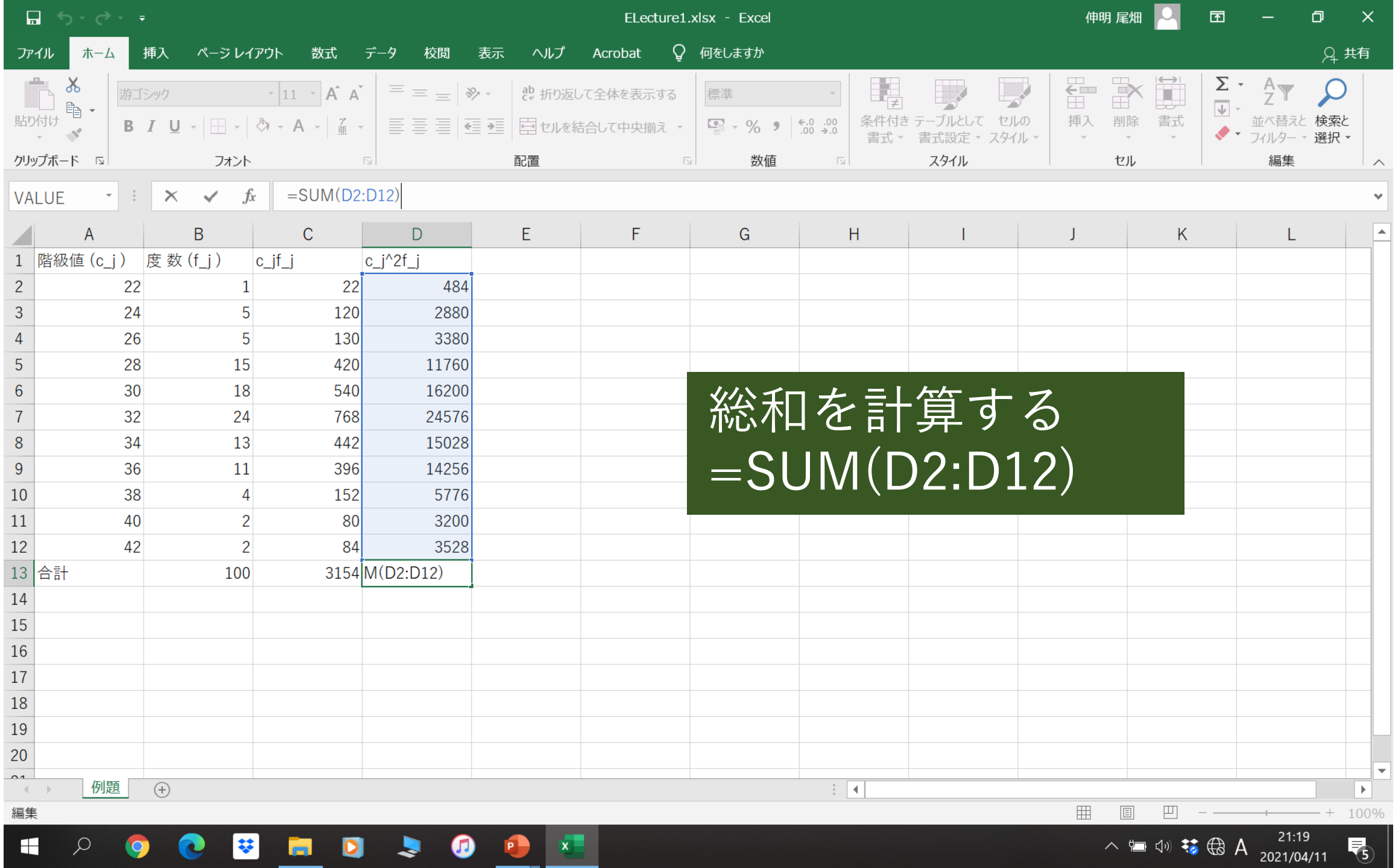
例題をExcel で扱う

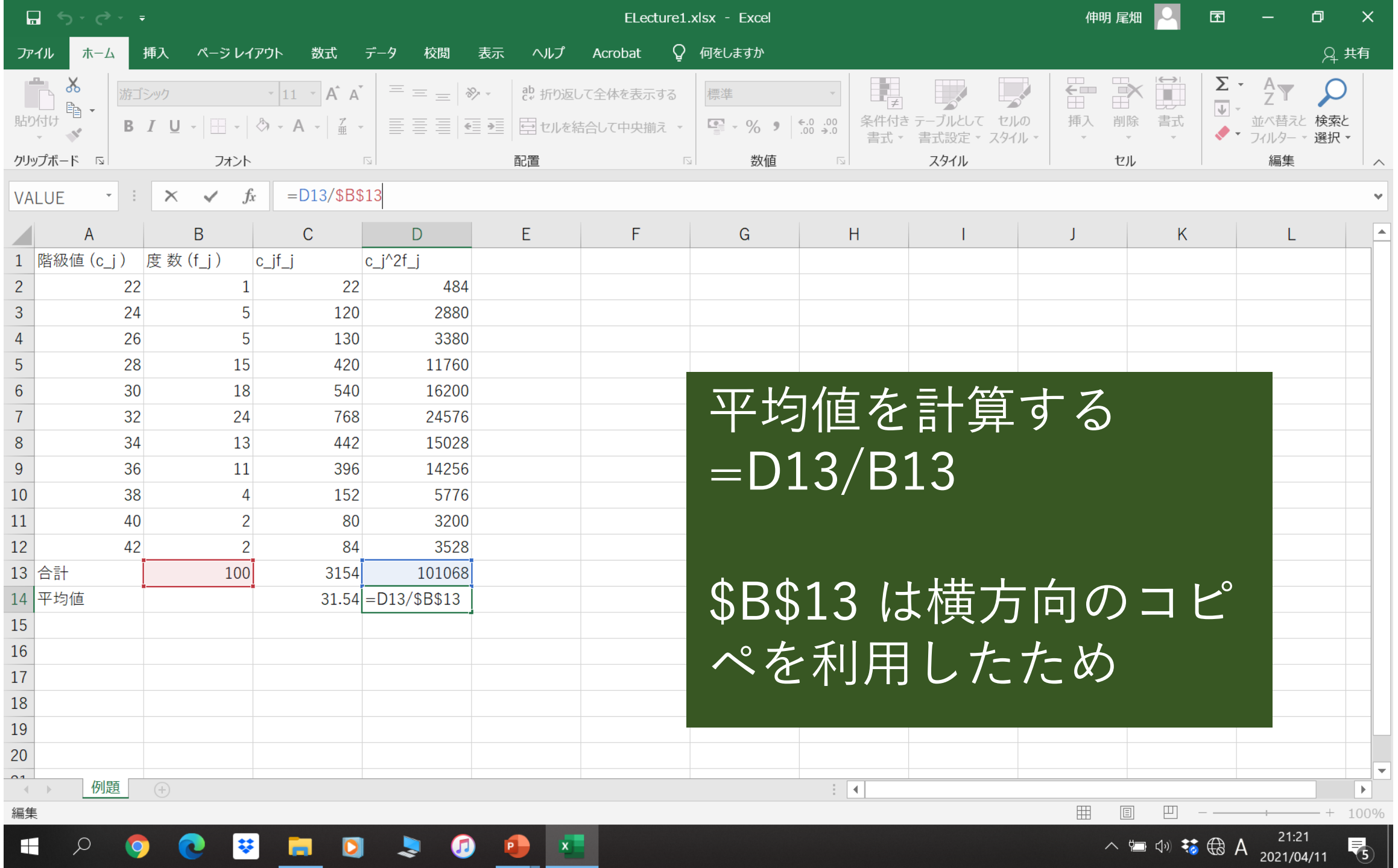
データを準備する

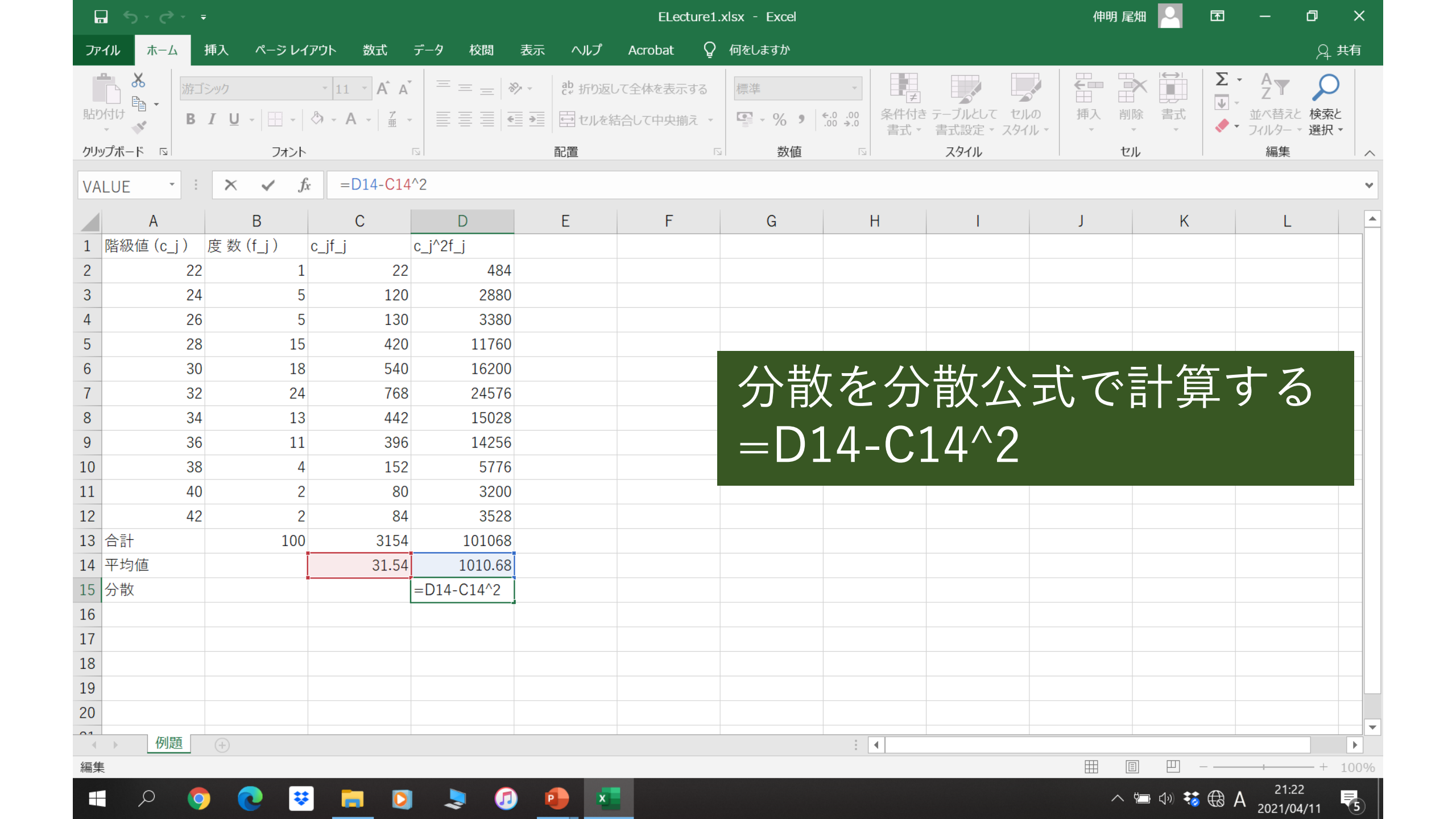


	A	B
1	階級値 (c_j)	度 数 (f_j)
2	22	1
3	24	5
4	26	5
5	28	15
6	30	18
7	32	24
8	34	13
9	36	11
10	38	4
11	40	2
12	42	2
13	合計	

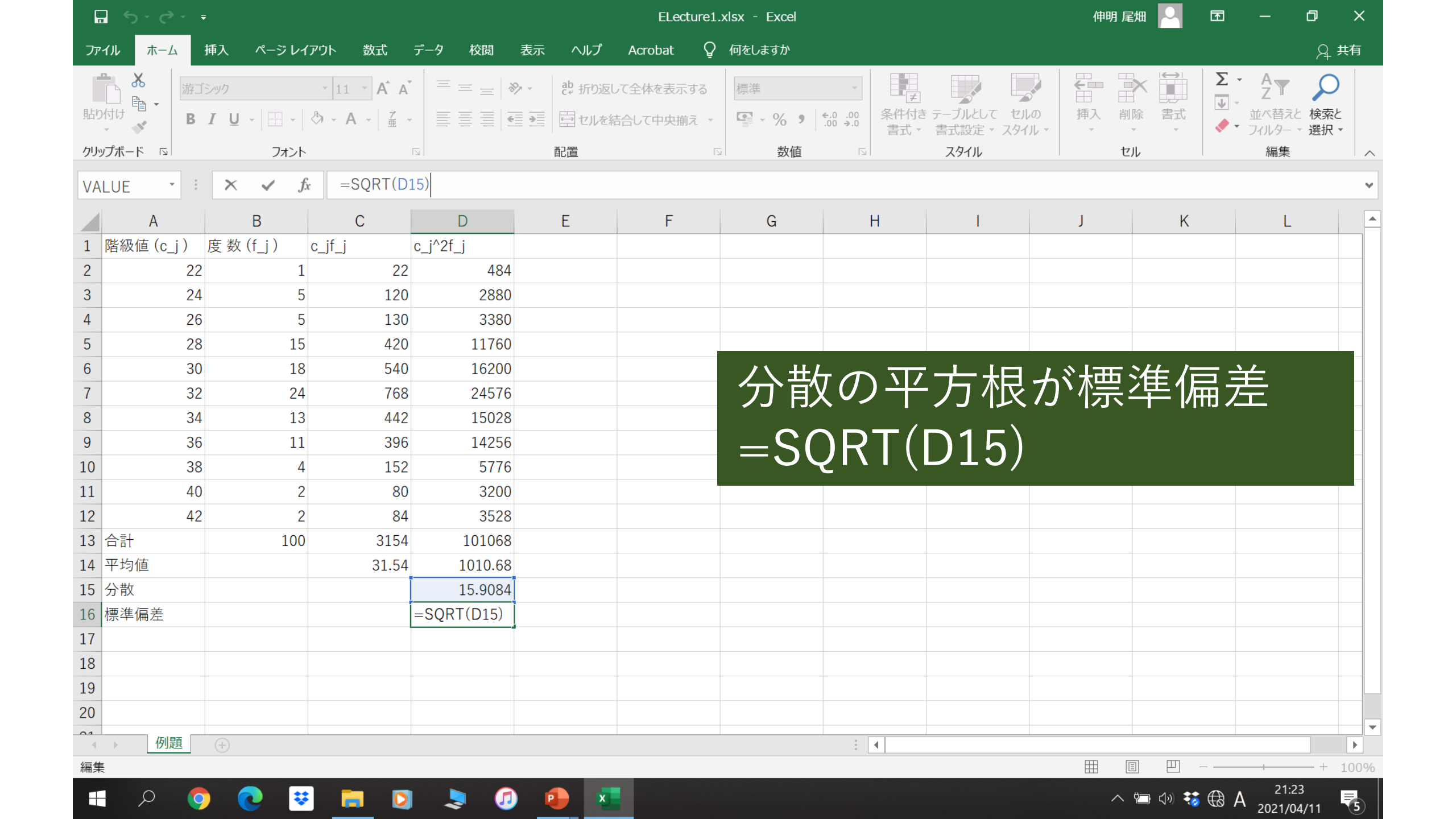

ELecture1.xlsx - Excel
伸明 尾畑
ファイル
ホーム
挿入
ページ レイアウト
数式
データ
校閲
表示
ヘルプ
Acrobat
何をしますか
共有
貼り付け
クリップボード
フォント
折り返して全体を表示する
セルを結合して中央揃え
配置
標準
数値
条件付き書式
テーブルとして書式設定
セルのスタイル
スタイル
挿入
削除
書式
セル
並べ替えとフィルター
検索と選択
編集
VALUE
=A2^2*B2
階級値 (c_j)
度 数 (f_j)
c_jf_j
c_j^2f_j
22 1 22 =A2^2*B2
24 5 120 2880
26 5 130 3380
28 15 420 11760
30 18 540 16200
32 24 768 24576
34 13 442 15028
36 11 396 14256
38 4 152 5776
40 2 80 3200
42 2 84 3528
合計 100
例題
編集
100%
c_j^2f_jを計算する
=A2^2*B2
コピペ
21:18
2021/04/11


ELecture1.xlsx - Excel
伸明 尾畑
ファイル
ホーム
挿入
ページ レイアウト
数式
データ
校閲
表示
ヘルプ
Acrobat
何をしますか
共有
貼り付け
クリップボード
フォント
配置
数値
スタイル
セル
編集
VALUE
=SUM(D2:D12)
階級値 (c_j)
度 数 (f_j)
c_jf_j
c_j^2f_j
22 1 22 484
24 5 120 2880
26 5 130 3380
28 15 420 11760
30 18 540 16200
32 24 768 24576
34 13 442 15028
36 11 396 14256
38 4 152 5776
40 2 80 3200
42 2 84 3528
合計 100 3154 M(D2:D12)
総和を計算する
=SUM(D2:D12)
例題
編集
100%
21:19
2021/04/11


ELecture1.xlsx - Excel
伸明 尾畑
ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 ヘルプ Acrobat 何をしますか 共有
貼り付け クリップボード 游ゴシック 11 フォント 折り返して全体を表示する セルを結合して中央揃え 配置 標準 数値 条件付き書式 テーブルとして書式設定 セルのスタイル スタイル 挿入 削除 書式 セル 並べ替えとフィルター 検索と選択 編集
VALUE
=D13/$B$13
A 階級値 (c_j) B 度 数 (f_j) C c_jf_j D c_j^2f_j
1 階級値 (c_j) 度 数 (f_j) c_jf_j c_j^2f_j
2 22 1 22 484
3 24 5 120 2880
4 26 5 130 3380
5 28 15 420 11760
6 30 18 540 16200
7 32 24 768 24576
8 34 13 442 15028
9 36 11 396 14256
10 38 4 152 5776
11 40 2 80 3200
12 42 2 84 3528
13 合計 100 3154 101068
14 平均値 31.54 =D13/$B$13
15
16
17
18
19
20
平均値を計算する
=D13/B13
$B$13 は横方向のコピペを利用したため
例題
編集
100%
21:21
2021/04/11



	A	B	C	D	E	F	G	H	I	J	K	L
1	階級値（c_j）	度 数（f_j）	c_jf_j	c_j^2f_j								
2	22	1	22	484								
3	24	5	120	2880								
4	26	5	130	3380								
5	28	15	420	11760								
6	30	18	540	16200								
7	32	24	768	24576								
8	34	13	442	15028								
9	36	11	396	14256								
10	38	4	152	5776								
11	40	2	80	3200								
12	42	2	84	3528								
13	合計	100	3154	101068								
14	平均値		31.54	1010.68								
15	分散			=D14-C14^2								
16												
17												
18												
19												
20												



	A	B	C	D
1	階級値 (c_j)	度 数 (f_j)	c_jf_j	c_j^2f_j
2	22	1	22	484
3	24	5	120	2880
4	26	5	130	3380
5	28	15	420	11760
6	30	18	540	16200
7	32	24	768	24576
8	34	13	442	15028
9	36	11	396	14256
10	38	4	152	5776
11	40	2	80	3200
12	42	2	84	3528
13	合計	100	3154	101068
14	平均値		31.54	1010.68
15	分散			15.9084
16	標準偏差			=SQRT(D15)
17				
18				
19				
20				

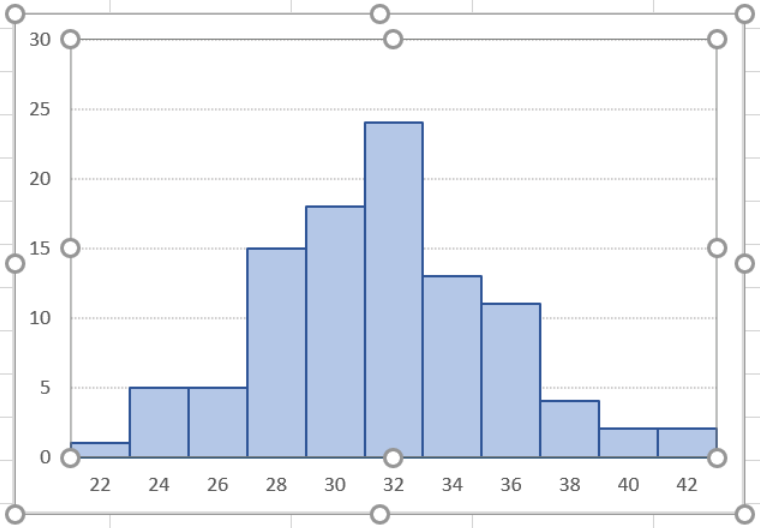
ELecture1.xlsx - Excel　　グラフ ツール　　伸明 尾畑

ファイル　ホーム　挿入　ページ レイアウト　数式　データ　校閲　表示　ヘルプ　Acrobat　デザイン　書式　何をしますか　共有

グラフ要素を追加　クイック レイアウト　色の変更　グラフのレイアウト　グラフ スタイル　行/列の切り替え　データの選択　データ　グラフの種類の変更　種類　グラフの移動　場所

グラフ 1

	A	B	C	D
1	階級値 (c_j)	度 数 (f_j)	c_jf_j	c_j^2f_j
2	22	1	22	484
3	24	5	120	2880
4	26	5	130	3380
5	28	15	420	11760
6	30	18	540	16200
7	32	24	768	24576
8	34	13	442	15028
9	36	11	396	14256
10	38	4	152	5776
11	40	2	80	3200
12	42	2	84	3528
13	合計	100	3154	101068
14	平均値		31.54	1010.68
15	分散			15.9084
16	標準偏差			3.988533565
17				
18				
19				
20				



ヒストグラム
縦棒グラフを選択して整形

例題

準備完了　100%

21:28
2021/04/11

ELecture1.xlsx - Excel

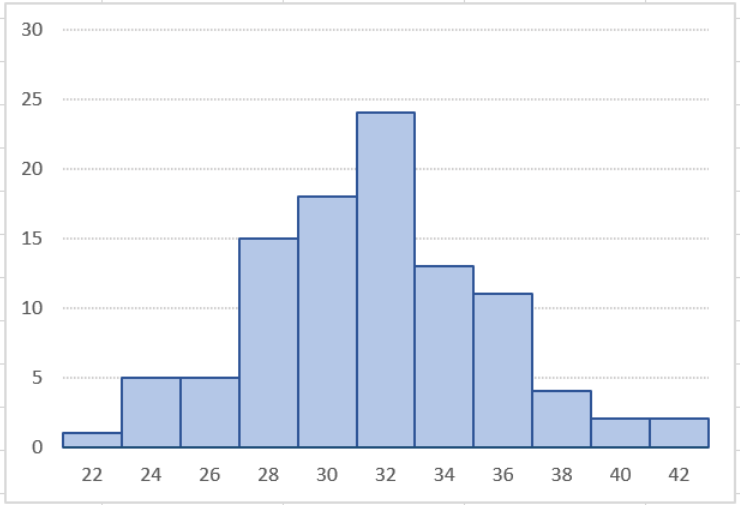
伸明 尾畑

ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 ヘルプ Acrobat 何をしますか 共有

貼り付け クリップボード 游ゴシック 11 フォント 折り返して全体を表示する セルを結合して中央揃え 配置 標準 数値 条件付き書式 テーブルとして書式設定 セルのスタイル スタイル 挿入 削除 書式 セル 並べ替えとフィルター 検索と選択 編集

A17

	A	B	C	D
1	階級値（c_j）	度 数（f_j）	c_jf_j	c_j^2f_j
2	22	1	22	484
3	24	5	120	2880
4	26	5	130	3380
5	28	15	420	11760
6	30	18	540	16200
7	32	24	768	24576
8	34	13	442	15028
9	36	11	396	14256
10	38	4	152	5776
11	40	2	80	3200
12	42	2	84	3528
13	合計	100	3154	101068
14	平均値		31.54	1010.68
15	分散			15.9084
16	標準偏差			3.988533565
17				
18				
19				
20				



解答完了

例題

準備完了

100%

21:32
2021/04/11

Lecture 1

おわり

Lecture 2

2 変量データの整理

2変量(2次元)データの例

新生児の身長，体重のデータ

番号	身長	体重
1	46.0	2700
2	49.5	3220
3	50.0	3360
4	50.0	3500
5	49.0	3120
6	50.0	3160
7	53.0	4150
8	48.0	3310
9	49.0	2880
10	50.5	3090
11	49.5	3020
12	49.0	3360
13	50.0	3110
14	50.0	3560
15	47.5	2990

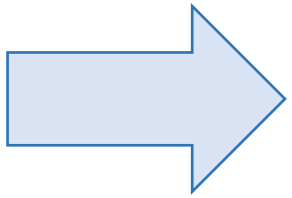
番号	身長	体重
16	50.5	3440
17	48.0	2920
18	49.0	3060
19	49.0	3360
20	50.0	3400
21	48.0	3200
22	50.5	2940
23	48.5	2850
24	50.5	3220
25	48.5	2750
26	49.0	3020
27	48.5	2570
28	48.5	3030
29	45.0	2410
30	51.0	3280

番号	身長	体重
31	50.5	3140
32	49.0	3040
33	52.0	3910
34	50.0	2770
35	46.5	2340
36	50.0	3140
37	50.5	3560
38	50.0	3390
39	50.0	3420
40	51.0	3450
41	49.5	3590
42	48.5	2830
43	48.0	3120
44	51.0	3190
45	50.0	3600

番号	身長	体重
46	47.0	2980
47	50.0	3090
48	51.0	3630
49	53.0	4060
50	50.0	3720
51	50.0	3400
52	50.5	3430
53	51.0	3250
54	48.0	2760
55	50.0	3320
56	49.0	2930
57	50.0	3320
58	48.0	2620
59	47.5	2860
60	48.0	2530

生データ ⇨ 相関表（2次元の度数分布表）

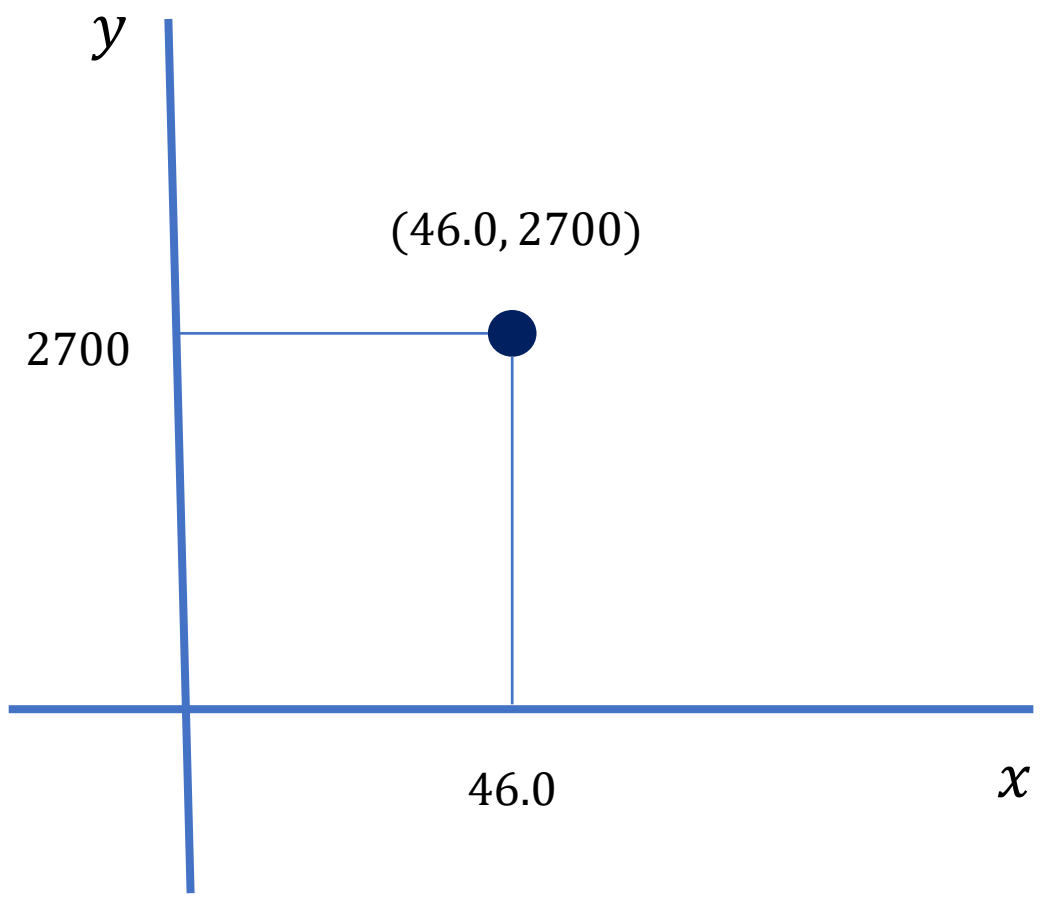
番号	身長 x	体重 y
1	46.0	2700
2	49.5	3220
3	50.0	3360
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
60	48.0	2530



y ＼ x	45	46	47	48	49	50	51	52	53	計
2400	1	1	0	0	0	0	0	0	0	2
2600	0	1	0	3	0	0	0	0	0	4
2800	0	0	1	4	1	1	0	0	0	7
3000	0	0	2	2	5	3	0	0	0	12
3200	0	0	0	2	2	5	3	0	0	12
3400	0	0	0	1	2	10	1	0	0	14
3600	0	0	0	0	1	3	1	0	0	5
3800	0	0	0	0	0	1	0	0	0	1
4000	0	0	0	0	0	0	0	1	1	2
4200	0	0	0	0	0	0	0	0	1	1
計	1	2	3	12	11	23	5	1	2	60

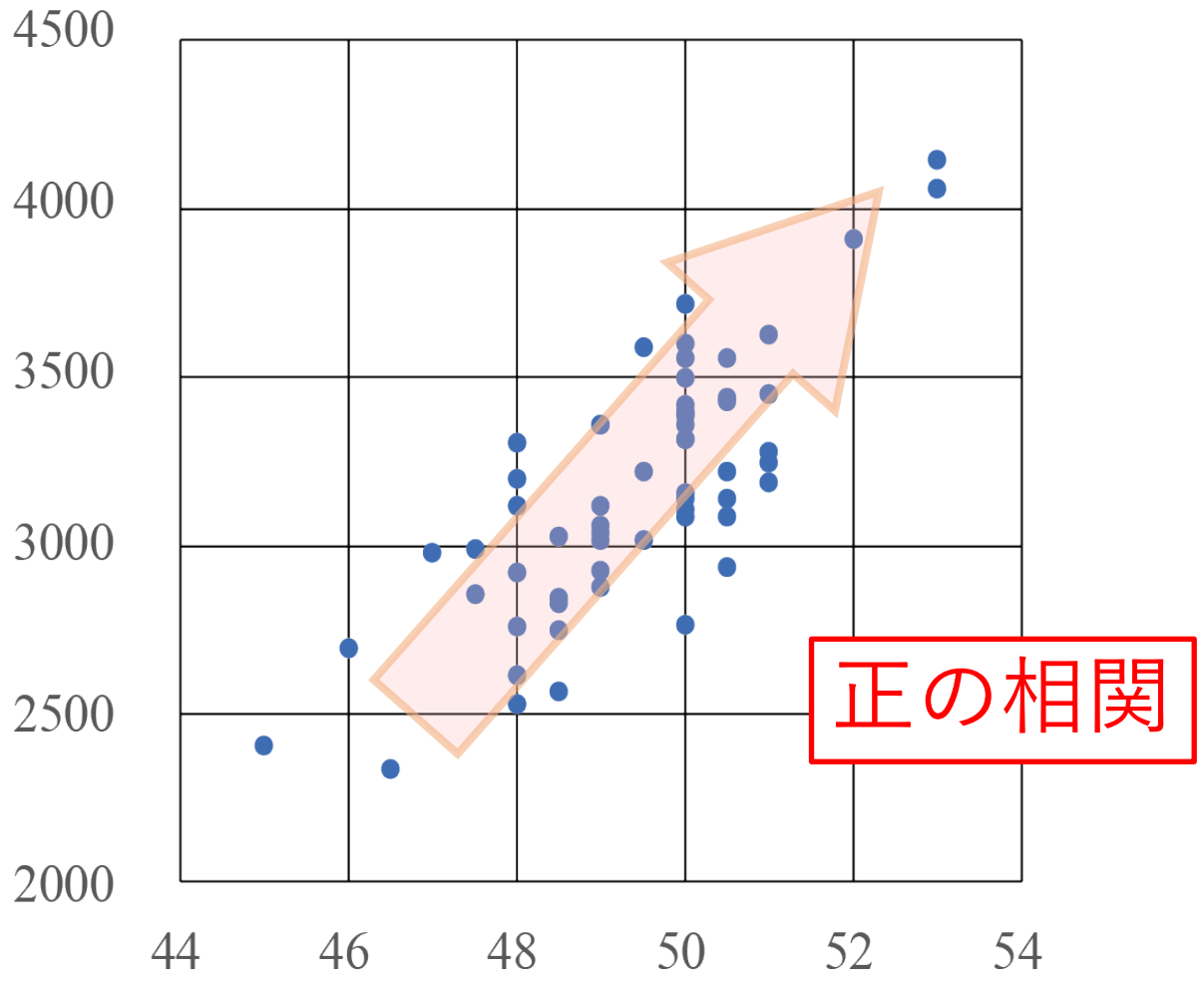
生データ ⇨ 散布図(Scatter Plot)

番号	身長 x 座標	体重 y 座標
1	46.0	2700
2	49.5	3220
3	50.0	3360
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
60	48.0	2530



生データ ⇨ 散布図(Scatter Plot)

番号	身長 x 座標	体重 y 座標
1	46.0	2700
2	49.5	3220
3	50.0	3360
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
60	48.0	2530



2 変量の統計量

番号	身長 x	体重 y
1	46.0	2700
2	49.5	3220
3	50.0	3360
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
60	48.0	2530

n：データの大きさ

x の平均値と分散

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

y の平均値と分散

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

x と y の共分散

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

x と y の相関係数

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

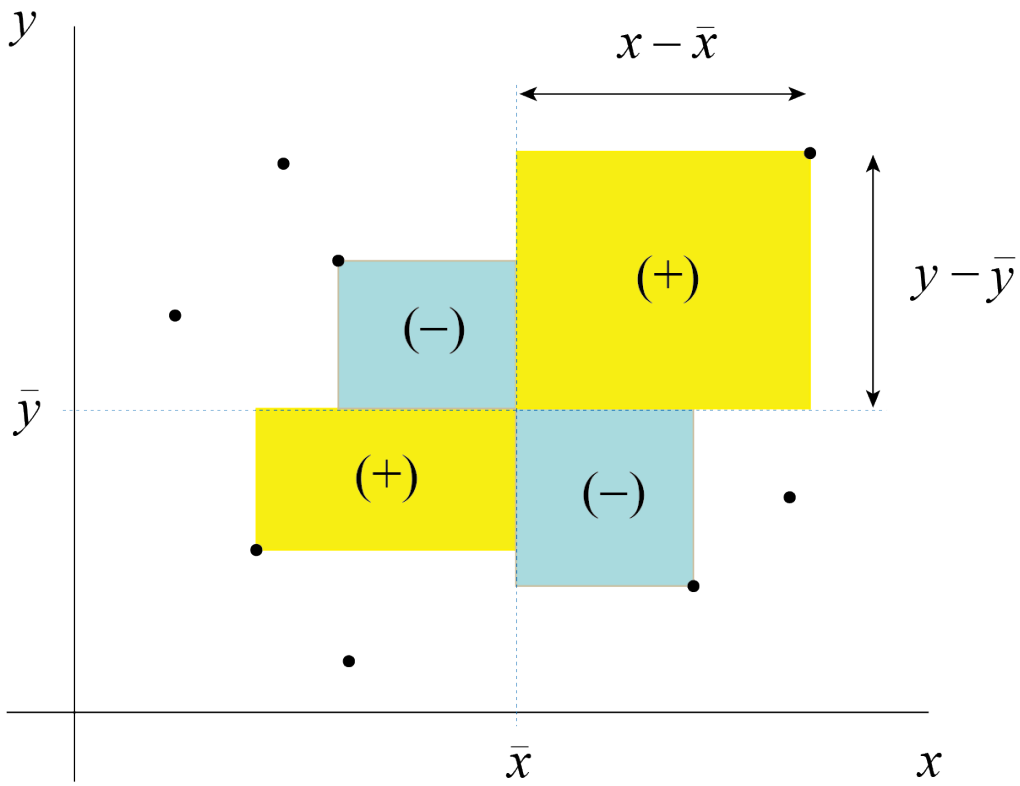
共分散・相関係数の意味

共分散

$$s_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$s_{xy} > 0$：概ね右上がりの傾向

$s_{xy} < 0$：概ね右下がりの傾向



相関係数

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n}\sum \frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

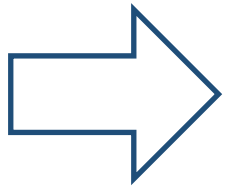
標準化された共分散といえる

変量の標準化（規準化）

生データ：　$x_1, \cdots, x_i, \cdots, x_n$

平均値：　$\bar{x}$

標準偏差：　$s = s_x$



標準化または規準化（z 変換）

$$z_i = \frac{x_i - \bar{x}}{s}$$

定 理　標準化された変数　$z_i = \dfrac{x_i - \bar{x}}{s}$　に対して

平均値 $= \bar{z} = 0$

標準偏差 $= s_z = 1$

定 理　標準化された変数　$z_i = \dfrac{x_i - \bar{x}}{s}$　に対して

平均値 $= \bar{z} = 0$

標準偏差 $= s_z = 1$

証 明　平均値の定義より

偏差の和はゼロ

$$\bar{z} = \frac{1}{n}\sum z_i = \frac{1}{n}\sum \frac{x_i - \bar{x}}{s} = \frac{1}{s}\frac{1}{n}\overbrace{\sum (x_i - \bar{x})} = \frac{1}{s}\left(\frac{1}{n}\sum x_i - \frac{1}{n}\sum \bar{x}\right)$$

$$= \frac{1}{s}(\bar{x} - \bar{x})$$

$$= 0$$

定 理　標準化された変数　$z_i = \dfrac{x_i - \bar{x}}{s}$　に対して

平均値 $= \bar{z} = 0$

標準偏差 $= s_z = 1$

証 明　分散の定義より

$$s_z^2 = \frac{1}{n}\sum (z_i - \bar{z})^2 = \frac{1}{n}\sum z_i^2 = \frac{1}{n}\sum \frac{(x_i - \bar{x})^2}{s^2} = \frac{1}{s^2} \times \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$= \frac{1}{s^2} \times s^2$$

$$= 1$$

変量の標準化（規準化）

定 理　標準化された変数

$$z_i = \frac{x_i - \bar{x}}{s}$$

に対して

平均値 $= \bar{z} = 0$

標準偏差 $= s_z = 1$

注 目　標準化によって，測定の内容や測定単位が異 なるデータの組を比較できる．

身長 x を cm で測定する

x_i　の単位は cm

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$　の単位は cm

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$　の単位は cm^2

s_x　の単位は cm

標準化された変数 z

$$z_i = \frac{x_i^{\text{cm}} - \bar{x}^{\text{cm}}}{s^{\text{cm}}}$$　は単位をもたない

標準化すれば，たとえば，身長のデータと体重のデータを比較することもできる

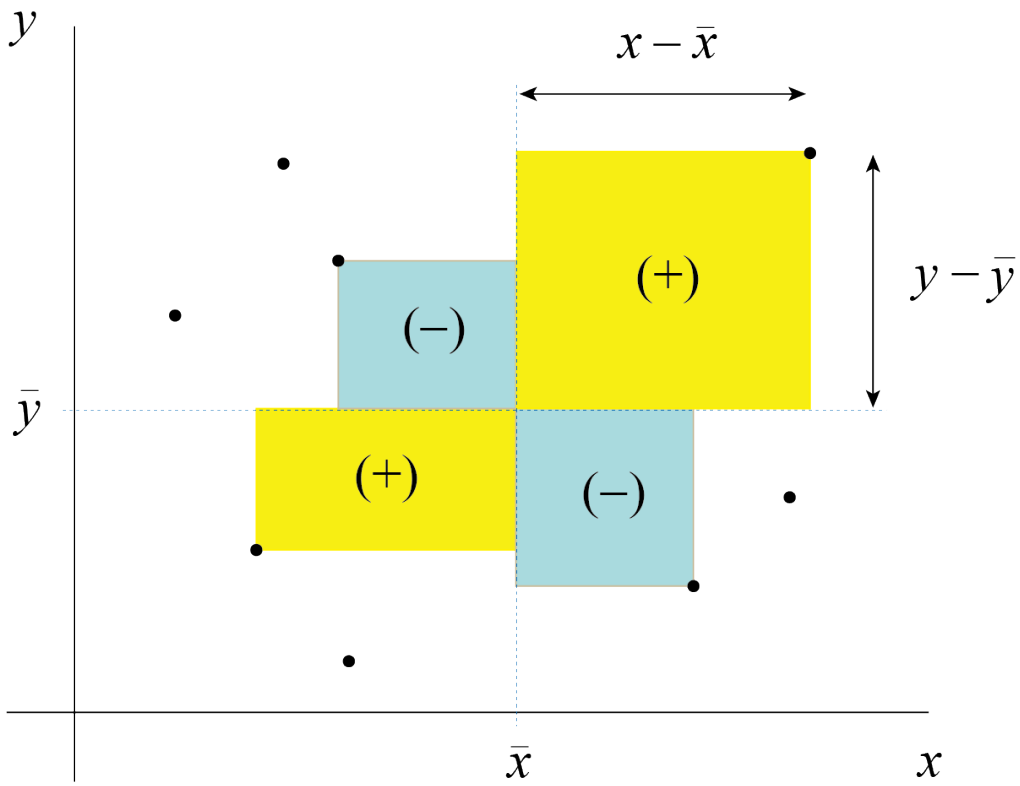
共分散・相関係数の意味

共分散

$$s_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$s_{xy} > 0$：概ね右上がりの傾向

$s_{xy} < 0$：概ね右下がりの傾向



相関係数

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n}\sum \frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

標準化された共分散といえる

納得!

定 理（相関係数の基本性質）

$$-1 \leq r_{xy} = r_{yx} \leq 1$$

証 明　すべての実数 t に対して　$\sum\{t(x_i - \bar{x}) + (y_i - \bar{y})\}^2 \geq 0$

$$t^2 \sum (x_i - \bar{x})^2 + 2t \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (y_i - \bar{y})^2 \geq 0$$

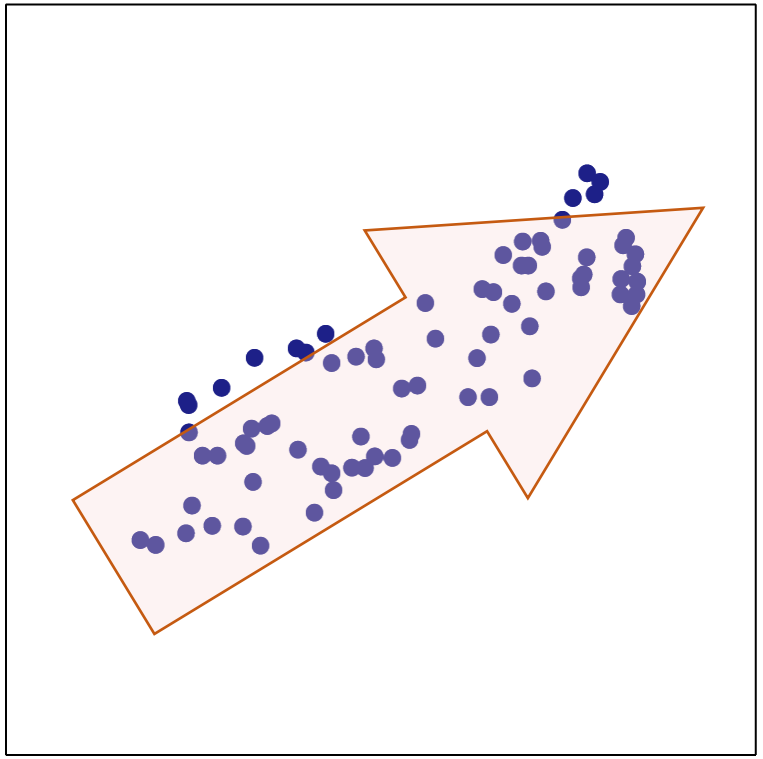
両辺を n で割って，　$t^2 s_x^2 + 2t s_{xy} + s_y^2 \geq 0$

これがすべての実数 t に対して成り立つので，判別式は　$\frac{D}{4} = s_{xy}^2 - s_x^2 s_y^2 \leq 0$

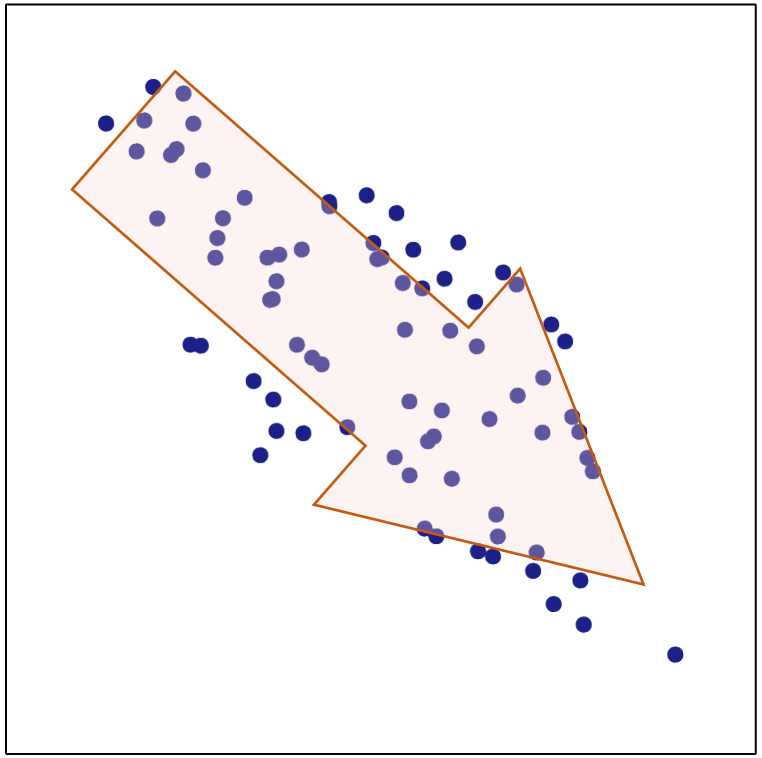
つまり，　$s_{xy}^2 \leq s_x^2 s_y^2$

相関係数は　$r_{xy} = \frac{s_{xy}}{s_x s_y}$　なので　$r_{xy}^2 \leq 1$　が得られ，　証明が終わる．

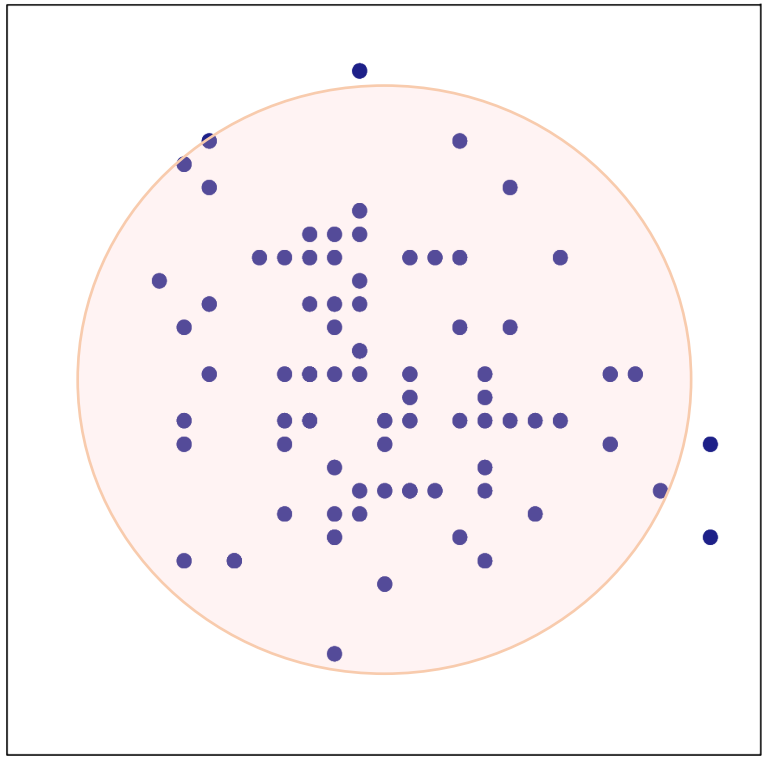
正の相関，負の相関，無相関



$r_{xy} > 0$：正の相関

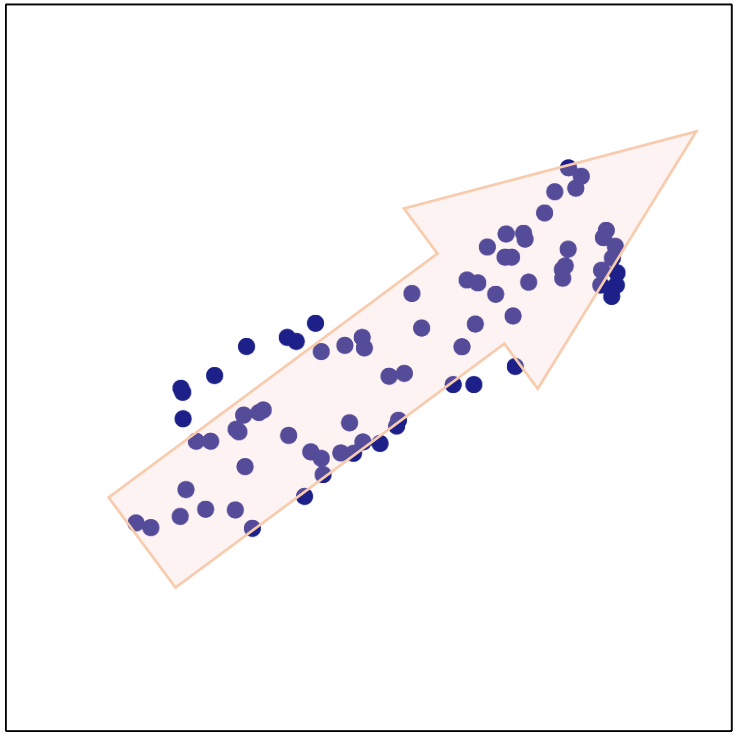


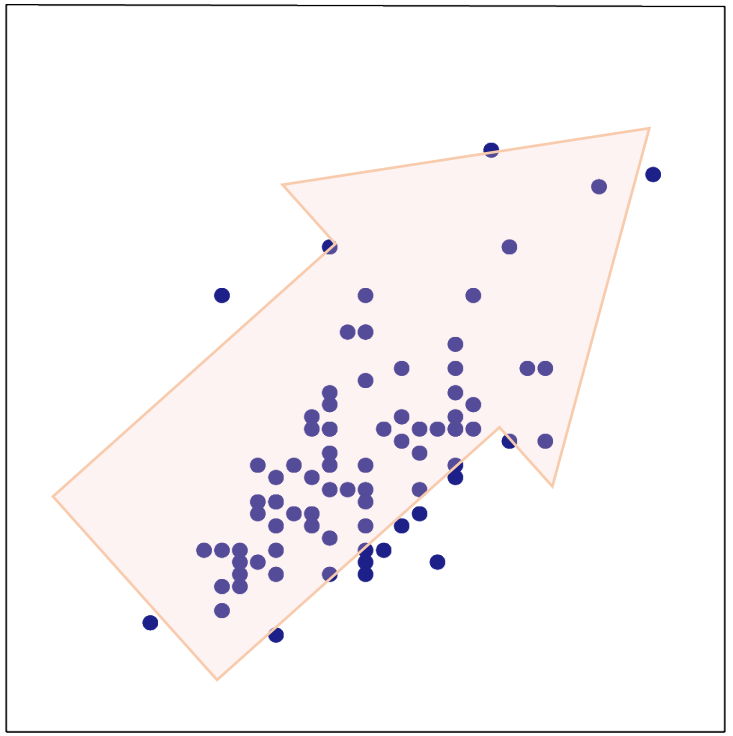
$r_{xy} < 0$：負の相関

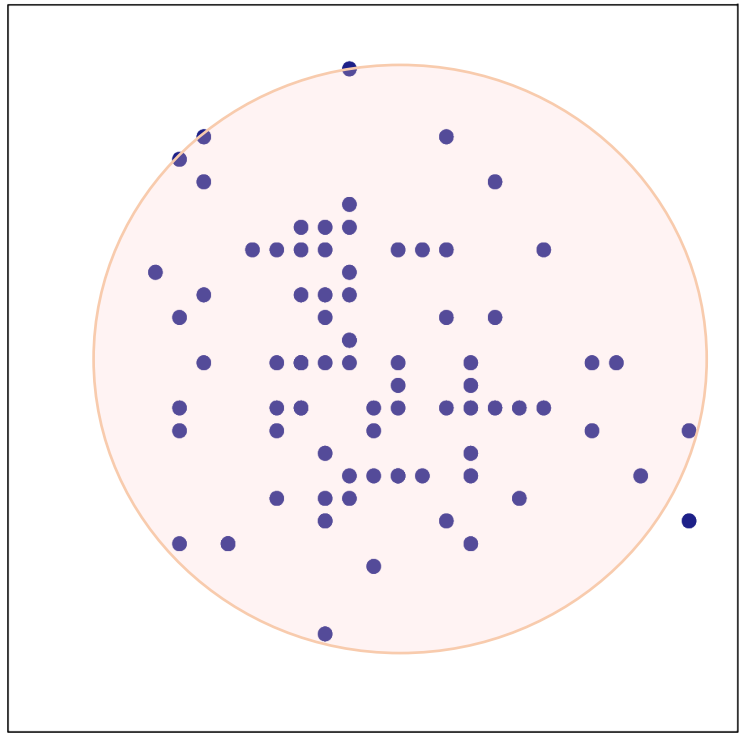


$|r_{xy}| \approx 0$：無相関

強い相関，弱い相関







$|r_{xy}|$

1　　0

強い相関　　弱い相関　　無相関

定理（共分散の公式）

共分散は積の平均値から平均値の積を引く

$$s_{xy} = \overline{xy} - \bar{x}\,\bar{y}$$

分散公式との関係

$$s_{xx} = \overline{xx} - \bar{x}\,\bar{x} = \overline{x^2} - \bar{x}^2 = s_x^2$$

証 明　共分散の定義より

$$s_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\,\frac{1}{n}\sum y_i - \bar{y}\,\frac{1}{n}\sum x_i + \frac{1}{n}\sum \bar{x}\bar{y}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y}$$

$$= \overline{xy} - \bar{x}\,\bar{y}$$

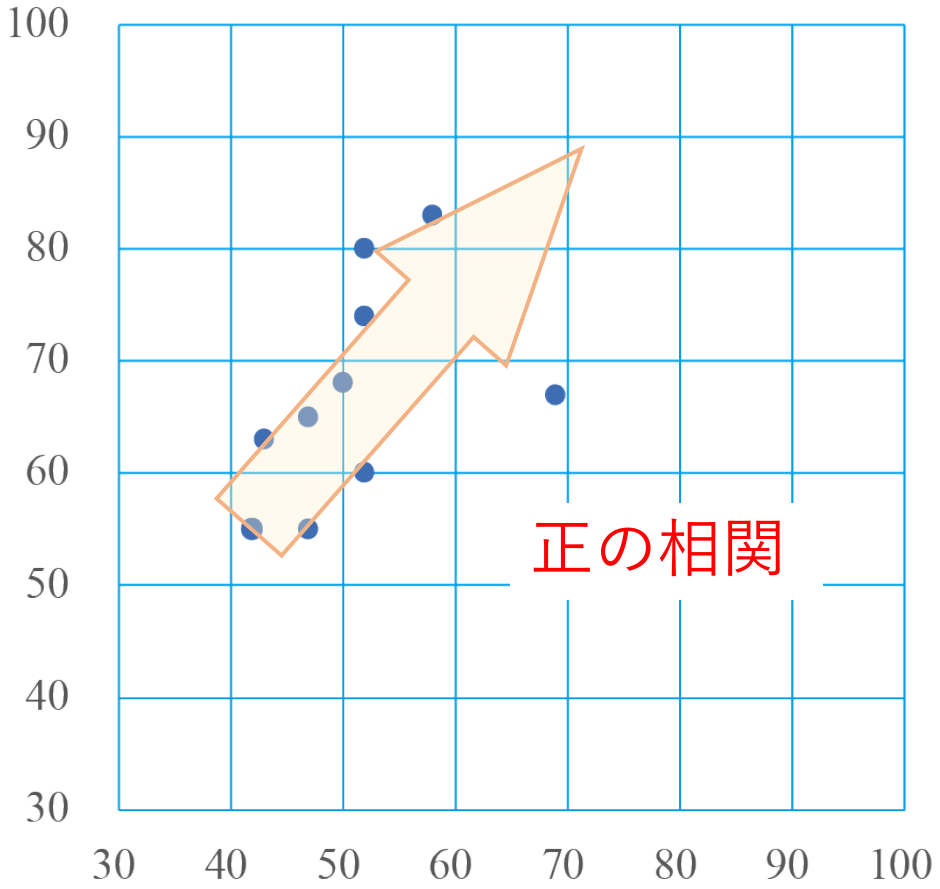
例題 2.1

受講生10名の中間試験と期末試験の結果は次の通りであった．中間試験と期末試験の得点について相関はあるか？

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

散布図



統計量の計算											合計
中間試験 (x)	50	58	52	52	43	47	52	69	47	42	512
x^2	2500	3364	2704	2704	1849	2209	2704	4761	2209	1764	26768
期末試験 (y)	68	83	74	80	63	55	60	67	65	55	670
y^2	4624	6889	5476	6400	3969	3025	3600	4489	4225	3025	45722
xy	3400	4814	3848	4160	2709	2585	3120	4623	3055	2310	34624

$$\bar{x} = \frac{1}{n}\sum x_j = 51.2$$

$$\bar{y} = \frac{1}{n}\sum y_j = 67.0$$

$$\overline{xy} = \frac{1}{n}\sum x_j y_j = 3462.4$$

$$\overline{x^2} = \frac{1}{n}\sum x_j^2 = 2676.8$$

$$\overline{y^2} = \frac{1}{n}\sum y_j^2 = 4572.2$$

$$s_{xy} = \overline{xy} - \bar{x}\cdot\bar{y}$$

$$= 3462.4 - 51.2 \times 67.0 = 32.0$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = 55.36$$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = 83.2$$

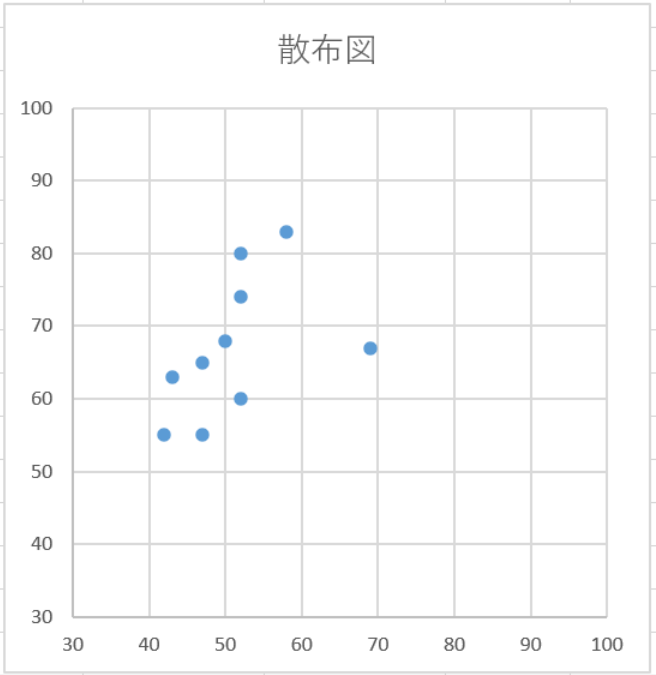
$$s_x = 7.44$$

$$s_y = 9.12$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.47$$

A19

	A	B	C	D	E	F
1	番号	中間試験 (x)	期末試験 (y)	x^2	y^2	xy
2	1	50	68	2500	4624	3400
3	2	58	83	3364	6889	4814
4	3	52	74	2704	5476	3848
5	4	52	80	2704	6400	4160
6	5	43	63	1849	3969	2709
7	6	47	55	2209	3025	2585
8	7	52	60	2704	3600	3120
9	8	69	67	4761	4489	4623
10	9	47	65	2209	4225	3055
11	10	42	55	1764	3025	2310
12	合計	512	670	26768	45722	34624
13	平均値	51.2	67	2676.8	4572.2	3462.4
14						
15	分散	55.36	83.2			
16	標準偏差	7.440430095	9.121403401			
17	共分散	32				
18	相関係数	0.471509312				
19						
20						



例題2.1

準備完了

100%

7:07
2021/04/12

ELecture2.xlsx - Excel

伸明 尾畑

ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 ヘルプ Acrobat 何をしますか 共有

貼り付け クリップボード フォント 配置 数値 スタイル セル 編集

VALUE =CORREL(B2:B11,C2:C11)

	A	B	C	D	E	F	G	H	I	J	K
1	番号	中間試験 (x)	期末試験 (y)	x^2	y^2	xy					
2	1	50	68	2500	4624	3400		統計コマンド			
3	2	58	83	3364	6889	4814		平均値	=average(B2:B11)	51.2	
4	3	52	74	2704	5476	3848		分散	=var.p(B2:B11)	55.36	
5	4	52	80	2704	6400	4160		標準偏差	=stdev.p(B2:B11)	7.440430095	
6	5	43	63	1849	3969	2709		共分散	=covariance.p(B2:B11,C2:C11)	32	
7	6	47	55	2209	3025	2585		相関係数	=correl(B2:B11,C2:C11)	2:B11,C2:C11)	
8	7	52	60	2704	3600	3120					
9	8	69	67	4761	4489	4623					
10	9	47	65	2209	4225	3055					
11	10	42	55	1764	3025	2310					
12	合計	512	670	26768	45722	34624					
13	平均値	51.2	67	2676.8	4572.2	3462.4					
14											
15	分散	55.36	83.2								
16	標準偏差	7.440430095	9.121403401								
17	共分散	32									
18	相関係数	0.471509312									
19											
20											

例題2.1

編集

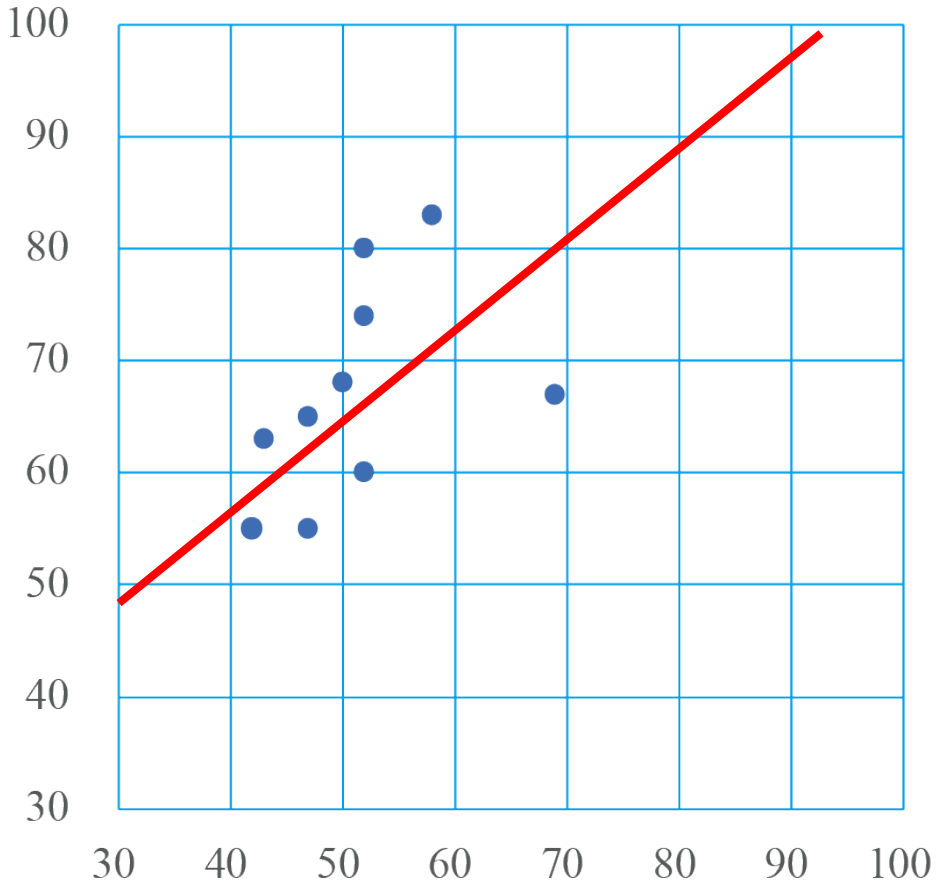
100%

7:16
2021/04/12

回帰分析

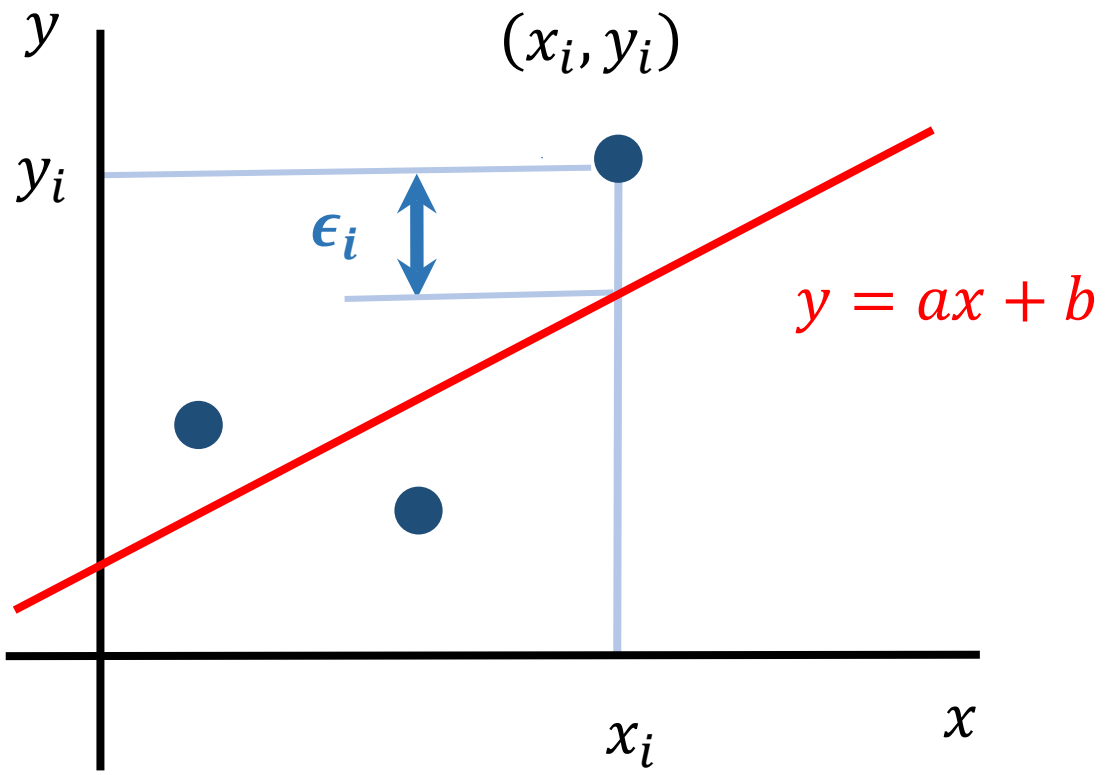
中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

散布図



データを最適な
1次関数で表現したい

最小２乗法



偏差 ϵ_i

$$y_i = ax_i + b + \epsilon_i$$

偏差平方和

$$Q = \sum \epsilon_i^2 = \sum (y_i - ax_i - b)^2$$

【最小２乗法】偏差平方和を最小にするように a, b を決める

- $Q = Q(a, b)$ は2次式なので、最小化は初等的にできる

線形回帰モデル（回帰直線）

$$\frac{\partial Q}{\partial a} = 2an(\sigma_x^2 + \bar{x}^2) - 2n(\sigma_{xy} + \bar{x}\bar{y}) + 2bn\bar{x}$$

$$\frac{\partial Q}{\partial b} = 2bn - 2n\bar{y} + 2an\bar{x}$$

連立方程式　$\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = 0$　を解いて

$$a_0 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy}s_y}{s_x}$$

$$b_0 = \bar{y} - a_0\bar{x}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$　（相関係数）

定 理　x を説明変数, y を目的変数とする線形回帰モデル（回帰直線）は

$$\frac{y - \bar{y}}{s_y} = r_{xy}\frac{x - \bar{x}}{s_x}$$

注 意　y を説明変数, x を目的変数とする線形回帰モデル（回帰直線）は

$$\frac{x - \bar{x}}{s_x} = r_{xy}\frac{y - \bar{y}}{s_y}$$

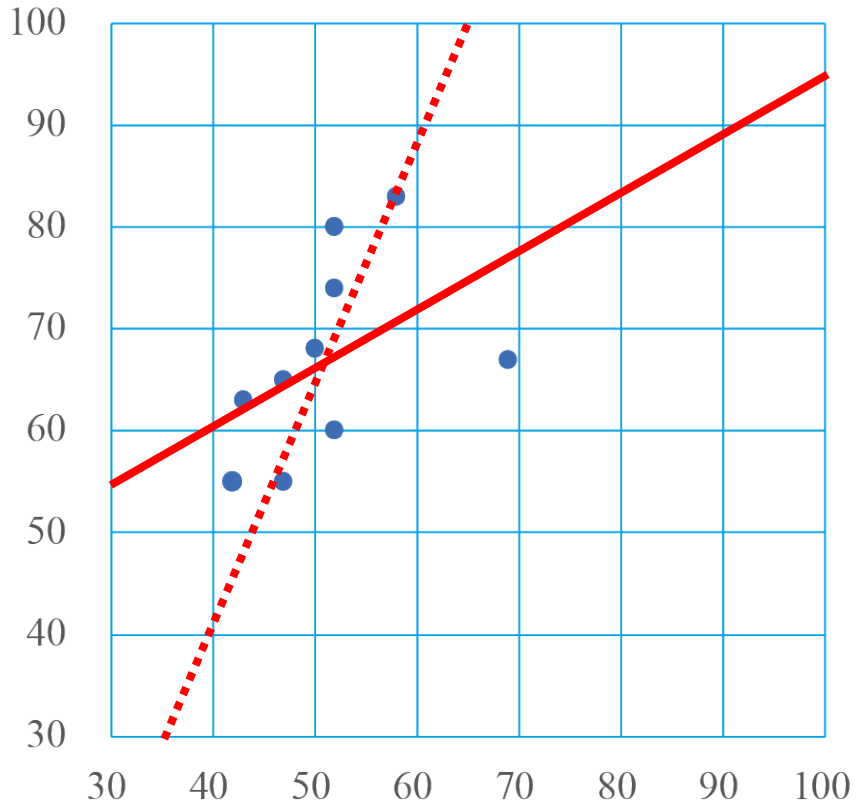
異なる

例題 2.2　受講生10名の中間試験と期末試験の結果から回帰直線を求めよ.

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

例題 2.2　受講生10名の中間試験と期末試験の結果から回帰直線を求めよ.

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55



$\bar{x} = 51.2$　　$\bar{y} = 67.0$

$s_x = 7.44$　　$s_y = 9.12$　　$r_{xy} = 0.47$

回帰直線（ x：説明変数）

$$\frac{y-\bar{y}}{s_y} = r_{xy}\frac{x-\bar{x}}{s_x} \quad \Longrightarrow \quad y = 0.58x + 37.3$$

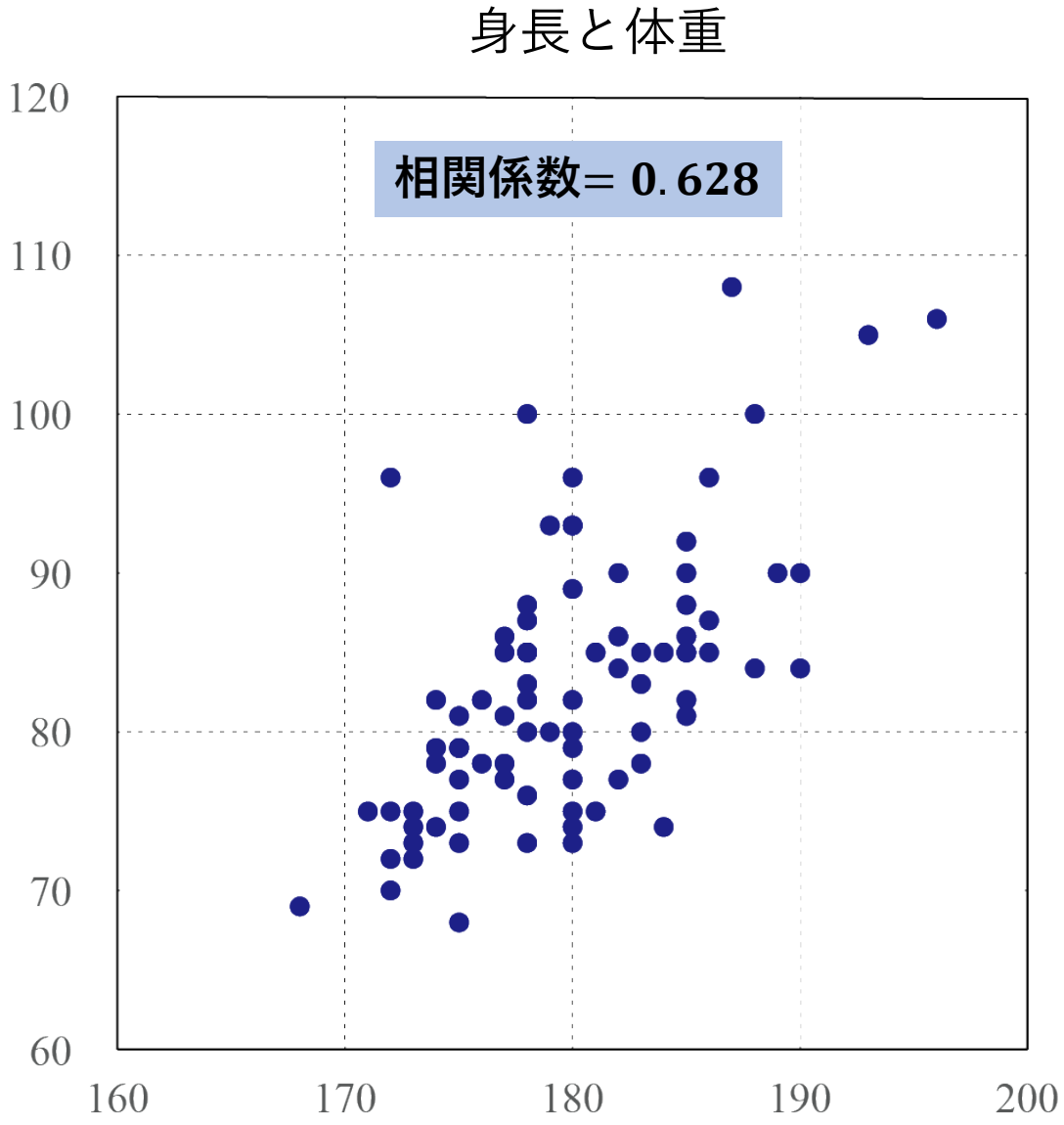
回帰直線（ y：説明変数）

$$\frac{x-\bar{x}}{s_x} = r_{xy}\frac{y-\bar{y}}{s_y} \quad \Longrightarrow \quad x = 0.38y + 25.5$$

身長, 体重, 年齢

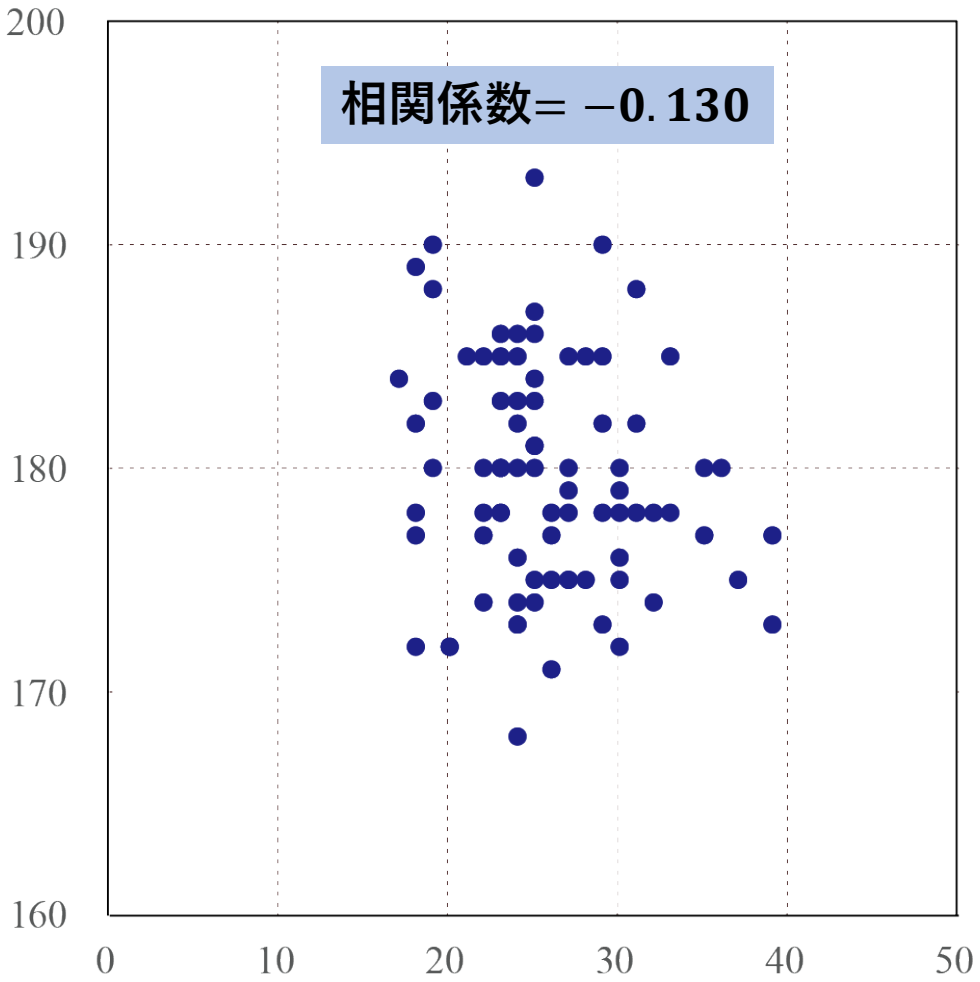
番号	選手名	身長	体重	年齢
1	ウィーラー	178	100	33
2	オコエ瑠偉	185	90	22
3	シャギワ	190	90	29
4	フェルナンド	175	79	27
⋮	⋮	⋮	⋮	⋮
81	鈴木翔天	185	82	23
82	和田恋	180	93	24
83	澤野聖悠	184	85	17

（あるスポーツチームのデータ）

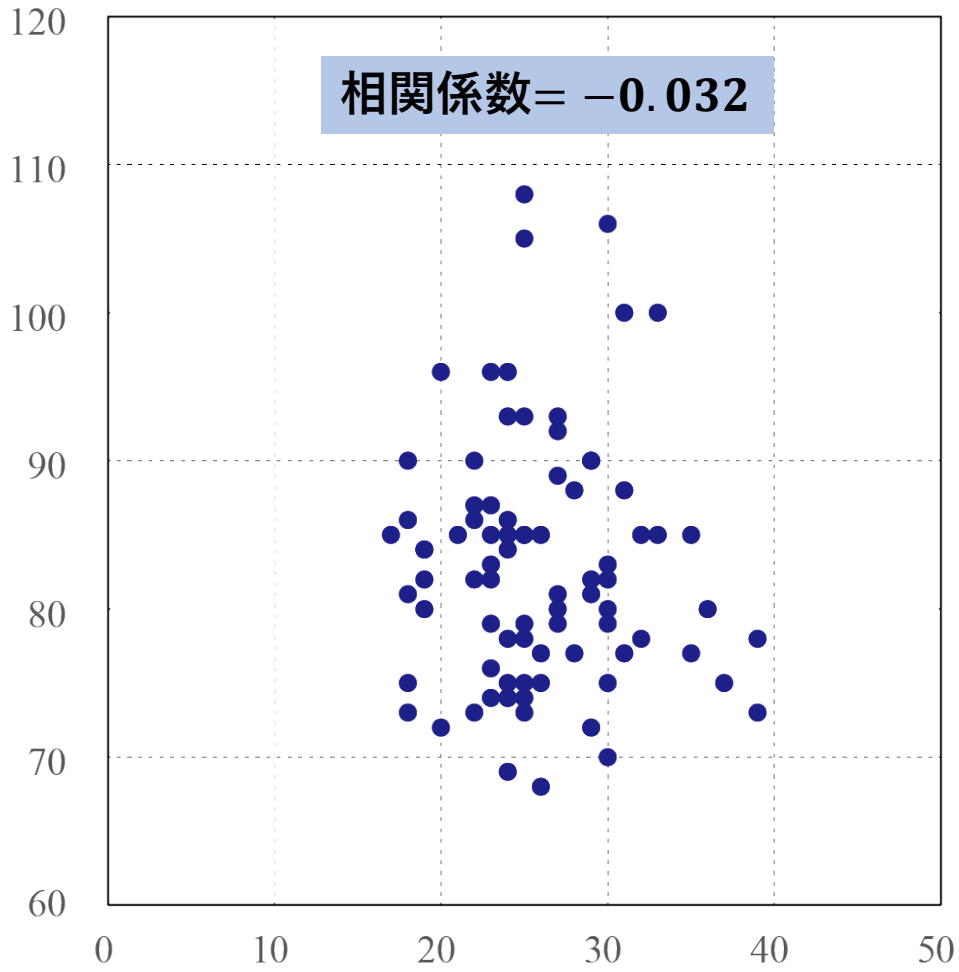


身長, 体重, 年齢

年齢と身長



年齢と体重



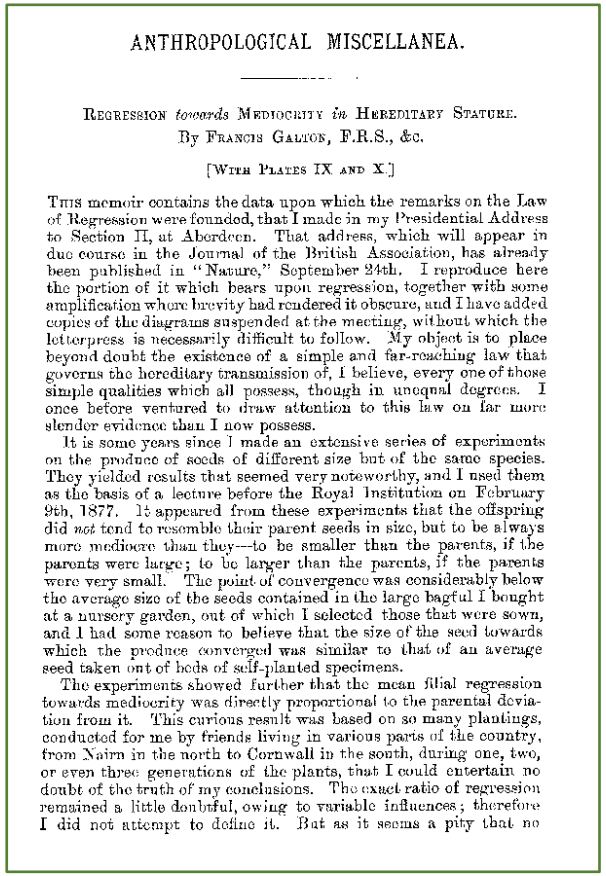
親の身長と子の身長 (x, y)

Heights of Adult Children (y)	Mid-Heights of Parents (x)											
	below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
above							5	3	2	4		14
73.2						3	4	3	2	2	3	17
72.2			1		4	4	11	4	9	7	1	41
71.2			2		11	18	20	7	4	2		64
70.2			5	4	19	21	25	14	10	1		99
69.2	1	2	7	13	38	48	33	18	5	2		167
68.2	1		7	14	28	34	20	12	3	1		120
67.2	2	5	11	17	38	31	27	3	4			138
66.2	2	5	11	17	36	25	17	1	3			117
65.2	1	1	7	2	15	16	4	1	1			48
64.2	4	4	5	5	14	11	16					59
63.2	2	4	9	3	5	7	1	1				32
62.2		1		3	3							7
below	1	1	1			1		1				5
sum	14	23	66	78	211	219	183	68	43	19	4	928

F. Galton（ゴルトン）

Regression towards mediocrity in hereditary stature, Anthropological Miscellanea (1886)

回帰分析の始まり



ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

Adult Children (y) \ Mid-height parents (x)	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	sum
73.2					3	4	3	2	2	14
72.2		1		4	4	11	4	9	7	40
71.2		2		11	18	20	7	4	2	64
70.2		5	4	19	21	25	14	10	1	99
69.2	2	7	13	38	48	33	18	5	2	166
68.2		7	14	28	34	20	12	3	1	119
67.2	5	11	17	38	31	27	3	4		136
66.2	5	11	17	36	25	17	1	3		115
65.2	1	7	2	15	16	4	1	1		47
64.2	4	5	5	14	11	16				55
63.2	4	9	3	5	7	1	1			30
62.2	1		3	3						7
sum	22	65	78	211	218	178	64	41	15	892

$\bar{x} = 68.3$　　$\bar{y} = 68.1$

$s_x^2 = 2.77$　　$s_y^2 = 5.62$

$s_x = 1.67$　　$s_y = 2.37$

$s_{xy} = 1.60$

$r_{xy} = 0.41$

回帰直線（ x：説明変数）

$$y = 0.58\,x + 28.36$$

1 inch = 2.54 cm　を
用いてcm で表すと

$$y = 0.58\,x + 72$$

例（1）　$x = 175 \rightarrow y = 173.5$　　例（2）　$x = 160 \rightarrow y = 164.8$

Lecture 2

2 変量データの整理

おわり

Lecture 3

確率の基本

確率とは

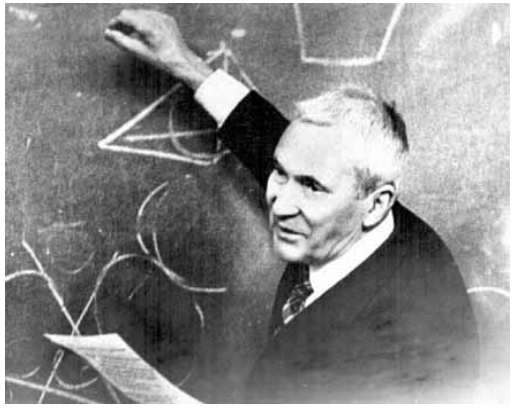
- 生起が決定論的に確定できない事象に対する確からしさの指標
- 0 (= 0%) から 1 (= 100%) までの数値で表現

$P(A)$　　事象 A の起こる確率（Probability）

カルダーノ（16c）、フェルマ、パスカル、ベルヌーイ（17c）、ラプラス（18-19c）

→ コルモゴロフ（20c）の公理論的確率

- 数学的に厳密な理論として出発する.
- 現代統計学の基礎になる.
- 確率モデルとして, 自然科学・生命科学・社会科学など
  あらゆる分野に波及している.



A. N. Kolmogorov (1903-1987)

確率とは

- 生起が決定論的に確定できない事象に対する確からしさの指標
- 0 (= 0%) から 1 (= 100%) までの数値で表現

$P(A)$　事象 A の起こる確率（Probability）

- 確率計算の原理

$$P(A) = \frac{|A|}{|\Omega|}$$

$|A|$：問題の事象の大きさ

$|\Omega|$：起こりうるすべての事象を網羅したものの大きさ

確率空間　$(\Omega, \mathcal{A}, P)$

Ω ：全事象または標本空間

- 根元事象を集めた集合

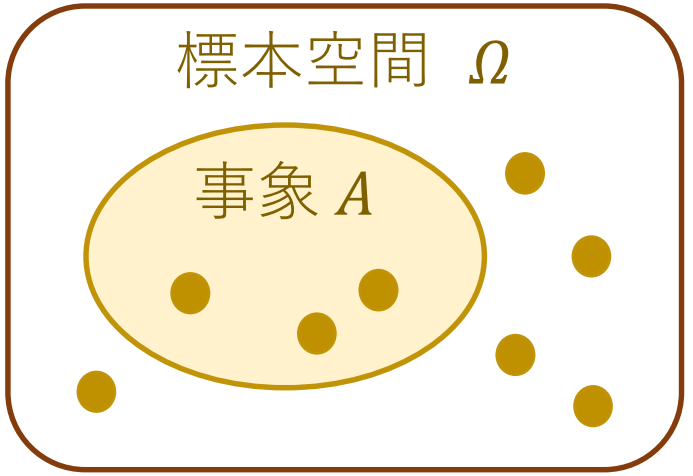
A ：事象

- 有限または無限個の根元事象からなり，確率を与える対象となる
- 標本空間の部分集合

$\mathcal{A}$ ：事象族

- 扱う事象を限定する（数学的理由）

$P(A)$ ：事象 A の起こる確率



計算の原理

$$P(A) = \frac{|A|}{|\Omega|}$$

各根元事象が
等確率で起こる

E : 偶数の目の出る事象

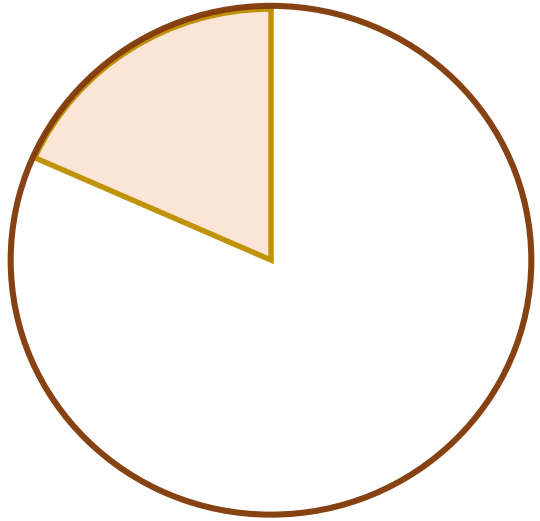
$$P(E) = \frac{3}{6}$$

例 ダーツの確率モデル

- 根元事象＝円板上の各点
- 標本空間 ＝ 円板そのもの

Ω ：半径 R の円板

A ：  に命中する事象



- 確率 $P(A)$

$$P(A) = \frac{|A|}{|\Omega|}$$

面積比

前提：どの点にも偏りなく命中する

標本空間が有限または可算集合の場合

- 根元事象に番号を付けることができる

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_n, \cdots\}$$

- 各根元事象の確率を考えることができる

$$p_k = P(\{\omega_k\}) = P(\omega_k)$$

2重かっこはうっとうしい

- 確率

$$P(A) = \sum_{k:\omega_k \in A} p_k$$

特に, 組合せ確率

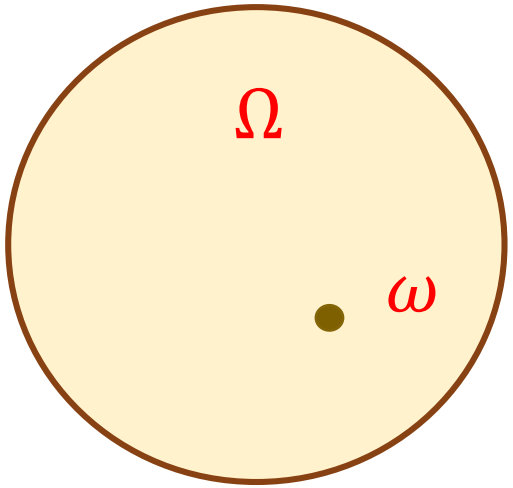
- Ω が有限集合
- $p_k = P(\omega_k)$ が一定

$$P(A) = \frac{|A|}{|\Omega|}$$

Ω と A の要素の個数の比

標本空間が連続無限集合の場合

✓ 各根元事象に確率を与えることでは先に進めない



$P(\{\omega\}) = p$ （一定）　　$0 \leq p \leq 1$

Ω から n 個の点 $\omega_1, \omega_2, \cdots, \omega_n$ を選ぶ

$E_n = \{\omega_1, \omega_2, \cdots, \omega_n\}$

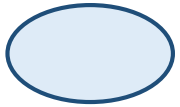
点 $\omega_1, \omega_2, \cdots, \omega_n$ のいずれかが選ばれる事象

$P(E_n) = np$

明らかに，　$0 \leq P(E_n) \leq 1$ ⇒ $0 \leq p \leq \frac{1}{n}$ ⇒ $p = 0$

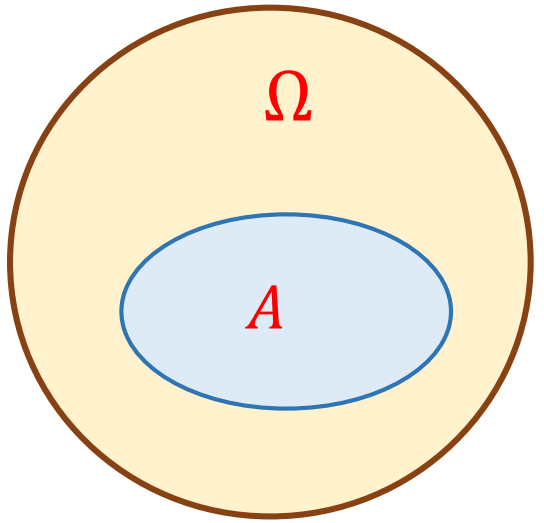
n は任意

円板からランダムに1点選ぶ

事象 A：  から点が選ばれること

$$P(A) = \frac{|A|}{|\Omega|}$$

Ω と A の面積比

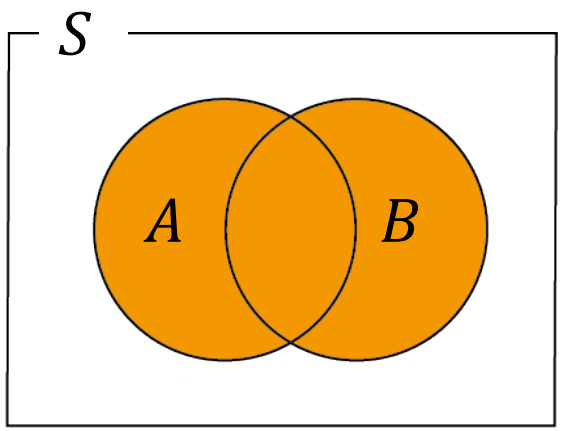


どの点にも偏りなくランダムに選ばれることをモデル化

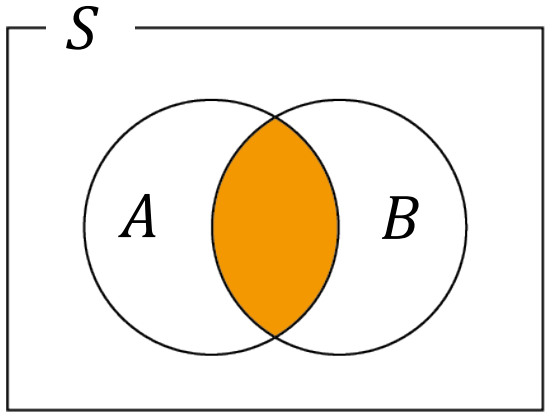
- ✓ 面積が同じならば、形や位置によらない
- ✓ 面積が2倍になれば, 確率も2倍になる

幾何学的確率：面積比以外に, 長さの比や体積比も使える

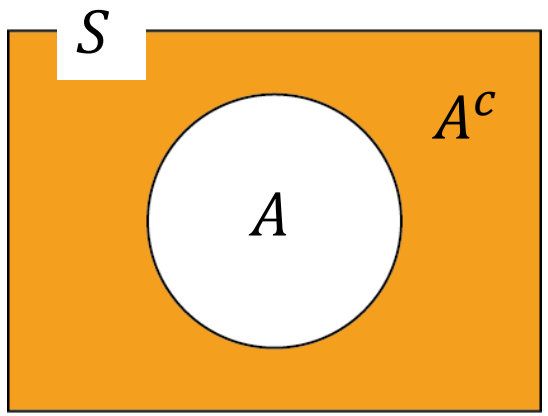
和事象　$A \cup B$



積事象　$A \cap B$



余事象 A^cまたは $\bar{A}$



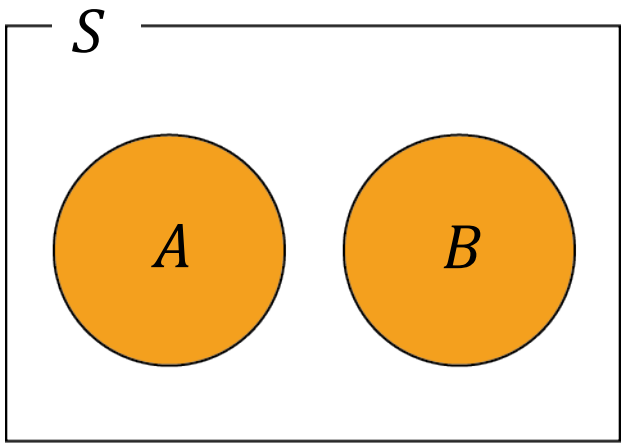
$\emptyset$: 空事象（起こりえない）

$$P(\emptyset) = 0$$

S : 全事象（必ず起こる事象）

$$P(S) = 1$$

排反な事象　$A \cap B = \emptyset$



確率の公理

裏に高度な数学あり

(0) 事象族 $\mathcal{A}$ は 可算加法族 をなす.

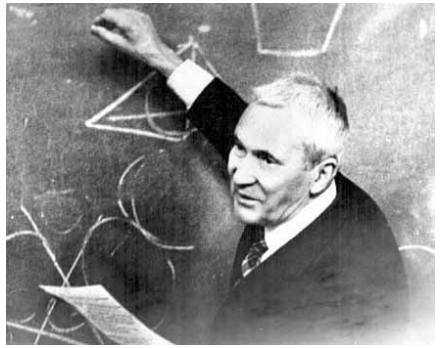
(1) $0 \leq P(A) \leq 1$

(2) $P(\emptyset) = 0,\ P(\Omega) = 1$

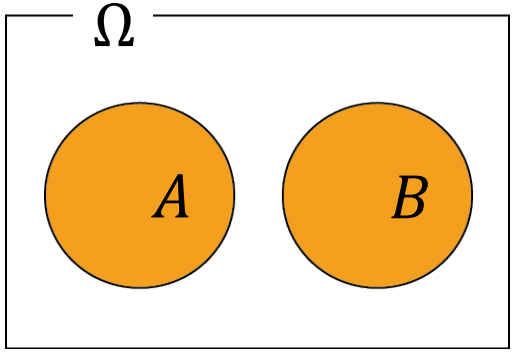
(3) $A_1, A_2, \cdots, A_n, \cdots$ が互いに排反であれば,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

事象の無限列を扱う
裏に高度な数学あり



A. N. Kolmogorov (1903-1987)



(3') A と B が排反$(A \cap B = \emptyset)$ならば,

$$P(A \cup B) = P(A) + P(B)$$

事象の有限列はカバーされるが
無限列を扱えない

定 理

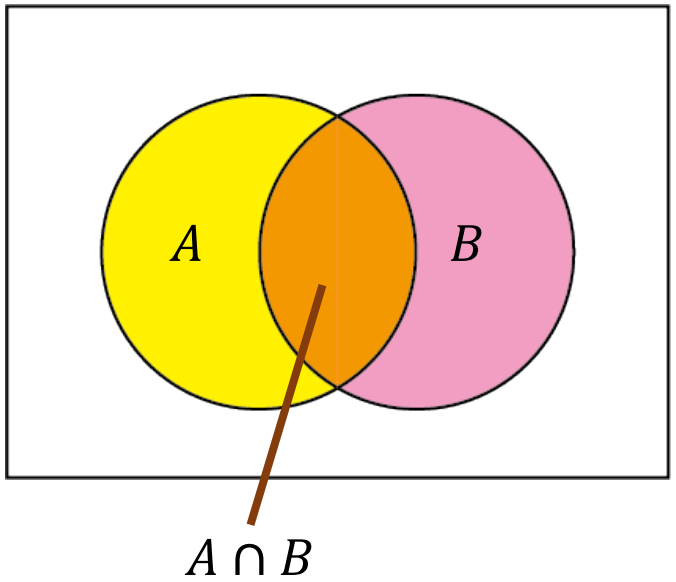
(1) 事象 A, B があるとき，A または B の事象の起こる確率 について

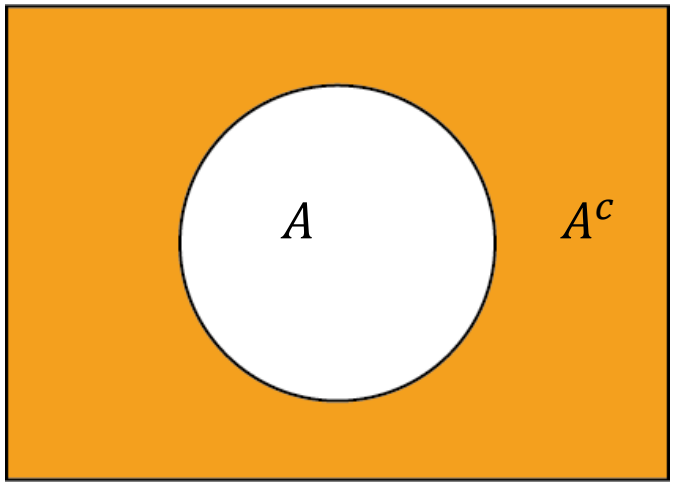
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(2) 事象 A の余事象が起こる確率 について

$$P(A^c) = 1 - P(A)$$

ポイント：
公理だけを用いて証明する





例 題 3.1　2 つのサイコロを投げ, 出た目の差を D とする.
D が $0, 1, 2, 3, 4, 5$ の場合についてその確率を求めよ.

例 題 3.1　2 つのサイコロを投げ, 出た目の差を D とする.
D が $0, 1, 2, 3, 4, 5$ の場合についてその確率を求めよ.

考え方

- 起こりうるすべての場合を考える

  サイコロの目は 6 通りの出方がある.
  したがって, 2 つのサイコロでは 36 通りの出方がある

- 記号を準備する

  2 つのサイコロの目 を X , Y とする. 試行の結果は (X, Y) となる.

  したがって, その差は $D = |X - Y|$ となる.

- 起こりうるそれぞれの場合の確率を考える

➢ 起こりうるすべての場合を考える

2つのサイコロの出目をそれぞれ X, Y とする

X \ Y	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

確率の決め方

公平なサイコロの

各根元事象は等確率

➢ 起こりうるすべての場合を考える

2つのサイコロの出目をそれぞれ X, Y とする

X \ Y	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

記号の使い方

$$P(X = 2, Y = 4) = \frac{1}{36}$$

$$P\big((X, Y) = (2,4)\big) = \frac{1}{36}$$

$$P(X + Y = 10) = \frac{3}{36} = \frac{1}{12}$$

各根元事象は等確率

➢ 起こりうるすべての場合を考える

2つのサイコロの出目をそれぞれ X, Y とする

X \ Y	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- 2個のサイコロの目の差

$$D = |X - Y|$$

- D の取りうる値は

$$0, 1, 2, 3, 4, 5$$

$$P(D = 0) = \frac{6}{36}$$

➢ 起こりうるすべての場合を考える

2つのサイコロの出目をそれぞれ X, Y とする

X \ Y	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

サイコロの目の差

$$D = |X - Y|$$

d	0	1	2	3	4	5
$P(D = d)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

この表を D の確率分布という

独立性

定 義　事象 A, B が独立である とは, 事象 A, B について

$$P(A \cap B) = P(A)P(B)$$

が成り立つことをいう．独立でないとき従属であるという．

例　サイコロを２回投げる．

X ：１回目に出る目，Y ：２回目の出る目

２つの事象 $A = \{X = 2\}$ と $B = \{Y = 5\}$ は独立である

$$P(A \cap B) = \frac{1}{36} \qquad P(A)P(B) = \frac{1}{6} \times \frac{1}{6}$$

ゆえに $P(A \cap B) = P(A)P(B)$ が成り立ち，A と B は独立．

例 題 3.2　52 枚のトランプから 1 枚をランダムに抜き取るとき, そのカードのスートを X, 数字を Y とする. X と Y は独立である.

例 題 3.2　52 枚のトランプから 1 枚をランダムに抜き取るとき, そのカードのスートを X, 数字を Y とする. X と Y は独立である.

考え方

（1）起こりうるすべての場合を考える

52枚のカードのうちの 1 枚なので 52 通りの起こり方がある

（2）記号を用いて適切に表示する

（3）起こりうるそれぞれの場合の確率を考える

例 題 3.2　52 枚のトランプから 1 枚をランダムに抜き取るとき, そのカードのスートを X, 数字を Y とする. X と Y は独立である.

起こりうるすべての場合を書き下す

X \ Y	1	2		11	12	13
1	♥ A	♥ 2	・・・	♥ J	♥ Q	♥ K
2	♠ A	♠ 2	・・・	♠ J	♠ Q	♠ K
3	♦ A	♦ 2	・・・	♦ J	♦ Q	♦ K
4	♣ A	♣ 2	・・・	♣ J	♣ Q	♣ K

♠ Q を引く確率

$$P(X=2, Y=12) = \frac{1}{52}$$

一方,

$$P(X=2) = \frac{13}{52} = \frac{1}{4}$$

$$P(Y=12) = \frac{4}{52} = \frac{1}{13}$$

ゆえに,

$$P(X=2, Y=12) = P(X=2)P(Y=12)$$

事象 $\{X=2\}$ と $\{Y=12\}$ は独立である

例 題 3.2　52 枚のトランプから 1 枚をランダムに抜き取るとき, そのカードのスートを X, 数字を Y とする. X と Y は独立である.

起こりうるすべての場合を書き下す

X \ Y	1	2		11	12	13
1	♥ A	♥ 2	・・・	♥ J	♥ Q	♥ K
2	♠ A	♠ 2	・・・	♠ J	♠ Q	♠ K
3	♦ A	♦ 2	・・・	♦ J	♦ Q	♦ K
4	♣ A	♣ 2	・・・	♣ J	♣ Q	♣ K

$\begin{matrix} \boldsymbol{a} \\ \boldsymbol{b} \end{matrix}$ を引く確率

$$P(X = \boldsymbol{a}, Y = \boldsymbol{b}) = \frac{1}{52}$$

一方,

$$P(X = \boldsymbol{a}) = \frac{13}{52} = \frac{1}{4}$$

$$P(Y = \boldsymbol{b}) = \frac{4}{52} = \frac{1}{13}$$

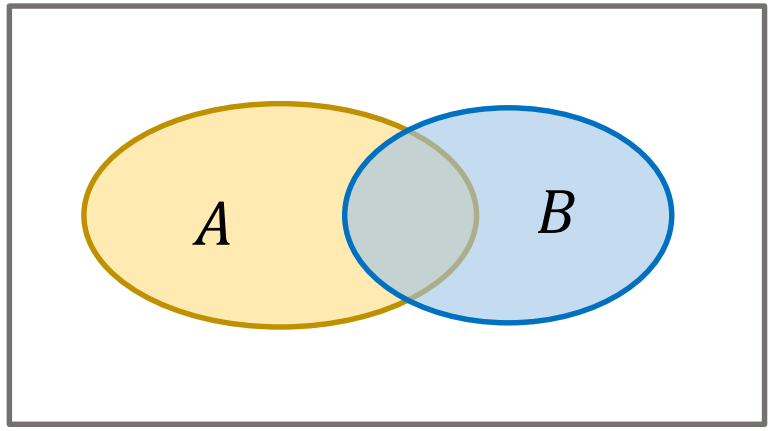
ゆえに,

$$P(X = \boldsymbol{a}, Y = \boldsymbol{b}) = P(X = \boldsymbol{a})P(Y = \boldsymbol{b})$$

確率変数 X, Y は独立である

条件付き確率

A, B：事象



A の下での B の条件付確率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

ただし，$P(A) > 0$

乗法定理

2 つの事象 A, B が同時に起こる確率

$P(A \cap B)$ は，条件付き確率を使って，

次のように表される．

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

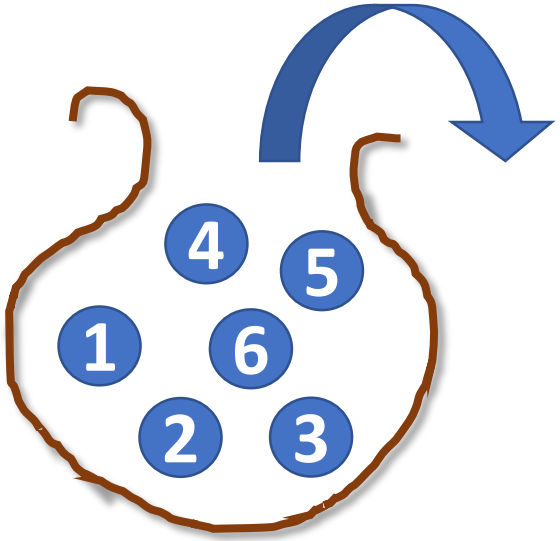
解釈

$P(B|A)$ は事象 A が起こったことを

知った上で，事象 B の起こる確率

例題 3.3

袋の中に球が 6 個が入っていて, それぞれに 1, 2, 3, 4, 5, 6 の数字が書いてあるとする. 袋の中から 2 回続けて1個の球を取り出す. 球の数字が順に i, j $(i, j = 1, \ldots, 6)$ となる確率を求めよ.

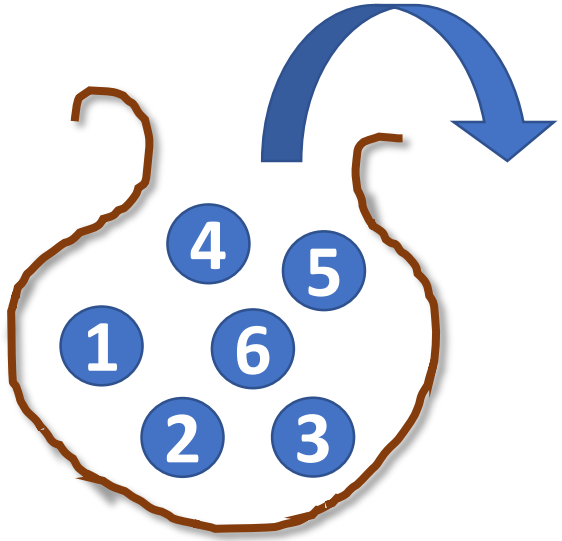


(1)【復元抽出】 1 回目に取り出した球を袋に戻し, よくかきまぜて, 改めて袋の中から球を 1 個取り出す.

(2)【非復元抽出】 1 回目に取り出した球をもとに戻さず, 続けて球を取り出す.

例題 3.3

袋の中に球が 6 個が入っていて, それぞれに 1, 2, 3, 4, 5, 6 の数字が書いてあるとする. 袋の中から 2 回続けて1個の球を取り出す. 球の数字が順に i, j $(i, j = 1, \ldots, 6)$ となる確率を求めよ.



(1)【復元抽出】 1 回目に取り出した球を袋に戻し, よくかきまぜて, 改めて袋の中から球を 1 個取り出す.

(2)【非復元抽出】 1 回目に取り出した球をもとに戻さず, 続けて球を取り出す.

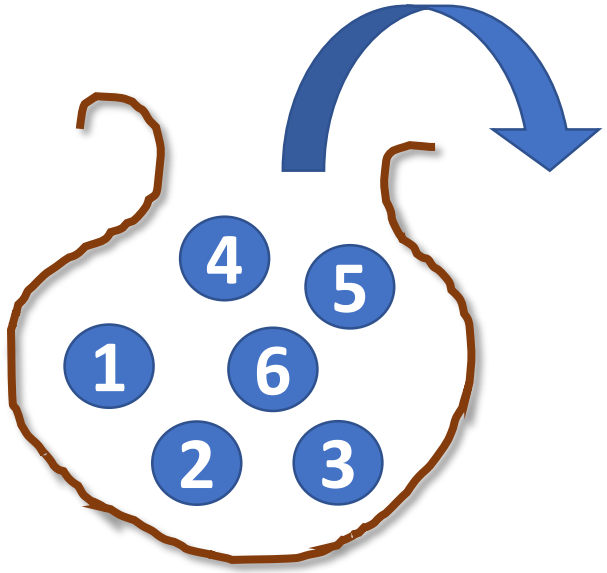
記号の準備

A_i ： 1 回目の球の数字が i となる事象

B_j ： 2 回目の球の数字が j となる事象

求める確率 $= P(A_i \cap B_j)$

（1）【復元抽出】



6個の球は等確率で選ばれるから　$P(A_i) = \frac{1}{6}$

2 回目の試行では，袋の中や取り出し方は1 回目と同じである．したがって，

$$P(B_j) = \frac{1}{6}$$

1回目と 2 回目の試行は独立であるから

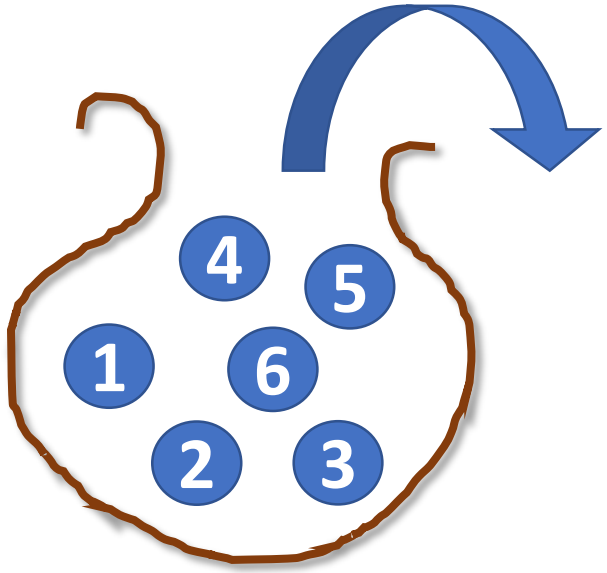
$$P(A_i \cap B_j) = P(A_i)P(B_j) = \frac{1}{36}$$

別解　1 回目が 2 回目に影響しないので　$P(B_j) = P(B_j|A_i)$

乗法定理によって

$$P(A_i \cap B_j) = P(A_i)P(B_j|A_i) = P(A_i)P(B_j) = \frac{1}{36}$$

（２）【非復元抽出】



乗法定理によって　$P(A_i \cap B_j) = P(A_i)P(B_j|A_i)$

題意から

$$P(B_j|A_i) = \begin{cases} 0, & i = j \\ \dfrac{1}{5}, & i \neq j \end{cases}$$

$i = j$ のとき，

$$P(A_i \cap B_j) = P(A_i)P(B_j|A_i) = 0$$

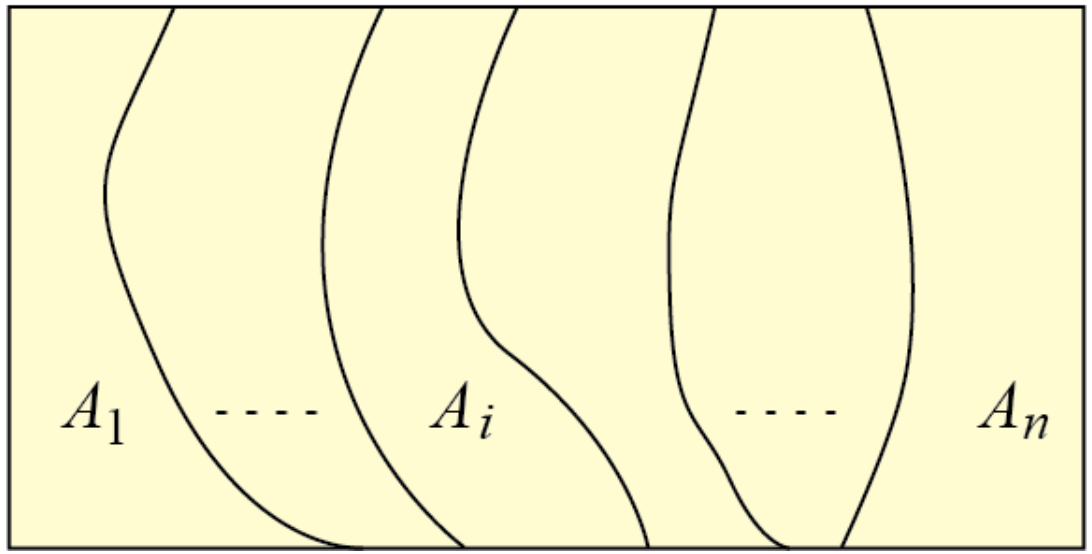
$i \neq j$ のとき，

$$P(A_i \cap B_j) = P(A_i)P(B_j|A_i) = \frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$$

注意　A_i と B_j は独立ではない．なぜなら $P(B_j) = \dfrac{1}{6}$ であるから（確認せよ）

ベイズの定理

全事象 S の層別 (stratification)



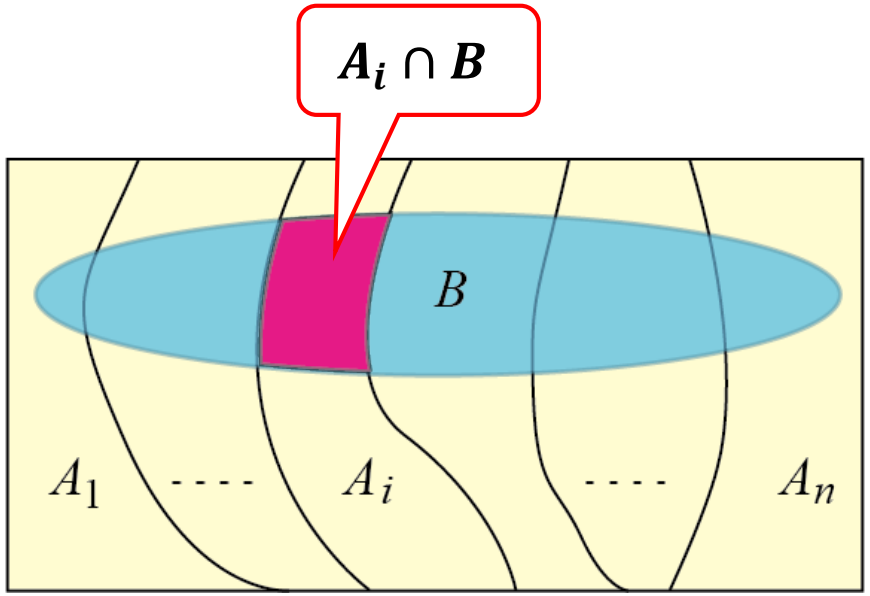
$$A_1 \cup A_2 \cup \cdots \cup A_n = S$$

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1$$

例　全人口を「10 代以下, 20 代, ..., 60 代, 70 代以上」と年代別に考える.

全確率の公式　全事象が $A_i\ (i = 1,\cdots,n)$ に層別されているとき，事象 B の確率 $P(B)$ は，A_i の事前確率 $P(A_i)$ と条件付き確 率 $P(B|A_i)$ を使って表される．

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$



証明

$$B = \bigcup_{i=1}^{n} A_i \cap B \quad \text{（互いに排反）}$$

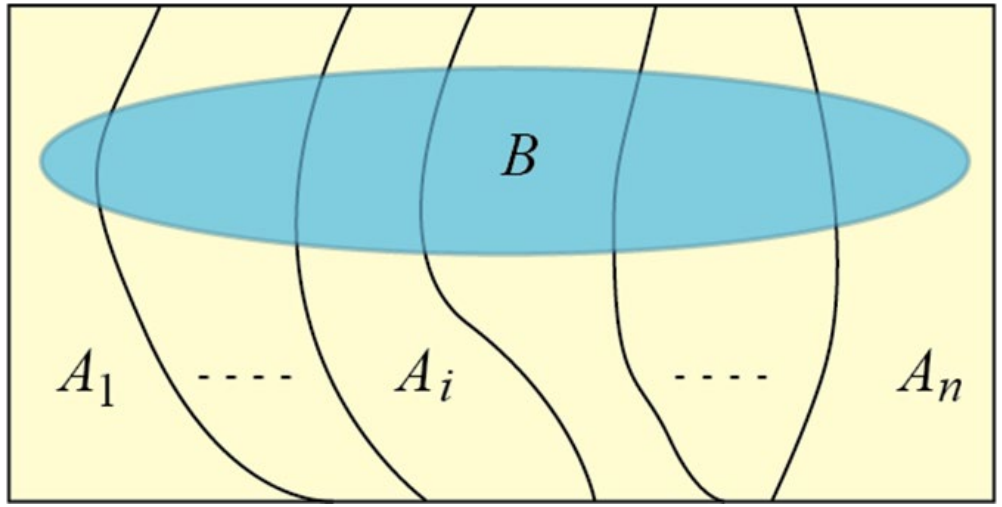
$$\therefore\ P(B) = \sum_{i=1}^{n} P(A_i \cap B)$$

一方，$P(A_i \cap B) = P(A_i)P(B|A_i)$ なので，　$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

ベイズの定理　全事象が $A_i\ (i = 1,\cdots,n)$ に層別されているとき，事象 B が起こったときの各層の条件付き確率 $P(A_i|B)$ を事後確率といい，事前確率 $P(A_i)$ と条件付き確率 $P(B|A_i)$ を使って表される．

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)}$$

証明



定義から　$P(A_i|B) = \dfrac{P(A_i \cap B)}{P(B)}$

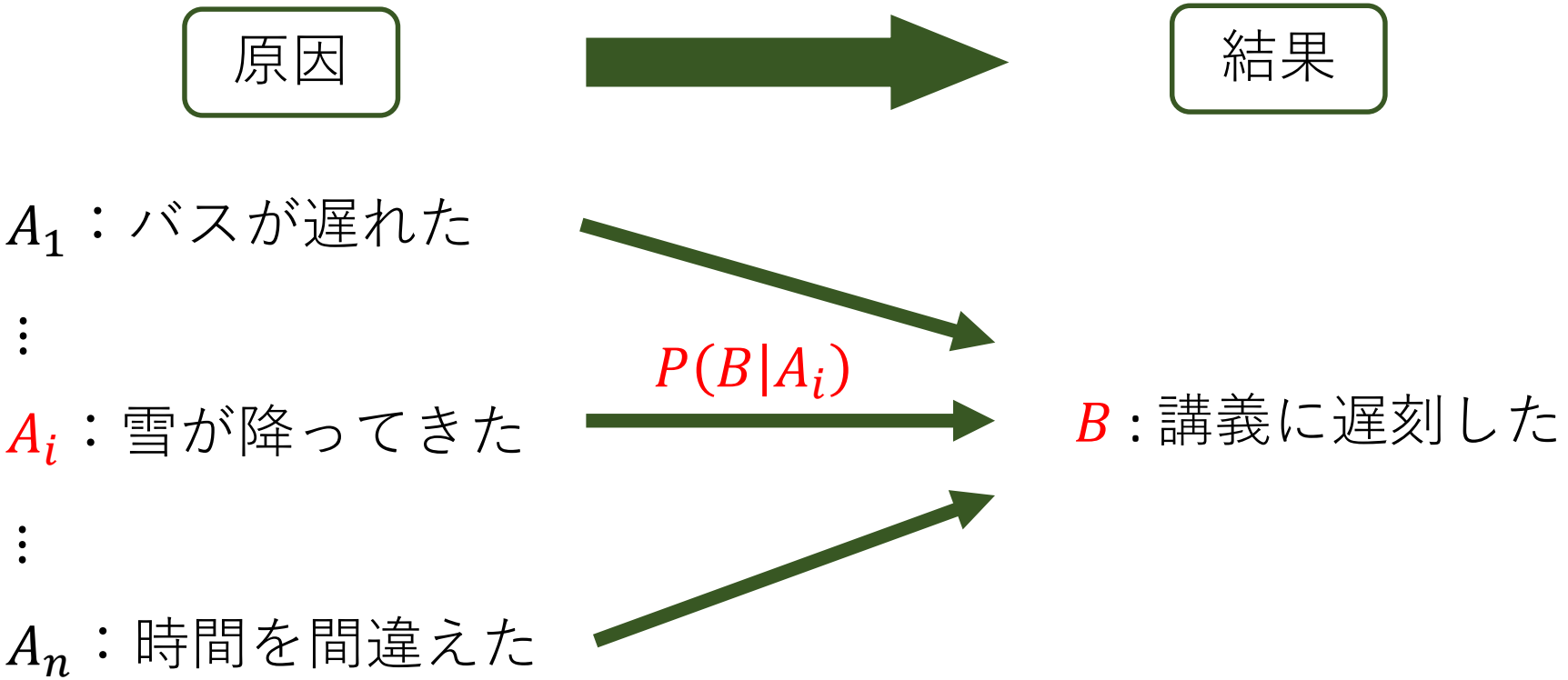
分子 $= P(A_i \cap B) = P(A_i)\,P(B|A_i)$

分母 $= P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$

これは全確率の公式

ベイズの公式＝結果から原因を知る公式

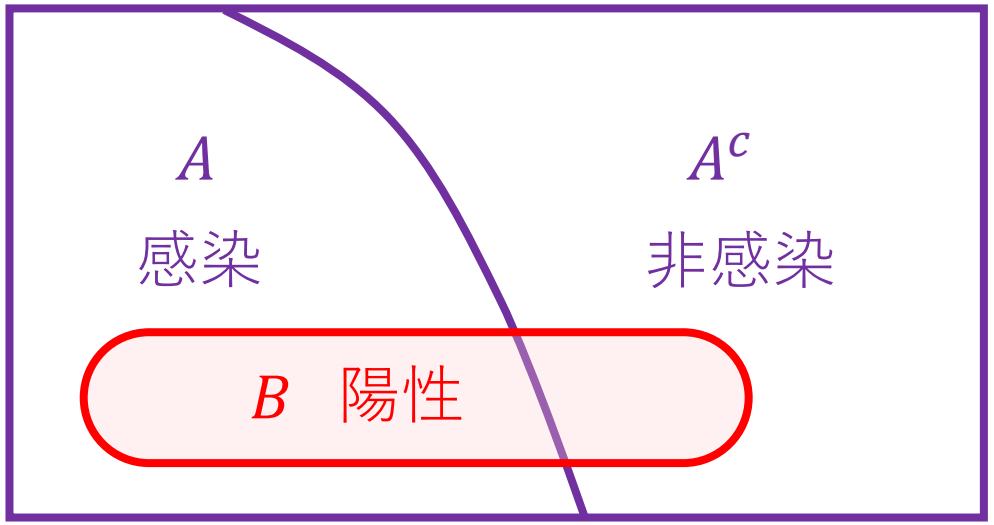
$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)}$$



例題 3.4 （擬陽性の問題）

- 病気 A の感染者は 500 人に 2 人の割合であるという.
- 検査 B は, 感染者の 95% に陽性反応を示すが,
  非感染者の 2% にも陽性反応が出てしまう.

（１）陽性反応が出れば感染しているか？
（２）陰性反応なら非感染か？



$$P(A) = \frac{2}{500}$$

$$P(B|A) = 0.95$$

$$P(B|A^c) = 0.02$$

（１）陽性反応なら感染か？

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$$

$$= \frac{\frac{2}{500} \times 0.95}{\frac{2}{500} \times 0.95 + \frac{498}{500} \times 0.02}$$

$$= \frac{1.9}{1.9 + 9.96} = 0.160$$

（２）陰性反応なら非感染か？

$$P(A^c|B^c) = \frac{P(A^c)P(B^c|A^c)}{P(A^c)P(B^c|A^c) + P(A)P(B^c|A)}$$

$$= \frac{\frac{498}{500} \times 0.98}{\frac{498}{500} \times 0.98 + \frac{2}{500} \times 0.05}$$

$$= \frac{488.04}{488.04 + 0.1} = 0.999795$$

Lecture 3

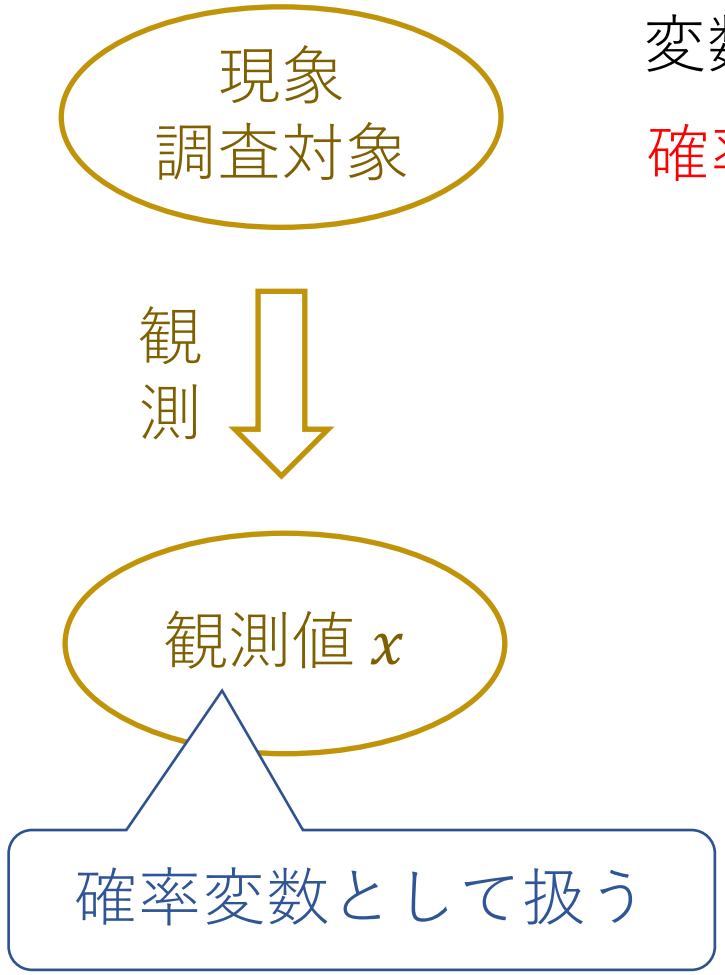
確率の基本

おわり

Lecture 4

離散型確率分布

確率変数(random variable)とは？



変数：ある範囲の値を代表する．$x, y, z, t, \ldots$

確率変数：確率を伴ってある範囲の値をとる．$X, Y, Z, T, \ldots$

- Discrete random variables (離散型確率変数)
  (1) コインを3回投げるとき表の出る回数.
  (2) 授業開始時の出席者数.

数える

- Continuous random variables (連続型確率変数)
  (1) 円の内部から1点をランダムに選んだとき，
  その点と中心との距離.
  (2) 新生児の体重.

測る/量る

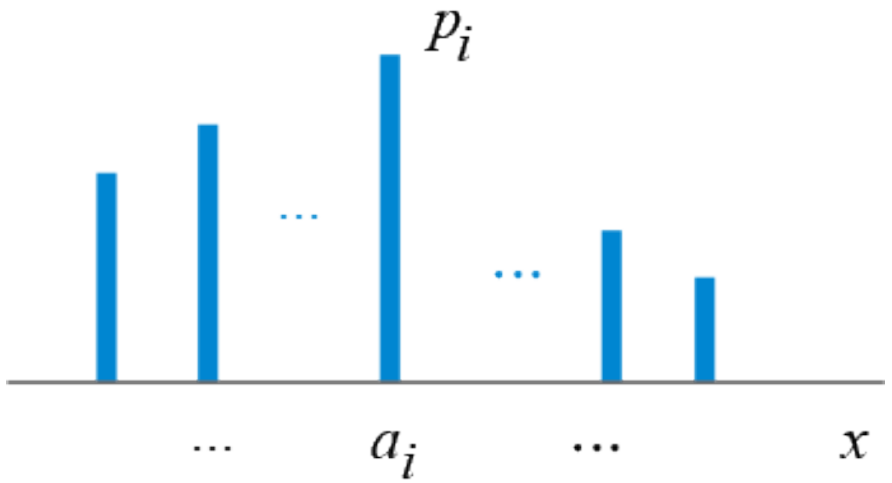
（注意）確率変数 X は特定の数を表すのではない.
その取りうる個別の値を X の実現値という.

離散型確率変数の分布(distribution)

取りうる値　$a_1, a_2, \dots, a_i, \dots$

その確率　$p_1, p_2, \dots, p_i, \dots$

x	a_1	$\cdots$	a_i	$\cdots$	合計
$P(X=x)$	p_1	$\cdots$	p_i	$\cdots$	1



数式による表記

$$P(X = a_i) = p_i$$

基本的な性質

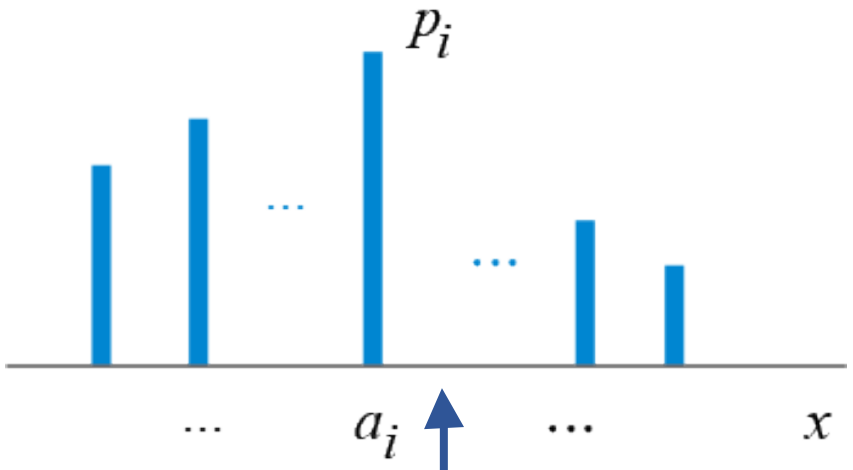
(1)　$p_i \geq 0$

(2)　$\sum p_i = 1$

離散型確率変数の平均値と分散

離散型確率変数 X の確率分布

x	a_1	$\cdots$	a_i	$\cdots$	合計
$P(X=x)$	p_1	$\cdots$	p_i	$\cdots$	1



重心

➢平均値 $\mathbf{E}[X] = m_X = \sum a_i p_i = \sum a_i P(X = a_i)$

➢分散 $\mathbf{V}[X] = \sigma_X^2 = \sum (a_i - m_X)^2 p_i = \sum a_i^2 p_i - m_X^2$

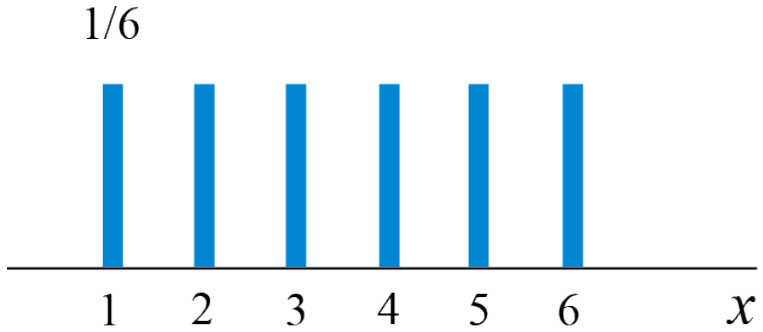
$= \mathbf{E}[(X - m_X)^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$　【分散公式】

例 題 4.1　サイコロを 1 回振って, 出た目を X は $\{1,2,3,4,5,6\}$ の範囲を動く確率変数である. X の確率分布, 平均値, 分散を求めよ.

確率分布

$$P(X=k)=\frac{1}{6},\quad k=1,2,\cdots,6$$

x	1	2	3	4	5	6	合計
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1



平均値

$$\mathbf{E}[X]=m_X=\sum a_i p_i=\sum_{k=1}^{6} k\frac{1}{6}=\frac{7}{2}$$

$$\mathbf{E}[X^2]=\sum a_i^2 p_i=\sum_{k=1}^{6} k^2\frac{1}{6}=\frac{91}{6}$$

分散

$$\mathbf{V}[X]=\mathbf{E}[X^2]-\mathbf{E}[X]^2=\frac{91}{6}-\left(\frac{7}{2}\right)^2=\frac{35}{12}$$

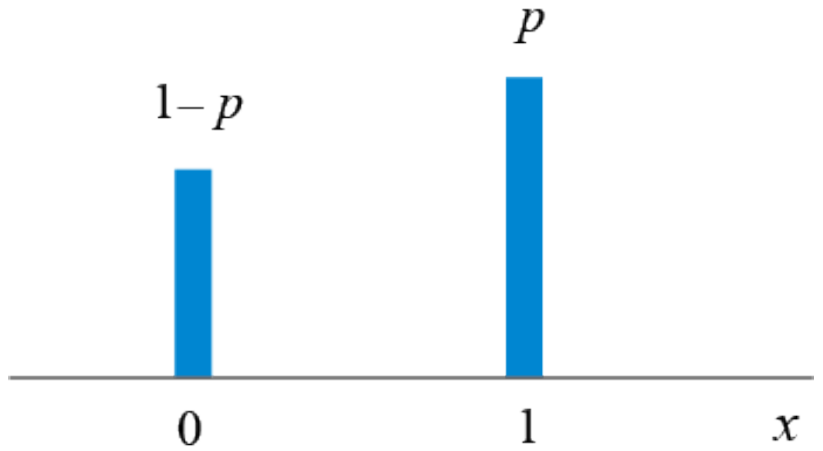
【分散公式】

名前の付いた重要な離散分布

- ✓ ベルヌイ分布　二項母集団の分布
- ✓ 二項分布　ベルヌイ試行列における成功回数の分布
- ✓ 幾何分布　ベルヌイ試行列における成功までの待ち時間
- ✓ ポアソン分布　二項分布の極限（少数の法則）

教科書等で見ておく

ベルヌイ分布 $B(1,p)$



2 値確率変数＝ベルヌイ型確率変数

$$Z=\begin{cases}1 & (\text{確率 } p \text{ で})\\ 0 & (\text{確率 } 1-p \text{ で})\end{cases}$$

x	0	1	計
$P(Z=x)$	$1-p$	p	1

数式だけで書くなら

$$P(Z=0)=1-p$$

$$P(Z=1)=p$$

平均値　$\mathbf{E}[Z]=0\times(1-p)+1\times p=p$

$\mathbf{E}[Z^2]=0^2\times(1-p)+1^2\times p=p$

分散　$\mathbf{V}[Z]=\mathbf{E}[Z^2]-\mathbf{E}[Z]^2=p-p^2=p(1-p)$

平均値　$m=p$

分散　$\sigma^2=p(1-p)$

二項分布 $B(n,p)$

X：表の出る確率 p のコインを n 回投げたときの表の回数

確率分布の計算

$P(X=k)$ ：○（表）が k 回，×（裏）が $n-k$ 回の確率

その一例　　○×・・・○×・・・×○×

その確率　　$p\ q \quad \cdots \quad p\ q \quad \cdots \quad q\ p\ q = p^k q^{n-k}$　（ただし，$q=1-p$）

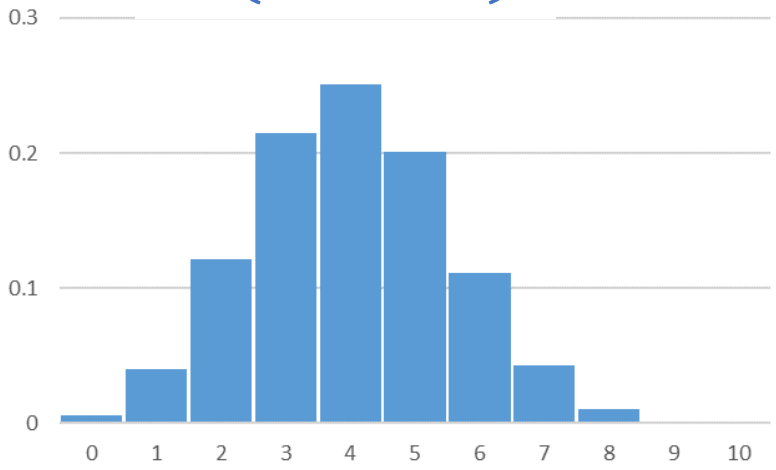
したがって，

二項係数

$$P(X=k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

二項分布の具体例

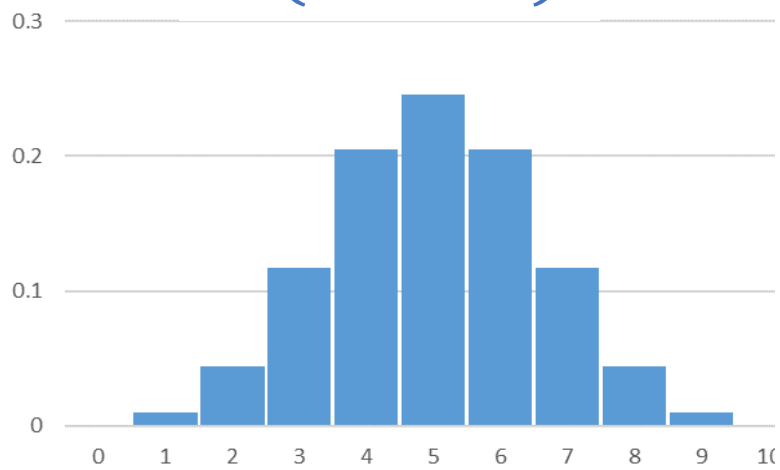
$B(10,0.4)$



$m = 4$

$\sigma^2 = 2.4$

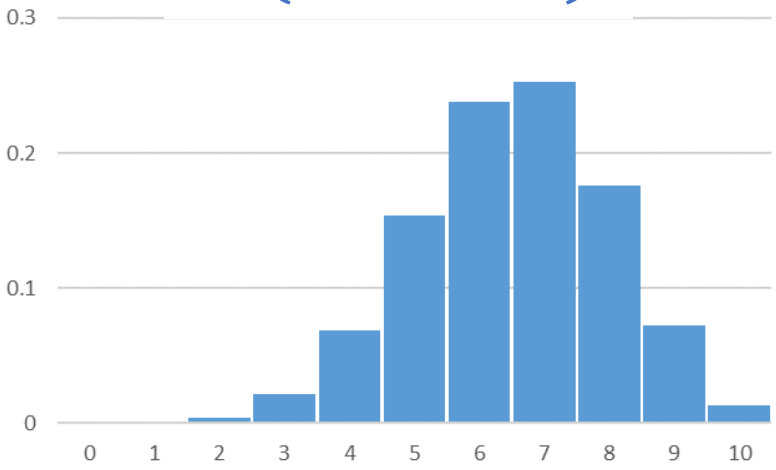
$B(10,0.5)$



$m = 5$

$\sigma^2 = 2.5$

$B(10,0.65)$

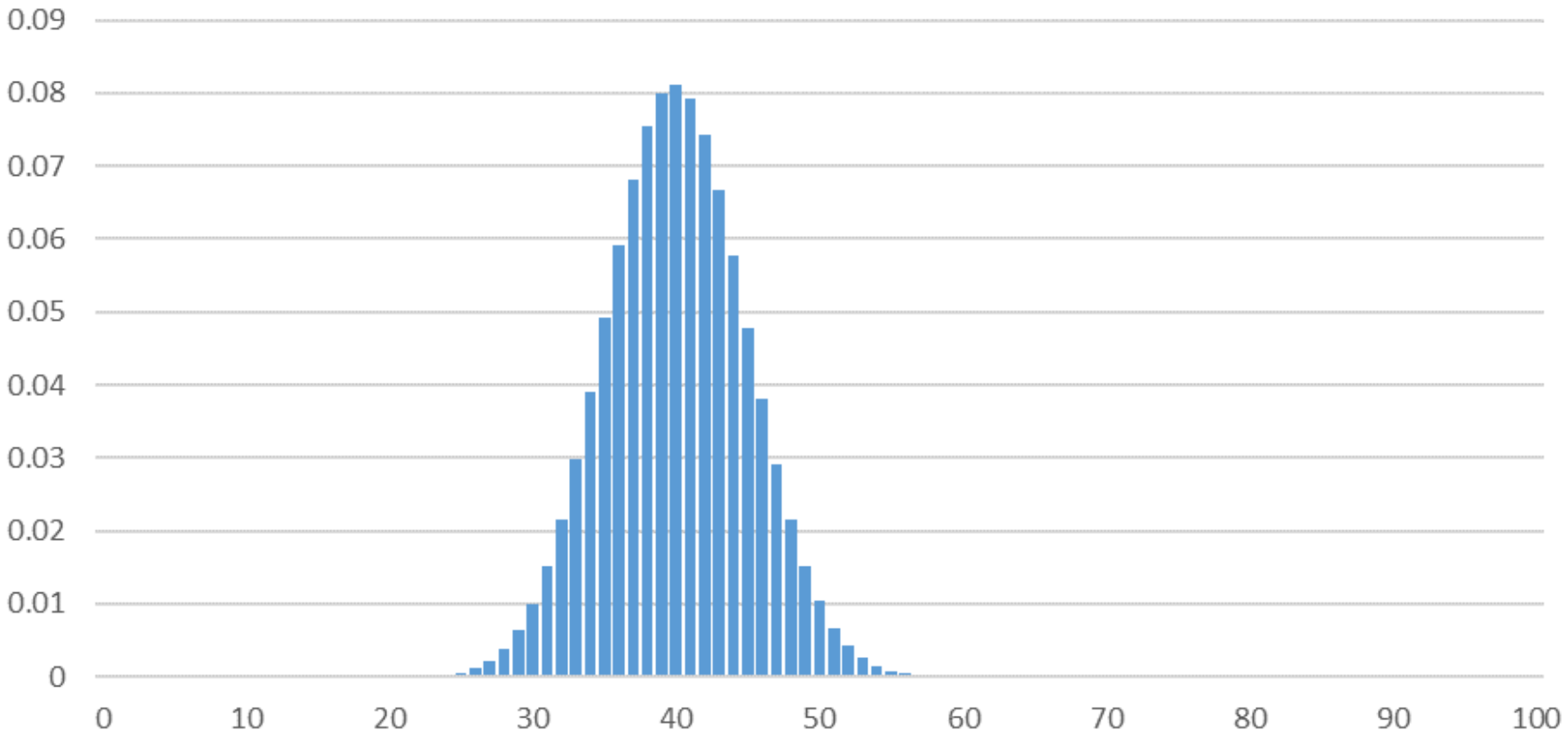


$m = 6.5$

$\sigma^2 = 2.275$

二項分布の具体例

$B(100,0.4)$　　$m = 40,\ \ \sigma^2 = 24$



二項分布 $B(n,p)$

平均値　　$m = np$

分散　　$\sigma^2 = np(1-p)$

$X \sim B(n,p)$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$\mathbf{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbf{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbf{V}[X] = \mathbf{E}[X^2] - \boldsymbol{E}[X]^2$$

※ $E[X^2]$ のかわりに

$$\mathbf{E}[X(X-1)] = \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k}$$

を計算する方が簡単

和の計算

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

$$= \sum_{k=1}^{n} k \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= n \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= n \sum_{k=0}^{n-1} \frac{(n-1)!}{k!\,\big(n-(k+1)\big)!} p^{k+1} (1-p)^{n-(k+1)}$$

$$= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!\,(n-1-k)!} p^k (1-p)^{n-1-k}$$

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

$$= np\big(p+(1-p)\big)^{n-1}$$

$$= np$$

和の計算

$$\sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=2}^{n} k(k-1)\frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

$$= \sum_{k=2}^{n} \frac{n!}{(k-2)!\,(n-k)!}p^k(1-p)^{n-k} = n(n-1)\sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!\,(n-k)!}p^k(1-p)^{n-k}$$

$$= n(n-1)p^2\sum_{k=0}^{n-2} \frac{(n-2)!}{k!\,(n-2-k)!}p^k(1-p)^{n-2-k}$$

$$= n(n-1)p^2\sum_{k=0}^{n-2} \binom{n-2}{k}p^k(1-p)^{n-2-k}$$

$$= n(n-1)p^2\big(p+(1-p)\big)^{n-2} = n(n-1)p^2$$

二項分布 $B(n,p)$

$X \sim B(n,p)$

$$P(X=k)=\binom{n}{k}\,p^k(1-p)^{n-k}$$

$$\mathbf{E}[X]=\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}=np$$

$$\mathbf{E}[X(X-1)]=\sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k}$$

$$=n(n-1)p^2$$

$$\mathbf{E}[X^2]=\mathbf{E}[X(X-1)]+\mathbf{E}[X]$$

$$=n(n-1)p^2+np$$

$$\mathbf{V}[X]=\mathbf{E}[X^2]-\boldsymbol{E}[X]^2$$

$$=n(n-1)p^2+np-(np)^2$$

$$=-np^2+np$$

$$=np(1-p)$$

平均値　$m=np$

分散　$\sigma^2=np(1-p)$

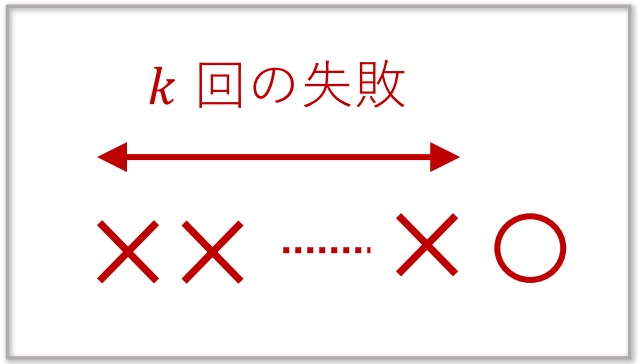
幾何分布

成功確率 p の試行を独立に繰り返す.
初めて成功するまでの失敗の回数を X とする.

$$P(X=k)=p(1-p)^k \qquad k=0,1,2,\ldots$$

この分布を**パラメータ p の幾何分布**という.

事象 $\{X=k\}$



（注意）$X+1$ は成功までの待ち時間

$$\mathbf{E}[X]=\sum_{k=0}^{\infty}kp(1-p)^k=\frac{1-p}{p}=\frac{1}{p}-1$$

$$\mathbf{E}[X^2]=\sum_{k=0}^{\infty}k^2p(1-p)^k=\frac{(2-p)(1-p)}{p^2}$$

$$\mathbf{V}[X]=\mathbf{E}[X^2]-\mathbf{E}[X]^2=\frac{1-p}{p^2}$$

平均値　$m=\dfrac{1-p}{p}$

分散　$\sigma^2=\dfrac{1-p}{p^2}$

確率母関数 (Generating function)

$n = 0,1,2, \ldots$ に値をとる確率変数 X に対して $p_n = P(X = n)$ とおく

確率母関数　$f(x) = \sum_{n=0}^{\infty} p_n x^n$

$$f'(x) = \sum_{n=0}^{\infty} n p_n x^{n-1}$$

$$f''(x) = \sum_{n=0}^{\infty} n(n-1) p_n x^{n-2}$$

$$f'(1) = \sum_{n=0}^{\infty} n p_n = E[X]$$　平均値

$$f''(1) = \sum_{n=0}^{\infty} n(n-1) p_n = \sum_{n=0}^{\infty} n^2 p_n - \sum_{n=0}^{\infty} n p_n$$

$$= E[X^2] - f'(1)$$

$$E[X^2] = f''(1) + f'(1)$$

分散　$V[X] = E[X^2] - E[X]^2$

$$= f''(1) + f'(1) - f'(1)^2$$

二項分布の確率母関数

確率母関数

$$f(x) = \sum_{k=0}^{\infty} p_k x^k$$

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

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

$$= (px + 1 - p)^n$$

$$f'(x) = np(px + 1 - p)^{n-1}$$

$$f'(1) = np$$

$$f''(x) = n(n-1)p^2(px + 1 - p)^{n-2}$$

$$f''(1) = n(n-1)p^2$$

平均値　$E[X] = f'(1) = np$

分散　$V[X] = f''(1) + f'(1) - f'(1)^2$

$$= n(n-1)p^2 + np - (np)^2$$

$$= np$$

ポアソン分布

$\lambda > 0$ を定数とする.
確率変数 X がパラメータ λ のポアソン分布に従うとは,

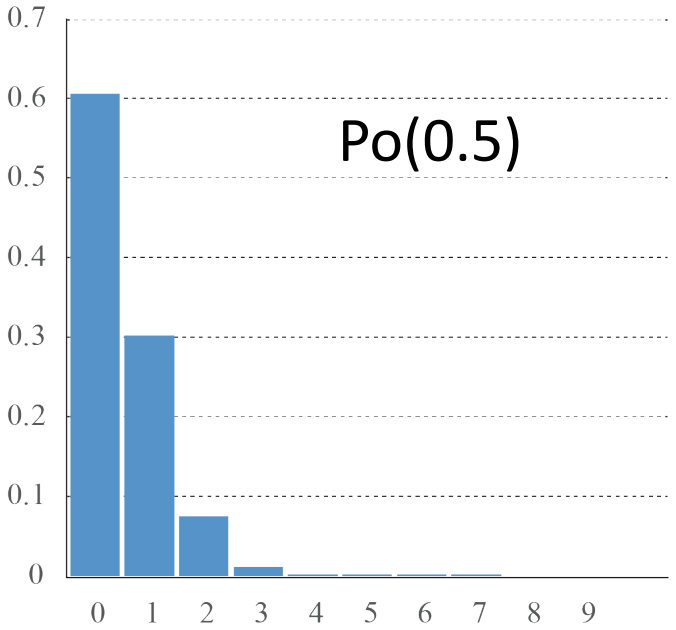
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad k = 0,1,2, \ldots$$

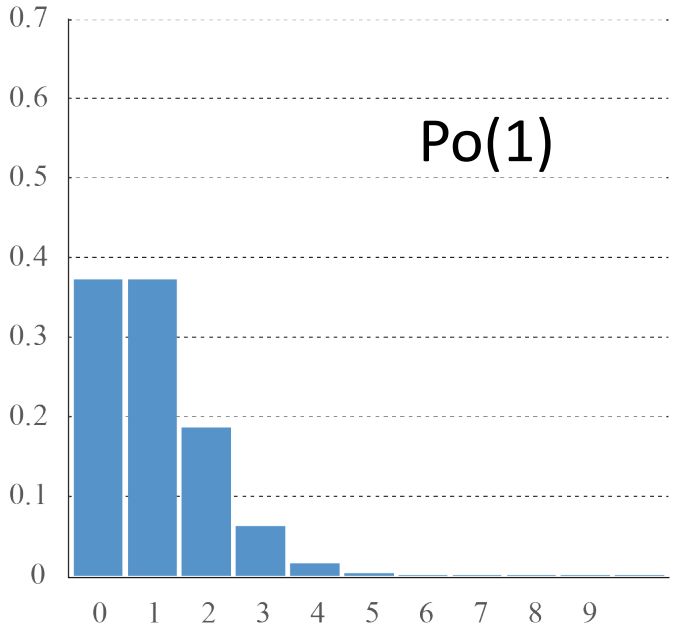
$X \sim \mathrm{Po}(\lambda)$ と書く.

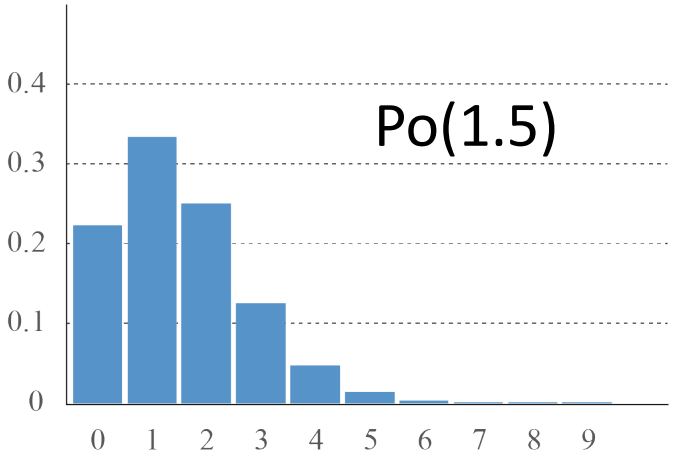
- 確かに確率分布になっている

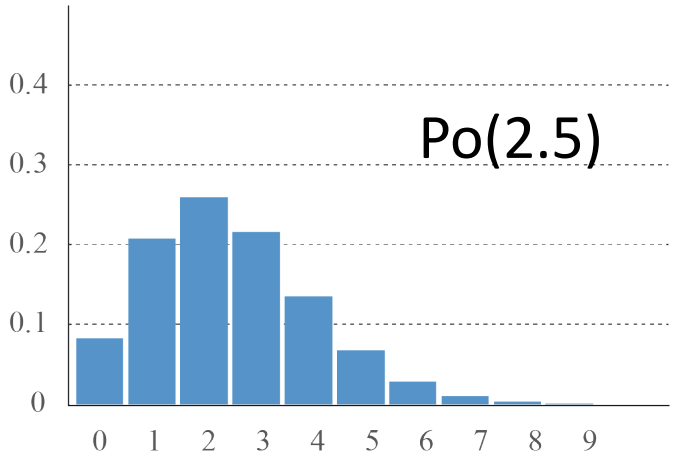
指数関数のテーラー展開

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \quad \Rightarrow \quad 1 = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}$$


Po(0.5)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9


Po(1)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9


Po(1.5)
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9


Po(2.5)
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9

ポアソン分布の確率母関数

確率分布

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

確率母関数

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

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

$$=e^{-\lambda}\,e^{\lambda x}$$

$$f'(x)=\lambda e^{-\lambda}e^{\lambda x} \qquad f'(1)=\lambda$$

$$f''(x)=\lambda^2 e^{-\lambda}e^{\lambda x} \qquad f''(1)=\lambda^2$$

平均値　$m=f'(1)=\lambda$

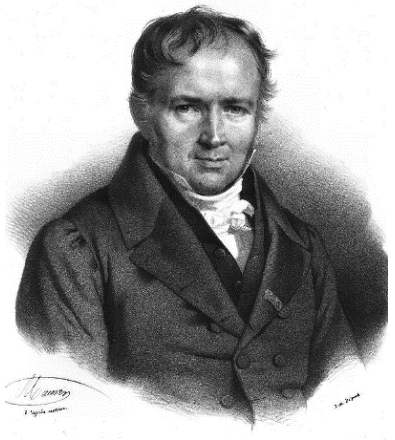
分散　$\sigma^2=f''(1)+f'(1)-f'(1)^2$

$$=\lambda^2+\lambda-\lambda^2=\lambda$$

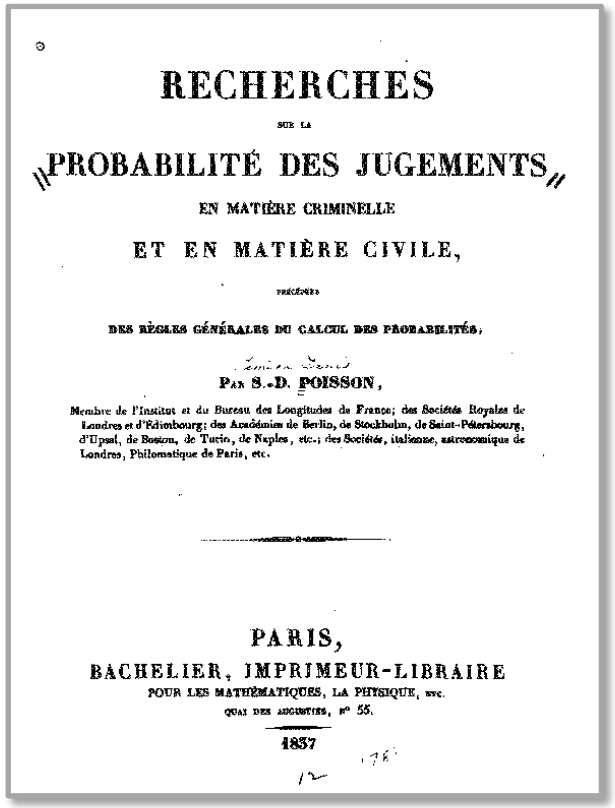
ポアソン分布の特徴

平均と分散が等しい

Siméon Denis Poisson (1781-1840)



- エコール・ポリテクニクでラプラスらに学ぶ.
- 数学・物理学に多大な貢献
- ポアソン○○と名のついた概念多数



RECHERCHES

SUR LA

PROBABILITÉ DES JUGEMENTS,

EN MATIÈRE CRIMINELLE

ET EN MATIÈRE CIVILE,

PRÉCÉDÉES

DES RÈGLES GÉNÉRALES DU CALCUL DES PROBABILITÉS,

PAR S.-D. POISSON,

Membre de l'Institut et du Bureau des Longitudes de France; des Sociétés Royales de Londres et d'Édimbourg; des Académies de Berlin, de Stockholm, de Saint-Pétersbourg, d'Upsal, de Boston, de Turin, de Naples, etc.; des Sociétés, italienne, astronomique de Londres, Philomatique de Paris, etc.

PARIS,

BACHELIER, IMPRIMEUR-LIBRAIRE

POUR LES MATHÉMATIQUES, LA PHYSIQUE, etc.

QUAI DES AUGUSTINS, n° 55.

1837

誤った有罪判決の回数について研究(1837)

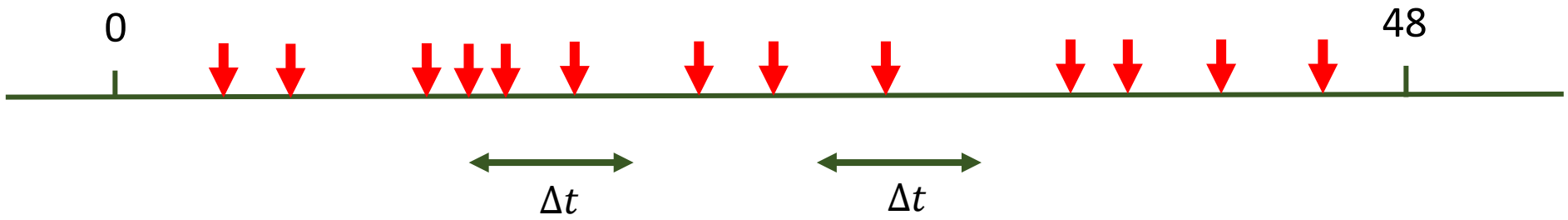
誤審が, ある一定期間に起こる回数 X の確率分布
（⇒ ポアソン分布）を導出

稀なイベントの発生回数の確率分布として広い応用

- 馬に蹴られて死亡した兵士数（ボルトキーヴィッチ）
- 電話の呼び出し, メールの着信
- サッカーのゴール数, 野球のホームラン数

メールの着信

メールの着信を観測して, 48時間に324回のメール着信があった．これをもとに, ある特定の 1 時間のメール着信回数を調べよう．

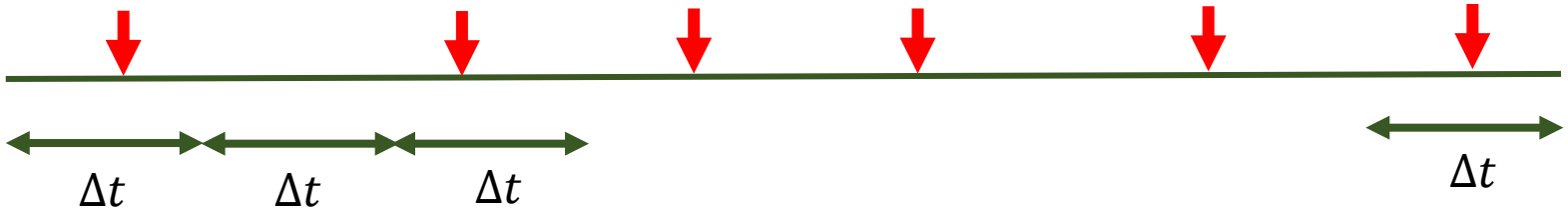


仮定

- 微小時間区間 Δt の中では着信は 1 回, または 0 回
- 微小時間区間 Δt がずれていれば着信の有無は独立

⇨ 微小時間区間 Δt 毎のメール着信をコイン投げとみなす

1時間当たりの着信回数 X をモデル化する



- 1時間を微小時間区間 Δt に N 分割

$$\Delta t = \frac{1}{N}$$

$p = \mathrm{P}$(時間区間 Δt にメール着信がある)

$1 - p = \mathrm{P}$(時間区間 Δt にメール着信がない)

- 1時間の間に N 回のコイン投げ

表の回数　$X \sim B(N, p)$

平均値 $\mu = Np$

- 観測から1時間当たり

$$\frac{324}{48} = 6.75 \text{ 回の着信}$$

- p の推定

$$\mu = Np = 6.75 \Rightarrow p = \frac{6.75}{N}$$

- こうして

$$X \sim B\left(N, \frac{6.75}{N}\right)$$

ポアソン分布の導出（ポアソンの少数の法則）

$B\left(N,\frac{\lambda}{N}\right)$ は $N\to\infty$ とすると $\mathrm{Po}(\lambda)$ に収束する.

言い換え　$B(N,p)$ は, $N\to\infty, p\to 0, Np\to\lambda$ とすると $Po(\lambda)$ に収束する.

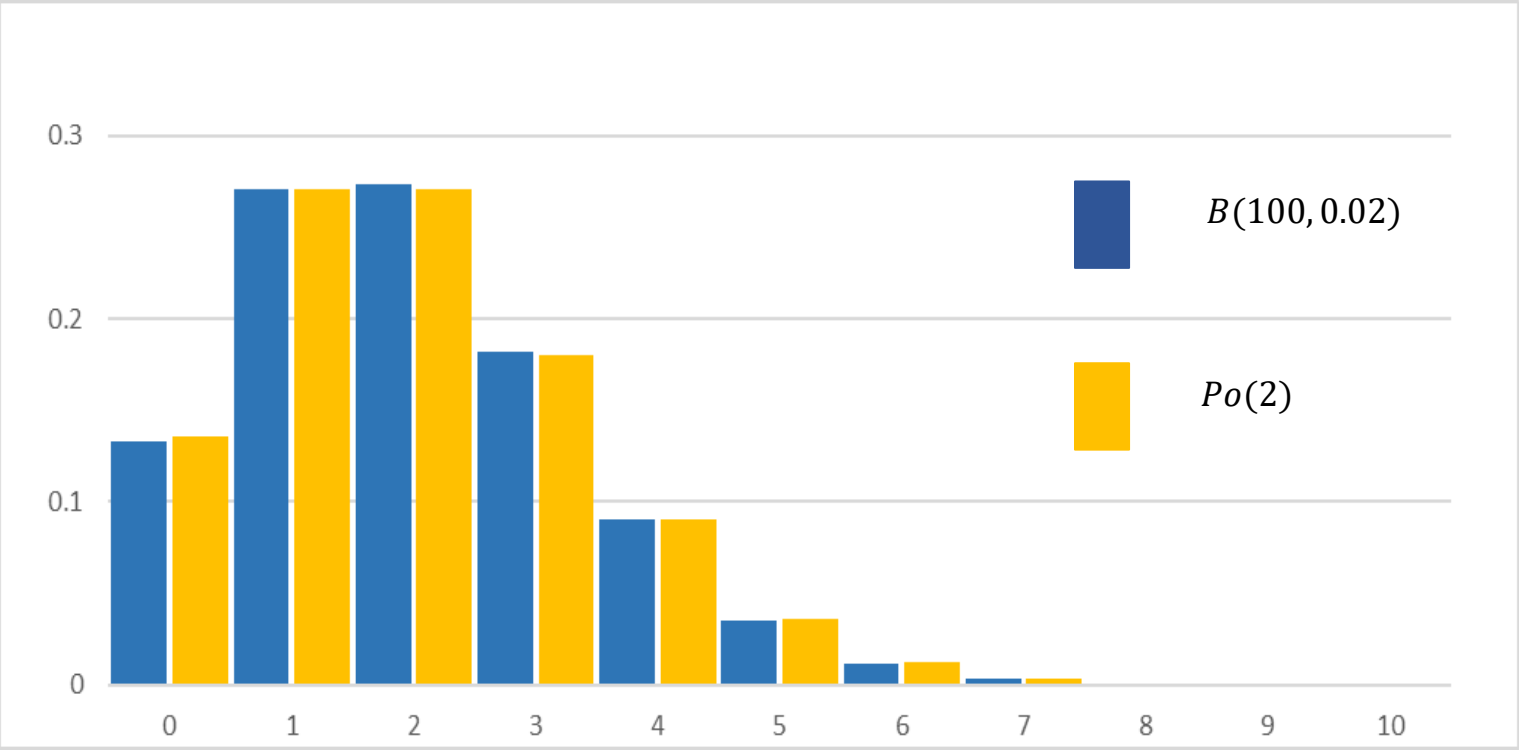
証 明　$X\sim B\left(N,\frac{\lambda}{N}\right)$ として

$$P(X=k)=\binom{N}{k}p^k(1-p)^{N-k}=\frac{N(N-1)\cdots(N-k+1)}{k!}\left(\frac{\lambda}{N}\right)^k\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-k}$$

$$\to\ \frac{\lambda^k}{k!}e^{-\lambda}\qquad(N\to\infty)$$

例　$B(100, 0.02) \approx Po(2)$

k	0	1	2	3	4	5	6	7	8	9
二項分布	0.13262	0.27065	0.27341	0.18228	0.09021	0.03535	0.01142	0.00313	0.00074	0.00015
ポアソン分布	0.13534	0.27067	0.27067	0.18045	0.09022	0.03609	0.01203	0.00344	0.00086	0.00019



メールの着信

メールの着信を観測して, 48時間に324回のメール着信があった．これをもとに, ある特定の 1 時間のメール着信回数を調べよう．

- X：1 時間当たりの着信回数
- 1時間を微小時間区間 Δt に N 分割してモデル化：　$X \sim B\left(N, \dfrac{6.75}{N}\right)$
- $N \to \infty$ として　$X \sim \mathrm{Po}(6.75)$
- たとえば

$$P(X \leq 2) = \frac{6.75^0}{0!} e^{-6.75} + \frac{6.75^1}{1!} e^{-6.75} + \frac{6.75^2}{2!} e^{-6.75}$$

$$= (1 + 6.75 + 22.78) \times 0.00117 = 0.036$$

N は人為的なので消去したい

Lecture 5

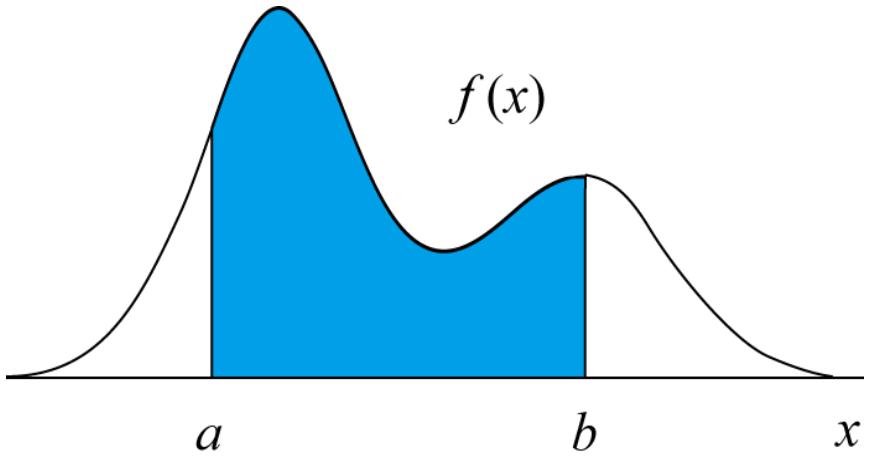
連続型確率分布

連続型確率変数 X

- 取りうる値は連続量
- 特定の値 a をとる確率：　$P(X = a) = 0$
- ある範囲（a より大きく b 以下）の値をとる確率：　$P(a < X \leq b)$

＜ でも ≤ でも同じ

- 確率密度関数を用いて面積で確率を表す



$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

$f(x)$ を X の確率密度関数という

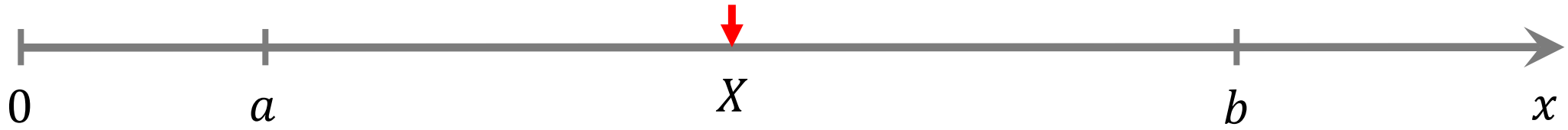
$$f(x) = f_X(x)$$

例 題 5.1

区間 $[a, b]$ からランダムに選ばれた点の座標を X とすれば，X は連続型確率変数になる．X の確率分布を求めよ．

例 題 5.1

区間 $[a, b]$ からランダムに選ばれた点の座標を X とすれば，X は連続型確率変数になる．X の確率分布を求めよ．

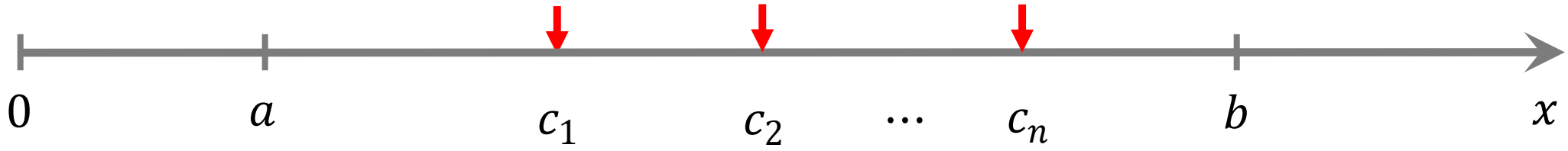


- 取りうる値（a 以上 b 以下の実数）は連続量
- ランダムって何？

1）決まった数学的定義はなく，文脈で理解する．

2）初めに思う状況は，「すべての根元事象が等確率で起こる」こと

➢ X が特定の値 c をとる確率：　$P(X = c) = 0$



n 個の異なる点を選ぶ：$c_1, c_2, \cdots, c_n$

$$P(X = c_1) = P(X = c_2) = \cdots = P(X = c_n) = p$$

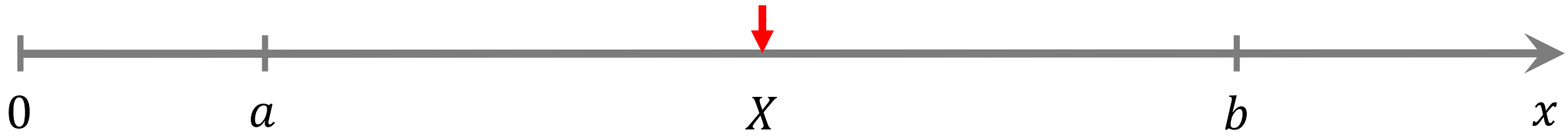
どの点も等確率で選ばれる

和事象の確率：　$P(X = c_1$ または $X = c_2$ または $\cdots X = c_n) = np$

確率は必ず ≤ 1 なので　$np \leq 1$　⇒　$p \leq \frac{1}{n}$

n はいくらでも大きくできるから　$p = 0$

➢ X が特定の値 c をとる確率：　$P(X = c) = 0$



➢ ある範囲（α より大きく β 以下）の値をとる確率：　$P(\alpha < X \leq \beta)$

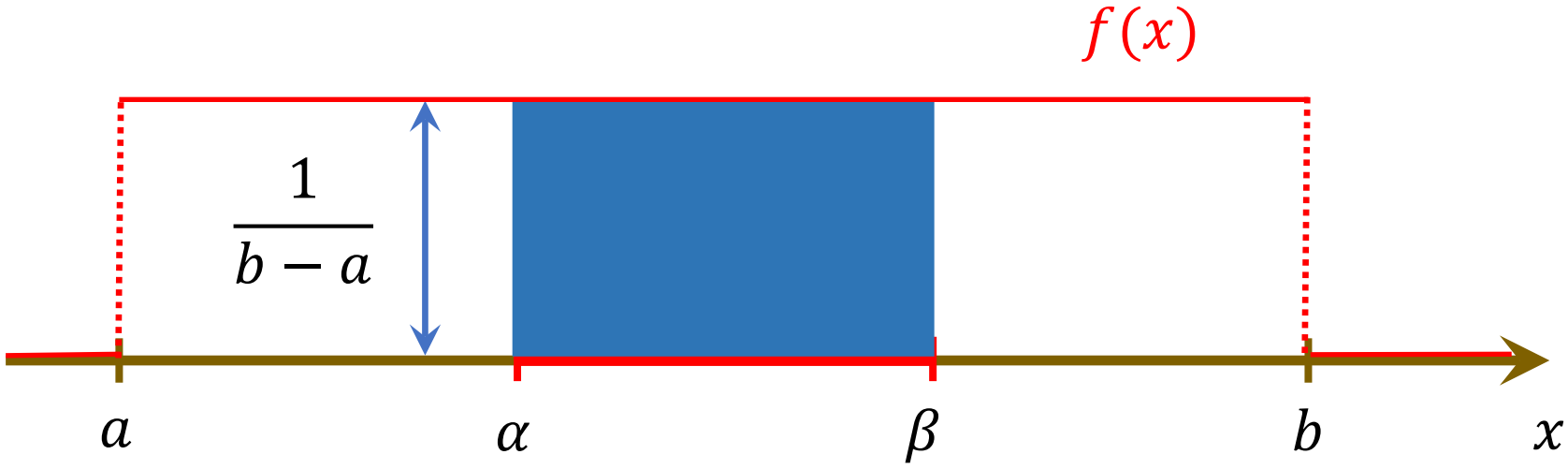
$$P(\alpha \leq X \leq \beta) = \frac{\beta - \alpha}{b - a}$$

長さの比

＜でも ≤ でも同じ

➢ どの点も同等に選ばれるという直感的状況が定式化できた！

➢ 確率 $P(\alpha \leq X \leq \beta)$ を面積で与える：



$$P(\alpha \leq X \leq \beta) = \frac{\beta - \alpha}{b - a} = \int_{\alpha}^{\beta} f(x)\, dx$$

➢ 確率 $P(\alpha \leq X \leq \beta)$ を面積で与える：

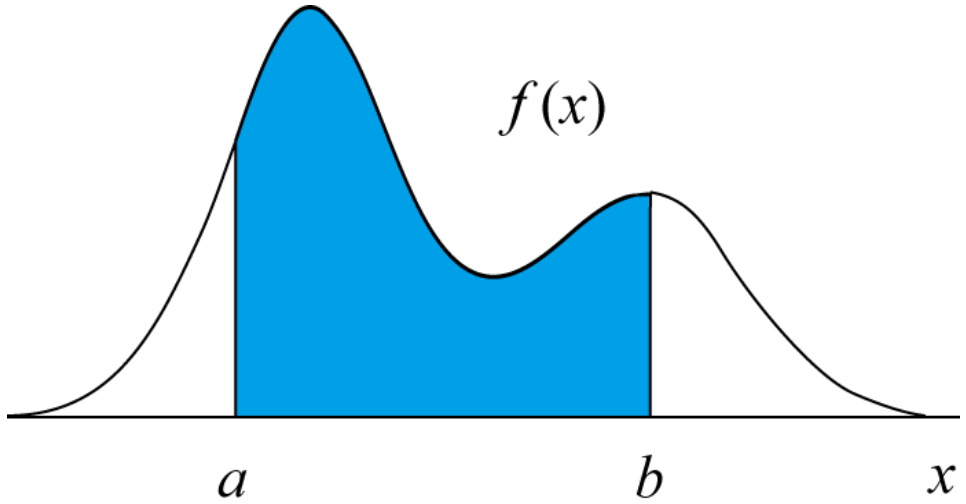
連続型確率変数 X の

確率密度関数は

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{その他} \end{cases}$$

一様分布という

確率密度関数



基本的な性質

(1)　$f(x) \geq 0$

(2)　$\int_{-\infty}^{\infty} f(x)dx = 1$

- 面積（定積分）で確率を表す

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

分布関数 (distribution function)

- 密度関数の補助として便利

定 義　確率変数 X が実数 x 以下の値をとる確率

$$F_X(x) = P(X \leq x)$$

を X の（累積）分布関数という. X は離散型でも連続型でもよい.

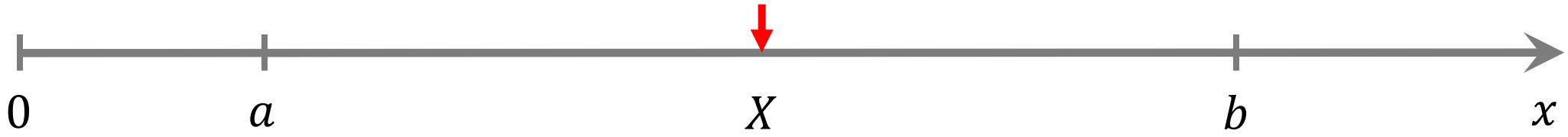
- X が密度関数 $f_X(x)$ をもつ連続型確率変数のとき,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \Longleftrightarrow \quad \frac{d}{dx}F_X(x) = F_X'(x) = f_X(x)$$

微分積分学の基本定理

例 題 5.1（再論）

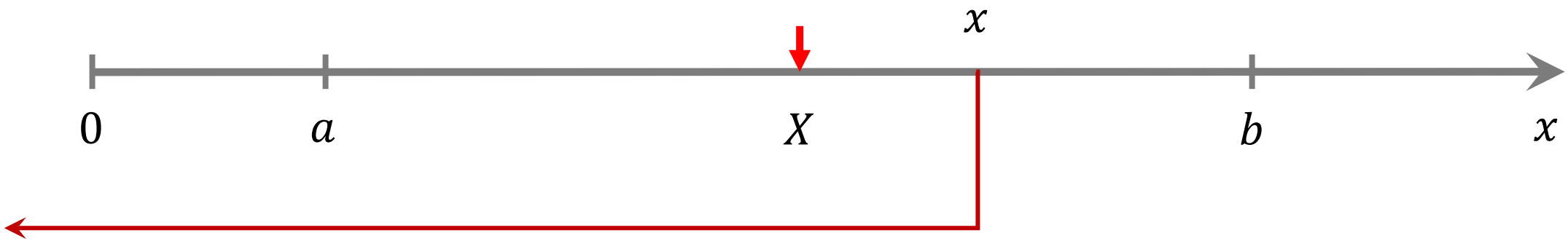
区間 $[a, b]$ からランダムに選ばれた点の座標を X とすれば，X は連続型確率変数になる．X の確率分布を求めよ．



➢ 分布関数を求める　　$F_X(x) = P(X \leq x)$

$x < a$ のとき $F(x) = 0$

$x > b$ のとき $F(x) = 1$



➢ 分布関数を求める

$$F_X(x) = P(X \leq x)$$

$x < a$ のとき $F_X(x) = 0$

$x > b$ のとき $F_X(x) = 1$

$\Rightarrow f_X(x) = 0$

$a \leq x \leq b$ のとき $F_X(x) = \dfrac{x-a}{b-a}$

$\Rightarrow f_X(x) = \dfrac{1}{b-a}$

➢ 密度関数を求める

$$f_X(x) = \frac{d}{dx} F_X(x) = F_X'(x)$$

連続型確率変数の平均値と分散

$f(x)$ ：連続型確率変数 X の密度関数

- 平均値　$\mathbf{E}[X] = \mu_X = \int_{-\infty}^{+\infty} x f(x)\, dx$

- 分散　$\mathbf{V}[X] = \sigma_X^2 = \int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x)\, dx = \mathbf{E}[(X-\mu_X)^2]$

$$= \int_{-\infty}^{+\infty} x^2 f(x)\, dx - \mu_X^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

分散公式

$$\mathbf{V}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

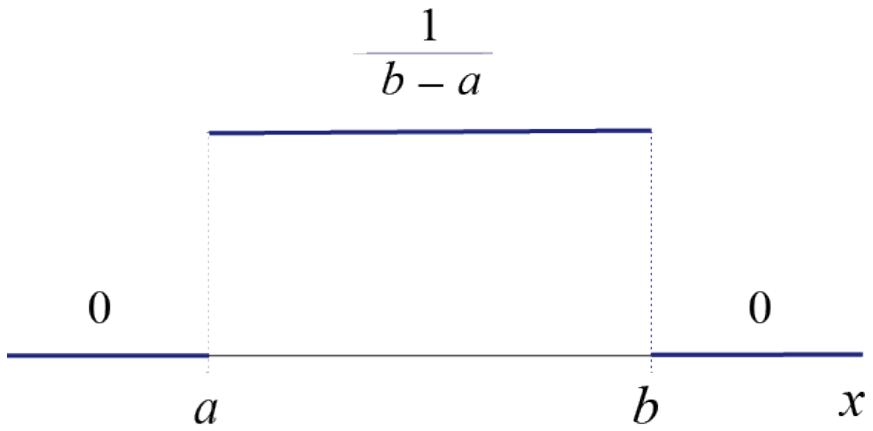
名前の付いた重要な連続分布

- ✓ 一様分布　二項母集団の分布
- ✓ 指数分布　待ち時間の分布
- ✓ 正規分布　統計学で最重要
- ✓ t-分布，　χ^2分布，　F分布　統計的推測で改めて扱う

一様分布

密度関数 $f(x)$ が区間 $[a, b]$ で与えられ, その値が一定であるような分布を一様分布という.

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{その他} \end{cases}$$



区間 $[a, b]$ から1点を, どの点も同程度の確からしさで選ぶとき, 選ばれた点の座標を X とする.

➡ X は連続型確率変数となり, その分布が一様分布

一様分布の平均値と分散

確率変数 X が区間 $[a, b]$ 上の一様分布に従うとき，

$$E[X] = \frac{a+b}{2} \qquad V[X] = \frac{(b-a)^2}{12}$$

平均値

$$\mu = \mathbf{E}[X] = \int_a^b x \frac{dx}{b-a} = \frac{a+b}{2}$$

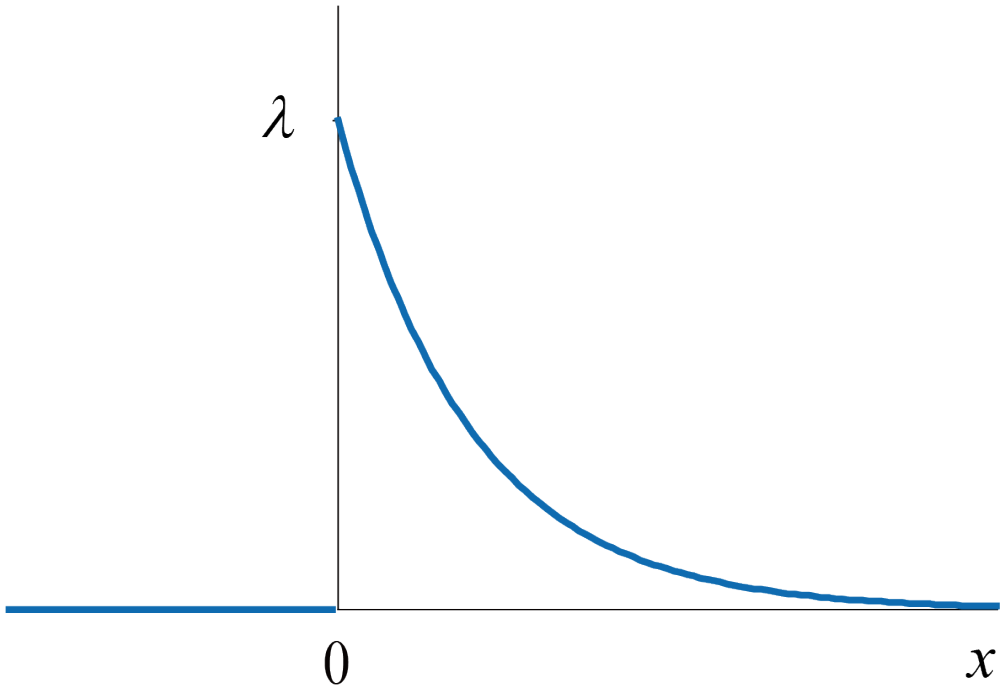
$$\mathbf{E}[X^2] = \int_a^b x^2 \frac{dx}{b-a} = \frac{a^2+ab+b^2}{3}$$

分散

$$\sigma^2 = \mathbf{E}[X^2] - \boldsymbol{E}[X]^2 = \frac{(b-a)^2}{12}$$

指数分布

$\lambda > 0$ を定数とする.
確率変数 X がパラメータ λ の指数分布に
従うとは, X が次の確率密度関数をもつこと



密度関数

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{その他} \end{cases}$$

平均値

$$\mu = \mathbf{E}[X] = \int_0^\infty x\, \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$\mathbf{E}[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

分散

$$\sigma^2 = \mathbf{E}[X^2] - \boldsymbol{E}[X]^2 = \frac{1}{\lambda^2}$$

正規分布＝ガウス分布 $N(\mu, \sigma^2)$

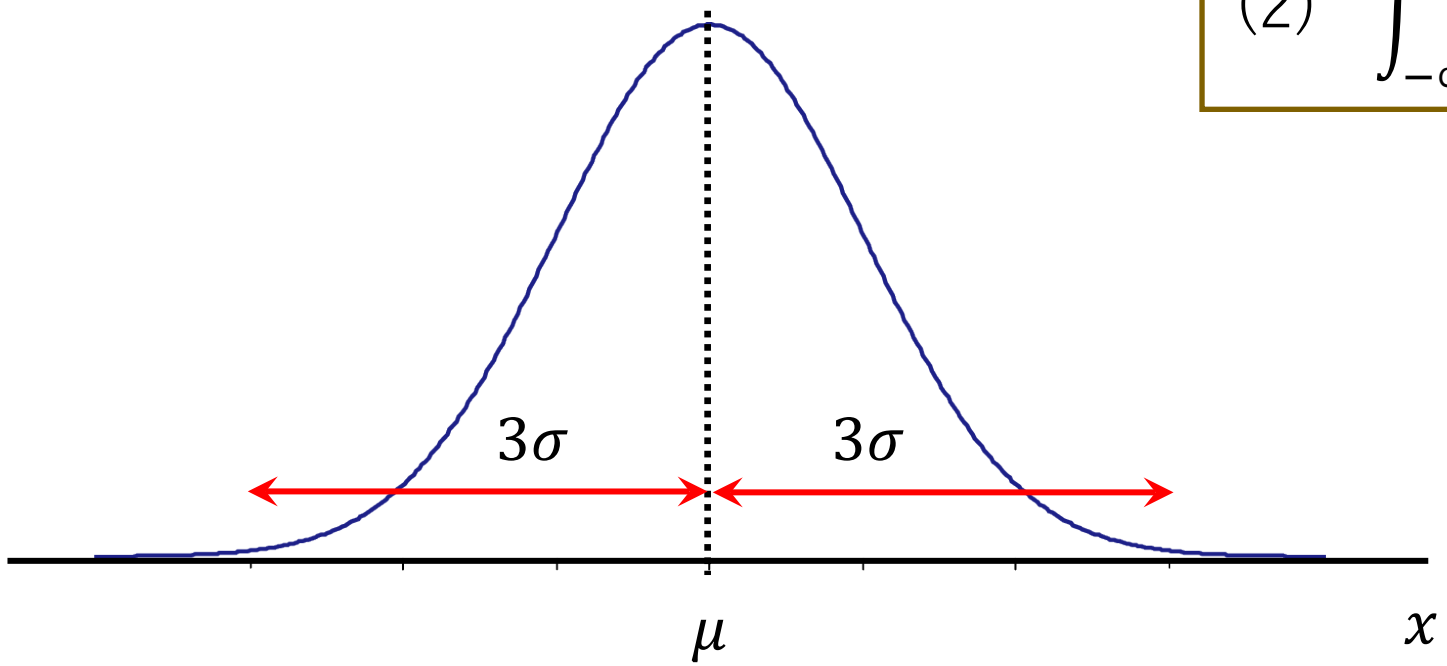
密度関数

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

基本的な性質

(1)　$f(x) \geq 0$

(2)　$\int_{-\infty}^{\infty} f(x)dx = 1$

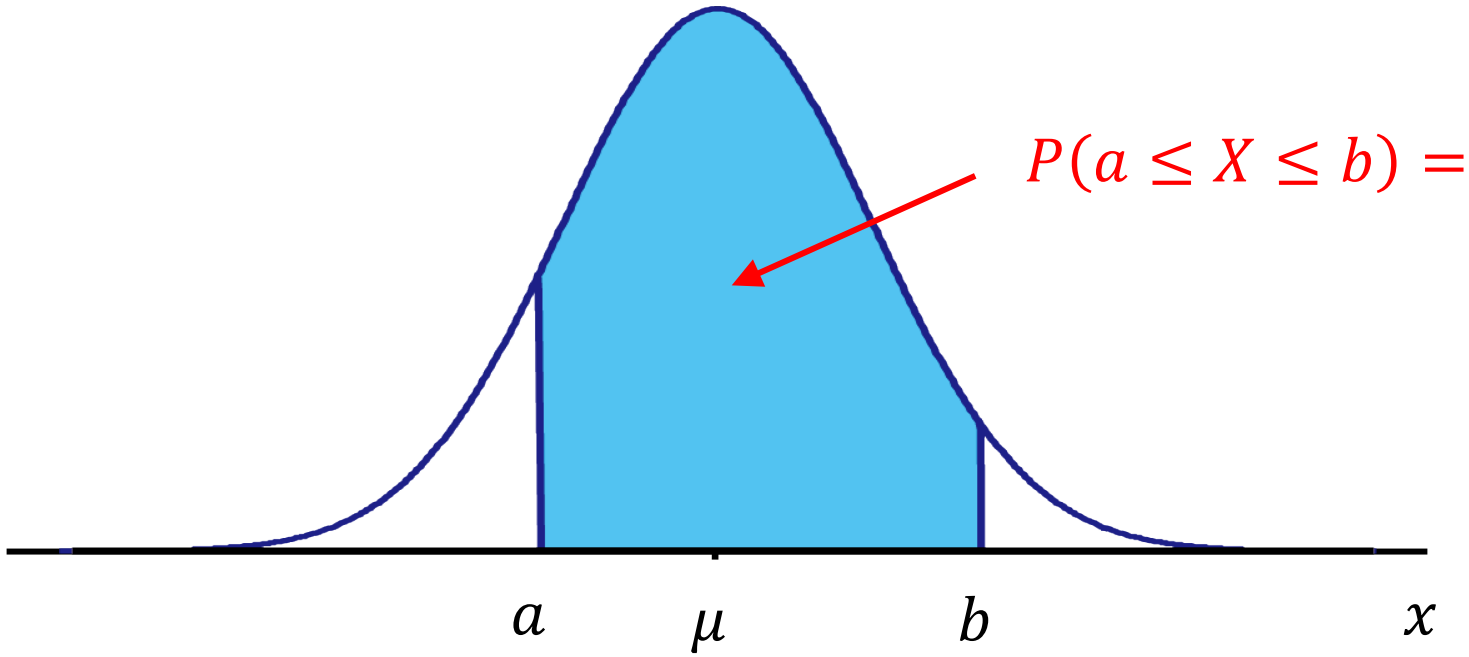


正規分布＝ガウス分布 $N(\mu, \sigma^2)$

密度関数

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

$X \sim N(\mu, \sigma^2)$



$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

正規分布の平均値と分散

密度関数　$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

密度関数

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

重積分と極座標の典型的練習問題

平均値

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

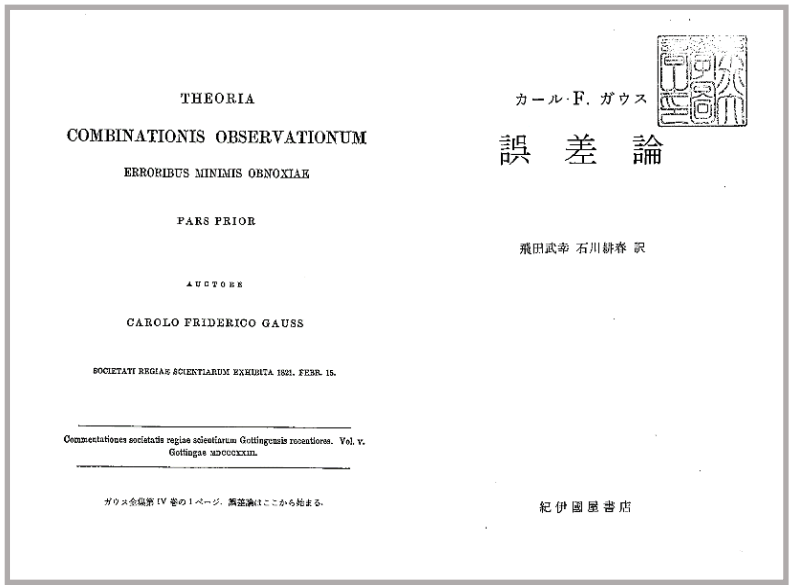
$N(\mu,\sigma^2)$

平均値と分散

分散

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

Johann Carl Friedrich Gauss (1777-1855)


THEORIA

COMBINATIONIS OBSERVATIONUM

ERRORIBUS MINIMIS OBNOXIAE

PARS PRIOR

AUCTORE

CAROLO FRIDERICO GAUSS

SOCIETATI REGIAE SCIENTIARUM EXHIBITA 1821. FEBR. 15.

Commentationes societatis regiae scientiarum Gottingensis recentiores. Vol. v.
Gottingae MDCCCXXIII.

ガウス全集第 IV 巻の 1 ページ．誤差論はここから始まる．

カール・F．ガウス

誤　差　論

飛田武幸 石川耕春 訳

紀伊國屋書店

19世紀以降の近代数学のほとんどの分野に影響を与えている：

- 素数定理、平方剰余の相互法則
- 正十七角形の作図、代数学の基本定理、円分多項式
- 最小二乗法(統計)
- 複素数、ガウス平面、複素関数論
- 電磁気などの物理学

1807年～終生
　ゲッティンゲンの天文台長（初代）

1809年 『天体運行論』
　最小二乗法を用いたデータ補正、正規分布

1821-26年 『誤差を最小にする観測』

標準正規分布 (standard normal distribution)

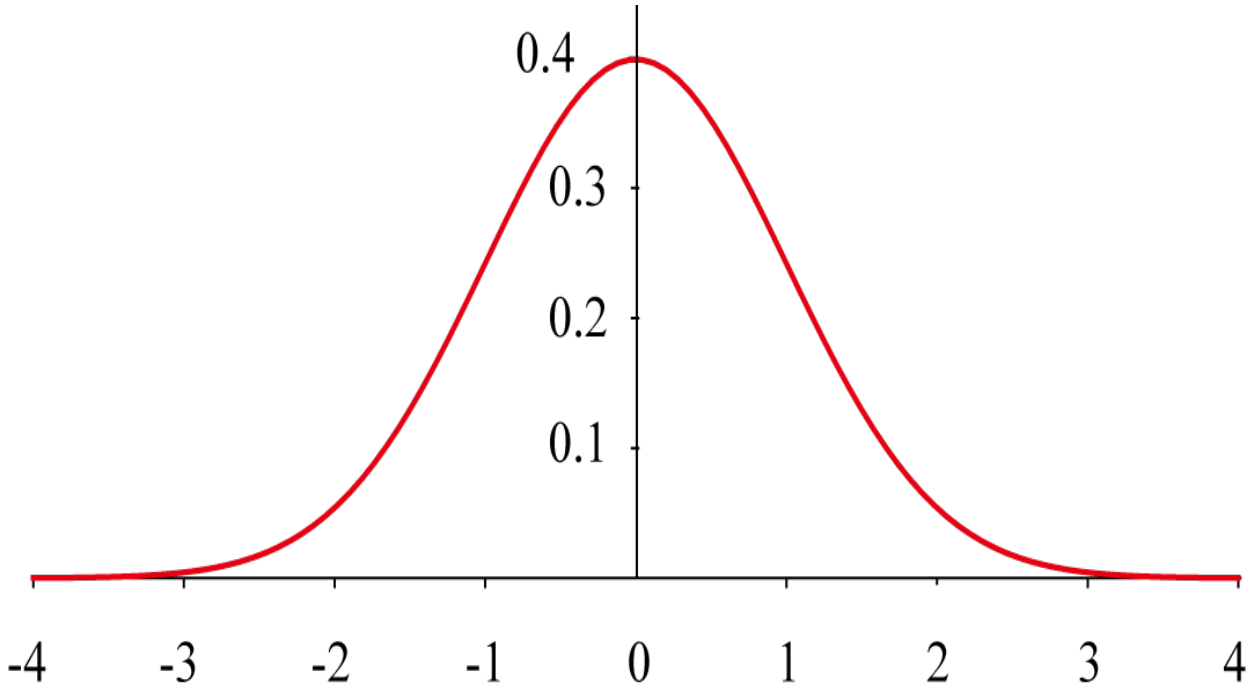
$N(0,1)$

平均値　$\mu = 0$

分散　$\sigma^2 = 1$

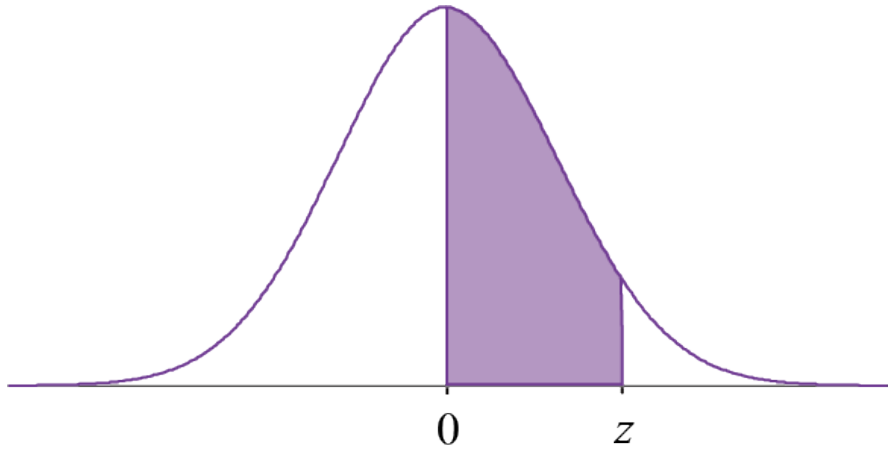
密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



標準正規分布表

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1311	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990



$$Z \sim N(0,1)$$

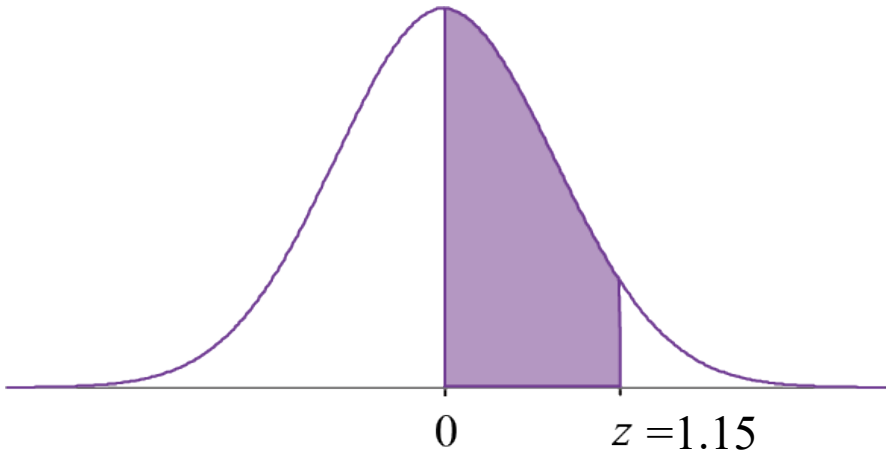
$$P(0 \leq Z \leq z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

この積分（$e^{-x^2/2}$ の原始関数）は初等関数ではない

標準正規分布表

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1311	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

例　$P(0 \leq Z \leq 1.15)$

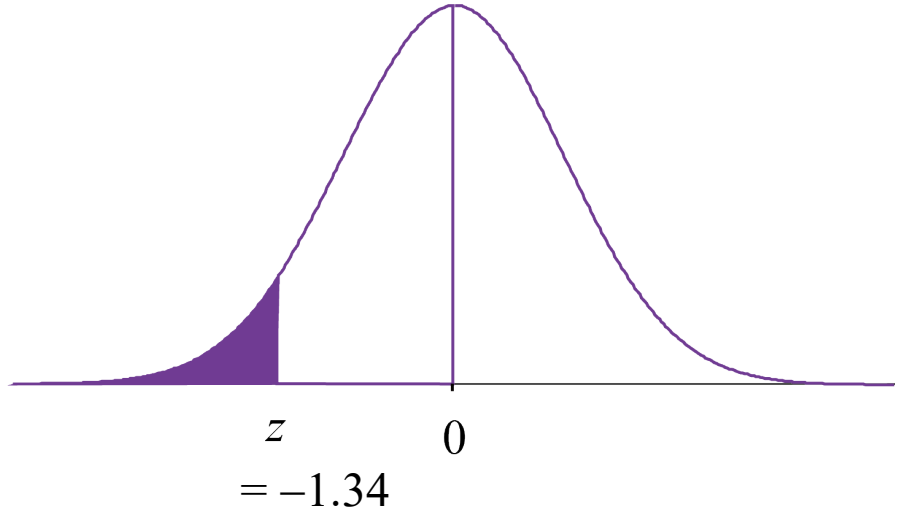


$$P(0 \leq Z \leq 1.15) = 0.3749$$

標準正規分布表

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1311	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

例　$P(Z \leq -1.34)$



$$P(Z \leq -1.34)$$

$$= 0.5 - P(0 \leq Z \leq 1.34)$$

$$= 0.5 - 0.4099$$

$$= 0.0901$$

定 理　確率変数 X が正規分布 $N(\mu,\sigma^2)$に従うとき, その 1 次変換 $a+bX$（a , b は定数）は正規分布 $N(a+b\mu\,,b^2\sigma^2)$ に従う.

$$X\sim N(\mu,\sigma^2) \Rightarrow a+bX\sim N(a+b\mu\,,b^2\sigma^2)$$

$$X\sim N(\mu,\sigma^2) \Rightarrow Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$　標準化

証 明　$X\sim N(\mu,\sigma^2)$ より

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

これを用いて, $Y=a+bX$ に対して $P(Y\le y)$ を求める

証 明　$X \sim N(\mu, \sigma^2)$ より

$$P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx \qquad \text{ただし，} \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

これを用いて，$Y = a + bX$ に対して $P(Y \leq y)$ を求める．

$b > 0$ の場合を扱う．（$b < 0$ の場合も同様, 不等式の向きに注意）

$$P(Y \leq y) = P(a + bX \leq y) = P\left(X \leq \frac{y-a}{b}\right) = \int_{-\infty}^{\frac{y-a}{b}} f(x)\,dx$$

$x = \frac{t-a}{b}$ によって変数変換すると，

$$= \int_{-\infty}^{\frac{y-a}{b}} f(x)\,dx = \int_{-\infty}^{y} f\left(\frac{t-a}{b}\right)\frac{dt}{b} = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi b^2\sigma^2}}\, e^{-\frac{(t-a-b\mu)^2}{2b^2\sigma^2}}\,dt$$

したがって，$Y \sim N(a + b\mu\,, b^2\sigma^2)$

定 理　確率変数 X が正規分布 $N(\mu, \sigma^2)$に従うとき, その 1 次変換 $Y = a + bX$ (a , b は定数)も正規分布 $N(a + b\mu , b^2\sigma^2)$ に従う.

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad a + bX \sim N(a + b\mu , b^2\sigma^2)$$

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$ 標準化

例　$X \sim N(2, 5^2)$ を標準化すると

$$Z = \frac{X - 2}{5} \sim N(0,1)$$

例 題 5.2　X が正規分布 $N(-1, 2^2)$ に従うとき，

（1）$P(X \leq 2.29)$ を求めよ．

（2）$P(X > x) = 0.01$ であるような x の値を求めよ．

例 題 5.2　X が正規分布 $N(-1, 2^2)$ に従うとき，

（1）　$P(X \leq 2.29)$ を求めよ．

（2）　$P(X > x) = 0.01$ であるような x の値を求めよ．

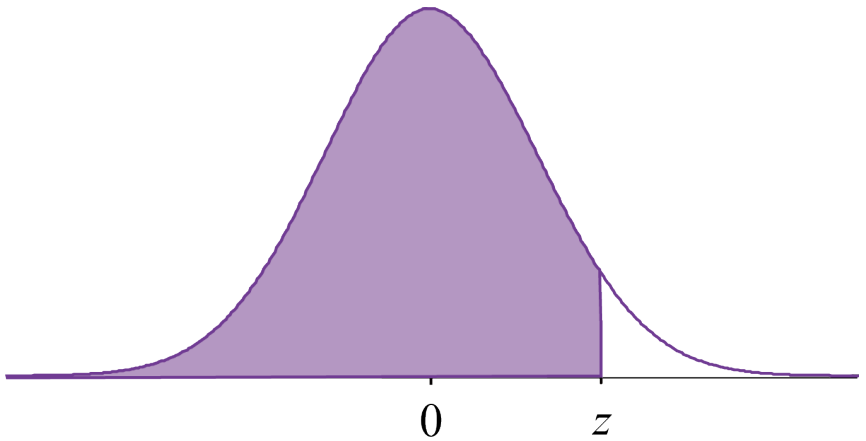
標準化　$X \sim N(\mu, \sigma^2)$ ➡ $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$

（1）　$P(X \leq 2.29) = P\left(\dfrac{X-(-1)}{2} \leq \dfrac{2.29-(-1)}{2}\right)$

$= P(Z \leq 1.645)$

$= 0.5 + 0.45$

$= 0.95$



例 題 5.2　X が正規分布 $N(-1, 2^2)$ に従うとき，

（1）　$P(X \leq 2.29)$ を求めよ．

（2）　$P(X > x) = 0.01$ であるような x の値を求めよ．

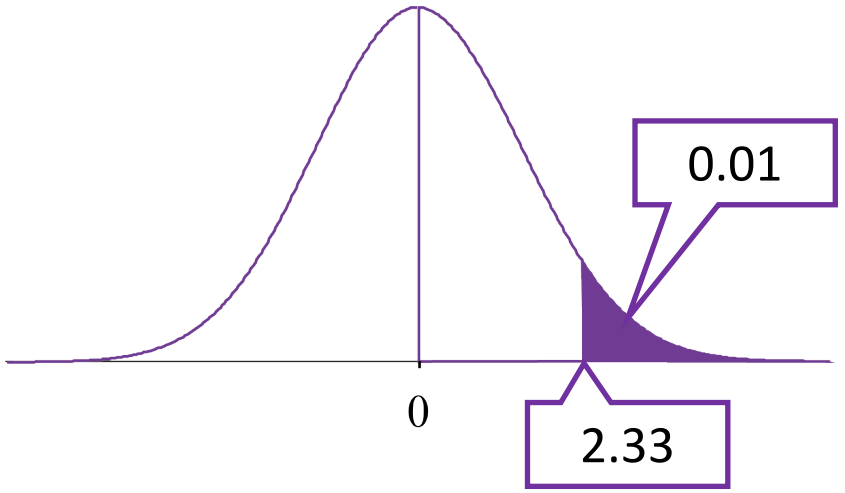
標準化　$X \sim N(\mu, \sigma^2)$ ⇒ $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$

（2）　$P(X > x) = P\left(\dfrac{X-(-1)}{2} > \dfrac{x-(-1)}{2}\right)$

$= P\left(Z > \dfrac{x+1}{2}\right)$ ｝ $\dfrac{x+1}{2} = 2.33$

一方，　$P(Z > 2.33) = 0.01$

$\therefore \quad x = 3.66$



例 題 5.3　ある年齢の女性の身長 [cm] は，平均値=156，標準偏差=5 の正規分布に従うと考えられるとき，153cm 以上 160cm 以下の女性は全体の何%を占めるか.

【演習（10分）】

例 題 5.3　ある年齢の女性の身長 [cm] は，平均値=156，標準偏差=5 の正規分布に従うと考えられるとき，153cm 以上 160cm 以下の女性は全体の何%を占めるか.

身長を X とすると，$X \sim N(156, 5^2)$

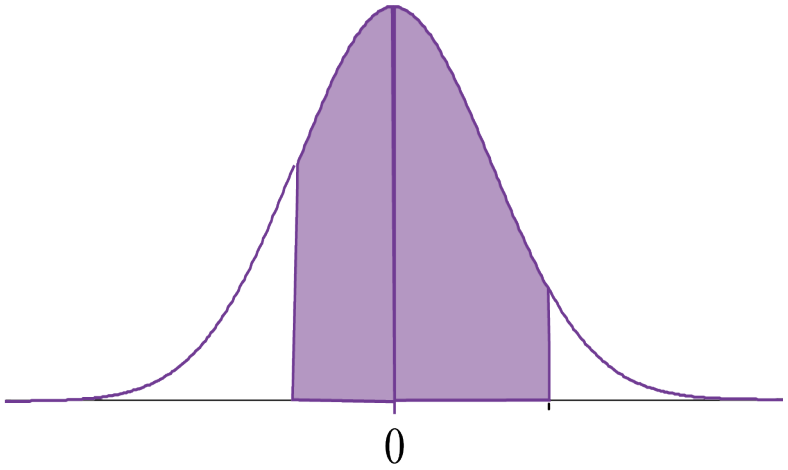
求める割合は $P\,(153 \leq X \leq 160)$

$$P\,(153 \leq X \leq 160) = P\left(\frac{153-156}{5} \leq \frac{X-156}{5} \leq \frac{160-156}{5}\right)$$

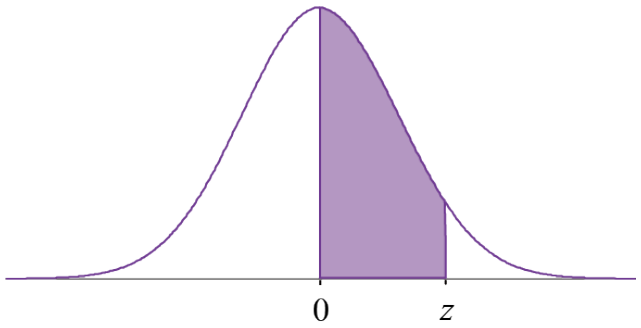
$$= P(-0.6 \leq Z \leq 0.8)$$

$$= P(0 \leq Z \leq 0.6) + P(0 \leq Z \leq 0.8)$$

$$= 0.2257 + 0.2881 = 0.5138$$



標準正規分布表



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1311	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Lecture 5

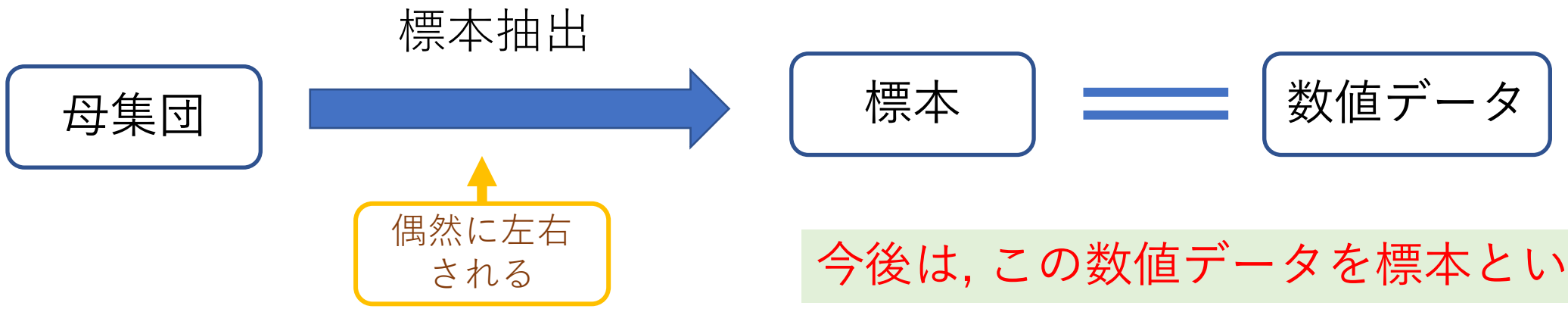
連続型確率分布

おわり

Lecture 6

標本抽出と正規分布

母集団と標本



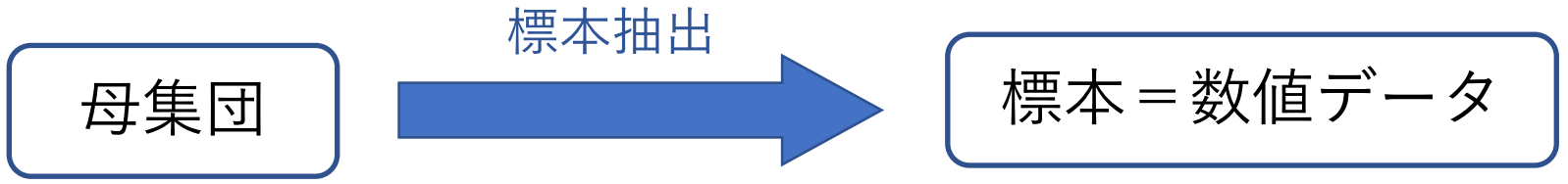
今後は，この数値データを標本という

母集団 (population) ＝調査対象の全体

実態として存在する集合（世論調査など）
または，理論上想定される（ふつうは無限）集合

課 題　得られた数値データから母集団の性質を信頼度付きで推定する

確率論による定式化



標本(sample)の取り出し方の基本：無作為復元抽出

- 母集団のどの要素も偏りなく選ばれる可能性を担保
- 取り出すたびに母集団は変化しない

1）母集団から n 個の標本 $x_1, x_2, \cdots, x_n$ を得る.

2）標本 x_i は確率変数 X_i として扱う.

3）無作為復元抽出なので，$X_1, X_2, \cdots, X_n$ はすべて母集団分布に従い，かつ独立である.

確率論による定式化

同一環境下における
繰り返し実験もこれ

定式化

n 個の標本は確率変数列 $X_1, X_2, \cdots, X_n$ であり,

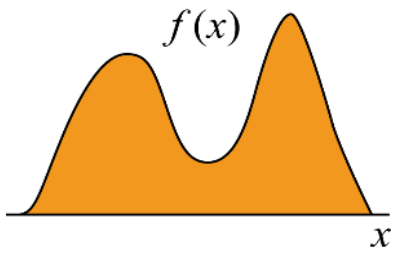
それらは独立で, 同分布である. その分布は,

$$X_i \sim \text{母集団分布}$$

実際得られる数値データ x_i は X_i の実現値と考える.

推測統計の目的

母集団



母集団分布

無作為
復元抽出

標本

$X_1, X_2, \ldots, X_n$

実際に知りたいのは，

母数 θ ＝母集団の統計量

平均値，分散，そのほか

θ を標本 $X_1, X_2, \ldots, X_n$ を用いて

何らかの合理的な計算で求める．

$$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$$

課題　$\hat{\theta}$ の確率分布を調べて，
信頼度付きで θ を推定する．

確率変数の和と積

$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$ のような確率変数の関数を扱う準備

確率変数 X, Y と 定数 a, b に対して

$X + Y$ ：和

XY ：積

aX ：スカラー倍（スカラー積）

$aX + bY$ ：線形和（線形結合）

定理　（平均値の線形性）

確率変数 X, Y と 定数 a に対して

（１）$E[aX] = aE[X]$

（２）$E[X + Y] = E[X] + E[Y]$

例　（１）X をサイコロの目とすると,

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

Y をサイコロの目を 10 倍して与えられる点数とすると,

$$E[Y] = E[10X] = 10E[X] = 10 \times 3.5 = 35$$

定理　（平均値の線形性）

確率変数 X, Y と 定数 a に対して

（１）$E[aX] = aE[X]$

（２）$E[X + Y] = E[X] + E[Y]$

例　（２）サイコロとコインを同時に投げて，
X をサイコロの目，Y をコインの表(1)裏(0)すると，

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5 \qquad E[Y] = \frac{1}{2}(0 + 1) = 0.5$$

Z をサイコロの目とコインを合わせた点数とすると，

$$E[Z] = E[X + Y] = E[X] + E[Y] = 3.5 + 0.5 = 4$$

証 明　離散型の場合を扱う（連続型では積分が必要）

平均値の定義

$$E[X] = \sum x_i P(X = x_i)$$

（1） X の取りうる値が x_i なら aX の取りうる値は ax_i

$$E[aX] = \sum ax_i P(X = x_i) = a \sum x_i P(X = x_i) = aE[X]$$

（2） X, Y の取りうる値が x_i, y_j なら $X + Y$ の取りうる値は $x_i + y_j$

$$E[X + Y] = \sum (x_i + y_j) P(X = x_i, Y = y_j) = \sum x_i P(X = x_i, Y = y_j) + \sum y_j P(X = x_i, Y = y_j)$$

第 1 項は,

$$\sum_{i,j} x_i P(X = x_i, Y = y_j) = \sum_i x_i \sum_j P(X = x_i, Y = y_j) = \sum_i x_i P(X = x_i) = E[X]$$

第 2 項も同様に変形して,　　$E[X + Y] = E[X] + E[Y]$

定理　確率変数 X, Y と 定数 a に対して,

（１）$V[aX] = a^2 V[X]$

（２）$V[X+Y] = V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$

ただし, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$ を共分散という.

分散公式

$$V[X] = E[X^2] - E[X]^2$$

証明　（１）

$$\begin{aligned} V[aX] &= E[(aX)^2] - E[aX]^2 \\ &= E[a^2X^2] - (aE[X])^2 \\ &= a^2E[X^2] - a^2E[X]^2 \\ &= a^2(E[X^2] - E[X]^2) \\ &= a^2V[X] \end{aligned}$$

定理　確率変数 X, Y と 定数 a に対して,

（１）$V[aX] = a^2 V[X]$

（２）$V[X+Y] = V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$

ただし, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$ を共分散という.

分散公式

$$V[X] = E[X^2] - E[X]^2$$

証明　（２）　$V[X+Y] = E[(X+Y)^2] - E[X+Y]^2$

$$= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2$$

$$= E[X^2] + 2E[XY] + E[Y^2] \ -E[X]^2 - 2E[X]E[Y] - E[Y]^2$$

$$= V[X] + V[Y] + 2(E[XY] - E[X]E[Y])$$

$$= V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$$

定理　(独立な確率変数)

確率変数 X, Y が独立であれば,

（1）$E[XY] = E[X]E[Y]$

（2）$V[X+Y] = V[X] + V[Y]$

例　サイコロ2回投げて出た目を X, Y とすると,

$$E[X] = E[Y] = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

Z をサイコロ2回投げて出た目の積とすると,

$$E[Z] = E[XY] = E[X]E[Y] = 3.5 \times 3.5 = 12.25$$

証 明　離散型の場合を扱う（連続型では積分が必要）

（１）X, Y の取りうる値が x_i, y_j なら XY の取りうる値は $x_i y_j$

$$E[XY] = \sum x_i y_j P(X = x_i, Y = y_j)$$

$$= \sum x_i y_j P(X = x_i) P(Y = y_j)$$

X, Y の独立性

$$= \sum x_i P(X = x_i) \sum y_j P(Y = y_j) = E[X]E[Y]$$

（２）　確率変数の和の分散　$V[X + Y] = V[X] + V[Y] + 2\mathrm{Cov}(X, Y)$

共分散　$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$

X, Y が独立であれば，　$E[XY] = E[X]E[Y]$ なので，　$\mathrm{Cov}(X, Y) = 0$

よって，　$V[X + Y] = V[X] + V[Y]$

正規確率変数の線形結合

定理　２つの正規確率変数 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ が独立であるとする.

このとき, 定数 a, b に対して, $aX + bY$ も正規確率変数であって,

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

- $aX + bY$ の分布が正規分布になることを証明するには, 密度関数の 積分を扱う必要がある（参考書等を見よ）
- ここでは, 平均値と分散だけを確認しておく.

$$E[aX + bY] = E[aX] + E[bY] = aE[X] + bE[Y] = a\mu_1 + b\mu_2$$

$$V[aX + bY] = V[aX] + V[bY] = a^2V[X] + b^2V[Y] = a^2\sigma_1^2 + b^2\sigma_2^2$$

標本平均は確率変数である

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

標本
$X_1, X_2, \ldots, X_n$

$X_1, X_2, \ldots, X_n$ は確率変数になる
（無作為標本ともいう）

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

【重要】　$\bar{X}$ も確率変数である

- 平均値や分散は？
- 確率分布は？

標本平均の平均値と分散

定 理　母平均 μ, 母分散 σ^2 の母集団から取り出した大きさ n の無作為標本

$X_1, X_2, \ldots, X_n$ の標本平均　$\bar{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$　に対して,

平均値：　$E[\bar{X}] = \mu$　　分散：　$V[\bar{X}] = \dfrac{\sigma^2}{n}$

証 明

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}E\left[\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

$$V[\bar{X}] = V\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}V\left[\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n} V[X_k] = \frac{1}{n^2}\sum_{k=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

標本平均の分布（正規母集団の場合）

正規母集団
$N(\mu, \sigma^2)$

→ 無作為復元抽出 →

標本
$X_1, X_2, \cdots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

定 理

正規母集団 $N(\mu, \sigma^2)$ から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

標本平均の分布（正規母集団の場合）

正規母集団
$N(\mu, \sigma^2)$

無作為復元抽出

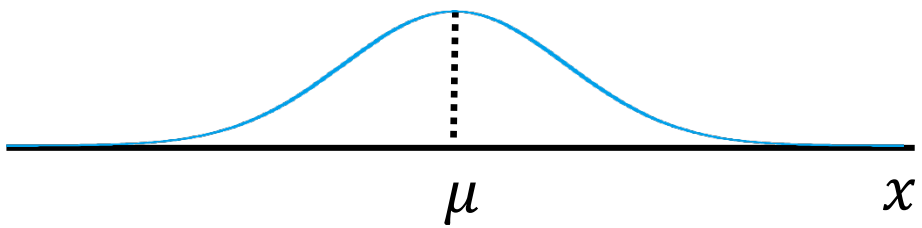
標本
$X_1, X_2, \cdots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

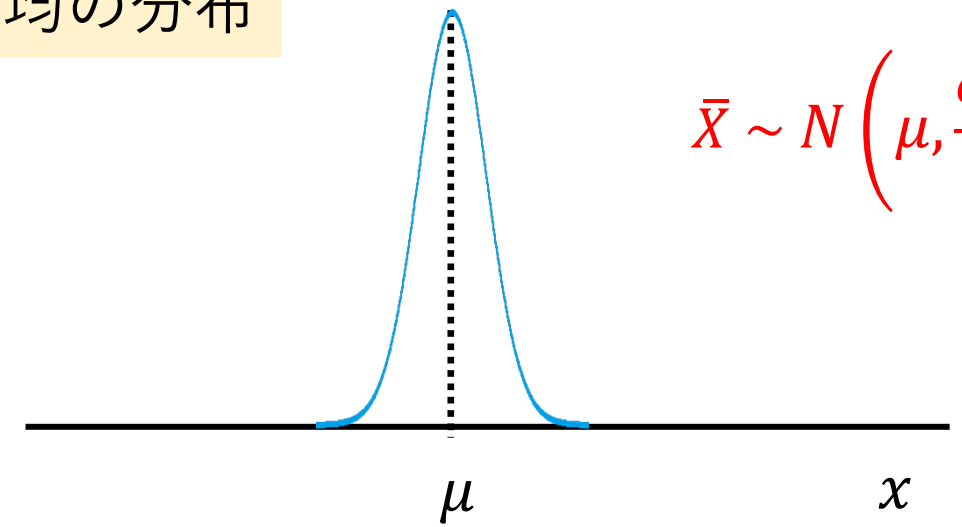
母集団分布

$$X_k \sim N(\mu, \sigma^2)$$



標本平均の分布

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



標本平均の分布（一般の母集団の場合）

一般の母集団
μ, σ^2

無作為復元抽出

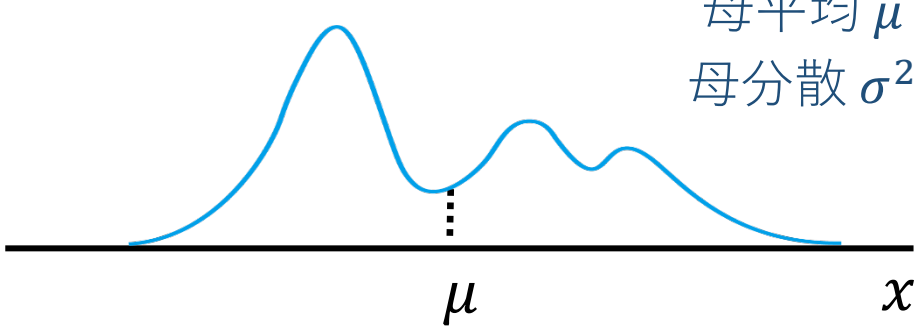
標本
$X_1, X_2, \cdots, X_n$

標本平均

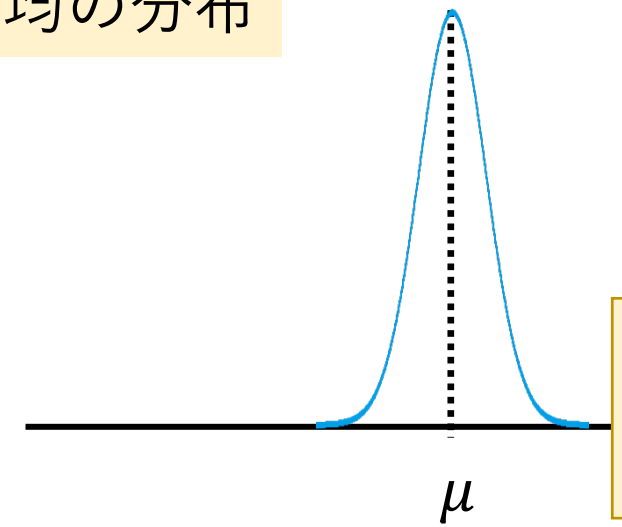
$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

母集団分布

$X_k \sim$ 母集団分布



標本平均の分布



$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

大きな n で近似的に成り立つ

中心極限定理（CLT）

定理　母平均 μ, 母分散 σ^2 のいっぱんの母集団から取り出した大きさ n の

無作為標本 $X_1, X_2, \ldots, X_n$ の標本平均　$\bar{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$　に対して,

n が大きいときは近似的に,

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \to \infty$$

- 二項母集団のとき：ベルヌーイ, ド・モアブル, …
- 一般の母集団：ラプラス（高度な微積分）

分布の特性関数

≈フーリエ変換 ≈ ラプラス変換 ≈ 連続分布に対する母関数

$$\varphi_X(t) = \int_{-\infty}^{+\infty} f_X(x) e^{itx}\, dx = E[e^{itX}]$$

- 特性関数は分布を一意的に決める
- 標準正規分布 $N(0,1)$ の特性関数

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} e^{itx}\, dx = e^{-t^2/2}$$

証 明

標本平均 $\bar{X}$ の標準化 Z

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Z_k$$

$$E[Z_k] = 0 \qquad V[Z_k] = E\left[Z_k^2\right] = 1$$

標準化 Z の特性関数

$$\varphi_Z(t) = E[e^{itZ}]$$

$$= E\left[\prod_{k=1}^{n} e^{it\frac{Z_k}{\sqrt{n}}}\right]$$

$$= \prod_{k=1}^{n} E\left[e^{it\frac{Z_k}{\sqrt{n}}}\right]$$

$$E\left[e^{it\frac{Z_k}{\sqrt{n}}}\right] = E\left[1 + it\frac{Z_k}{\sqrt{n}} + \frac{1}{2}\left(it\frac{Z_k}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right]$$

$$= 1 + 0 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

$$\varphi_Z(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-t^2/2} \qquad (n \to \infty)$$

これは標準正規分布 $N(0,1)$ の特性関数

Z の分布 $\to N(0,1)$　　$(n \to \infty)$

$$Z \approx N(0,1)$$

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

例題 6.1　サイコロを 50 回振るとき, 出目の平均値 $\bar{X}$ はどのような分布に従うと考えられるか.

例題 6.1　サイコロを 50 回振るとき, 出目の平均値 $\bar{X}$ はどのような分布に従うと考えられるか.

母集団
母平均 μ
母分散 σ^2

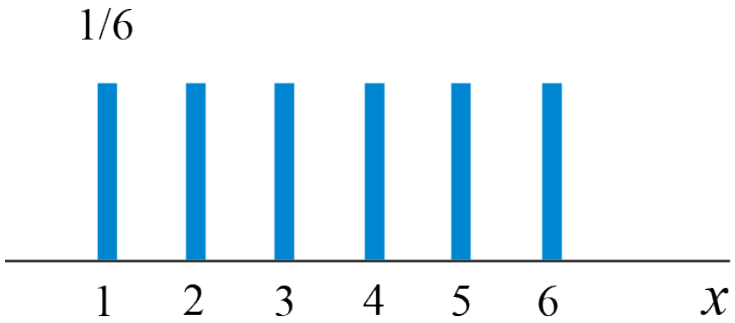
→ 無作為復元抽出 →

標本
$X_1, X_2, \ldots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$X_k \sim$ 母集団分布



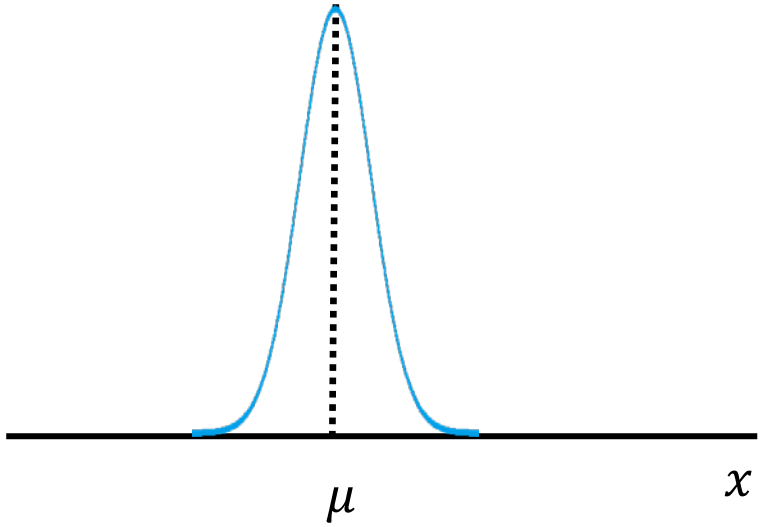
$$\mu = \frac{7}{2} \qquad \sigma^2 = \frac{35}{12}$$

標本平均の分布

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$= N\left(\frac{7}{2}, \frac{35}{12n}\right)$$

$$= N\left(\frac{7}{2}, \frac{35}{600}\right)$$



二項分布の正規分布近似

定 理（ドモアブル-ラプラス）

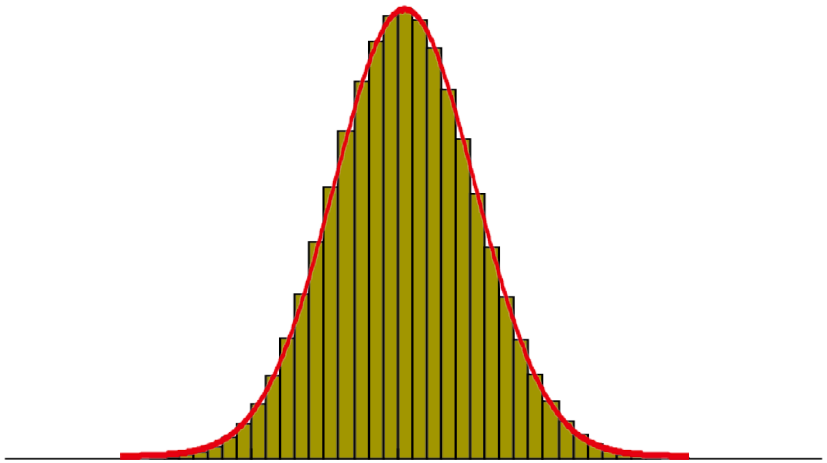
二項分布 $B(n,p)$ は，n が大きいとき，同じ平均値と分散をもつ正規分布 $N(np\,,np(1-p))$ で近似できる．

証 明　$Z_1, Z_2, \cdots, Z_n, \cdots$ ：ベルヌーイ試行列．　平均値 $= p$，　分散 $= p(1-p)$

$X_n = Z_1 + Z_2 + \cdots + Z_n \sim B(n,p)$　これは定義！

一方，CLT により，　$\displaystyle\sum_{k=1}^{n} Z_k \approx N(np, np(1-p))$

したがって，　$B(n,p) \approx N(np, np(1-p))$



例 題 6.2　公平なコインを400回投げたとき，表が225回以上出る確率を求めよ.

例 題 6.2　公平なコインを400回投げたとき，表が225回以上出る確率を求めよ.

X：表の枚数　$X \sim B(400,0.5) \approx N(200, 10^2)$

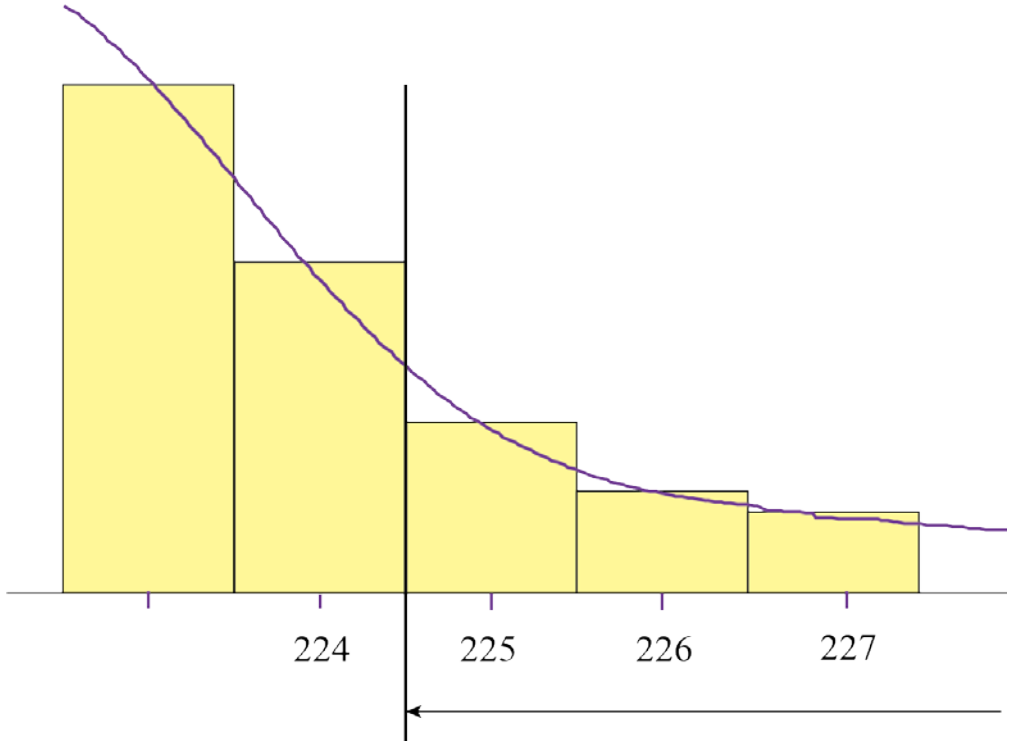
$P(X \geq 225) = P(X \geq 224.5)$

連続補正（半目補正）

$$= P\left(\frac{X-200}{10} \geq \frac{224.5-200}{10}\right)$$

$$= P(Z \geq 2.45)$$

$$= 0.5 - 0.4929 = 0.0071$$



中心極限定理（CLT）の変形

定 理　（中心極限定理＝CLT）

一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は，n が大きいときは近似的に，

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{X}$ の標準化（z-変換）

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

よく使う変形

$$\sum_{k=1}^{n} X_k \approx N(n\mu, n\sigma^2)$$

$$\sum_{k=1}^{n} X_k - n\mu \approx N(0, n\sigma^2)$$

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} \approx N(0,1)$$

$$Z_k = \frac{X_k - \mu}{\sigma}$$　（標準化）

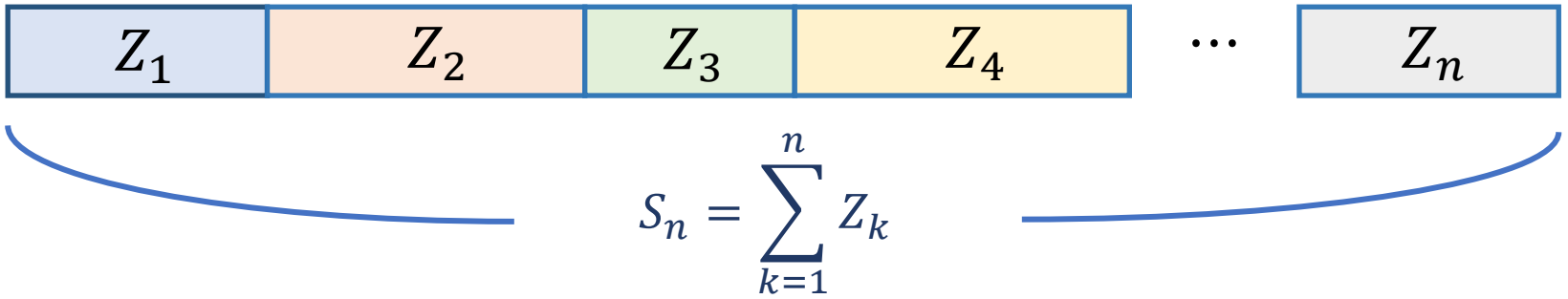
$$E[Z_k] = 0, \quad V[Z_k] = 1$$

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k \approx N(0,1)$$

実世界の揺らぎは細かい誤差の集積

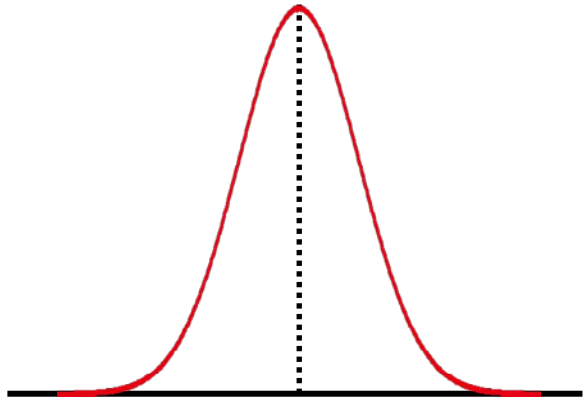
$Z_1, Z_2, \cdots, Z_n, \cdots$　：独立同分布な確率変数列

それぞれの平均値 $= \mu$,　分散 $= \sigma^2$



CLTによって,

$$S_n - n\mu = S_n - E[S_n] \approx N(0, n\sigma^2)$$



つまり，S_n は平均値のまわりに正規分布に従って分布する

大数の法則 (LLN)

定 理（中心極限定理＝CLT）

母平均 μ，母分散 σ^2 の一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は，n が大きいときは近似的に，

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \to \infty$$

$n \to \infty$ で分散がゼロになるので，揺らぎが消えて μ に収束することが示唆される．このことは，次のように数学の定理として証明される（やや高度）

定 理（大数の法則＝LLN）

母平均 μ の一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ について，

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mu\right) = 1$$

コイン投げのシミュレーション

Jakob Bernoulli (1654-1705)

コイン投げ：$Z_1, Z_2, \cdots, Z_n, \cdots$

$Z_k = 1$ (表のとき)，　$Z_k = 0$ (裏のとき)，

母平均と母分散（$= Z_k$の平均値と分散）

$$\mu = E[Z_k] = 1 \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{2}$$

$$\sigma^2 = E\left[Z_k^2\right] - E[Z_k]^2 = \mu - \mu^2 = \frac{1}{4}$$

CLTによって

$$T_n = \boxed{\frac{1}{n}\sum_{k=1}^{n} Z_k} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\frac{1}{2}, \frac{1}{4n}\right)$$

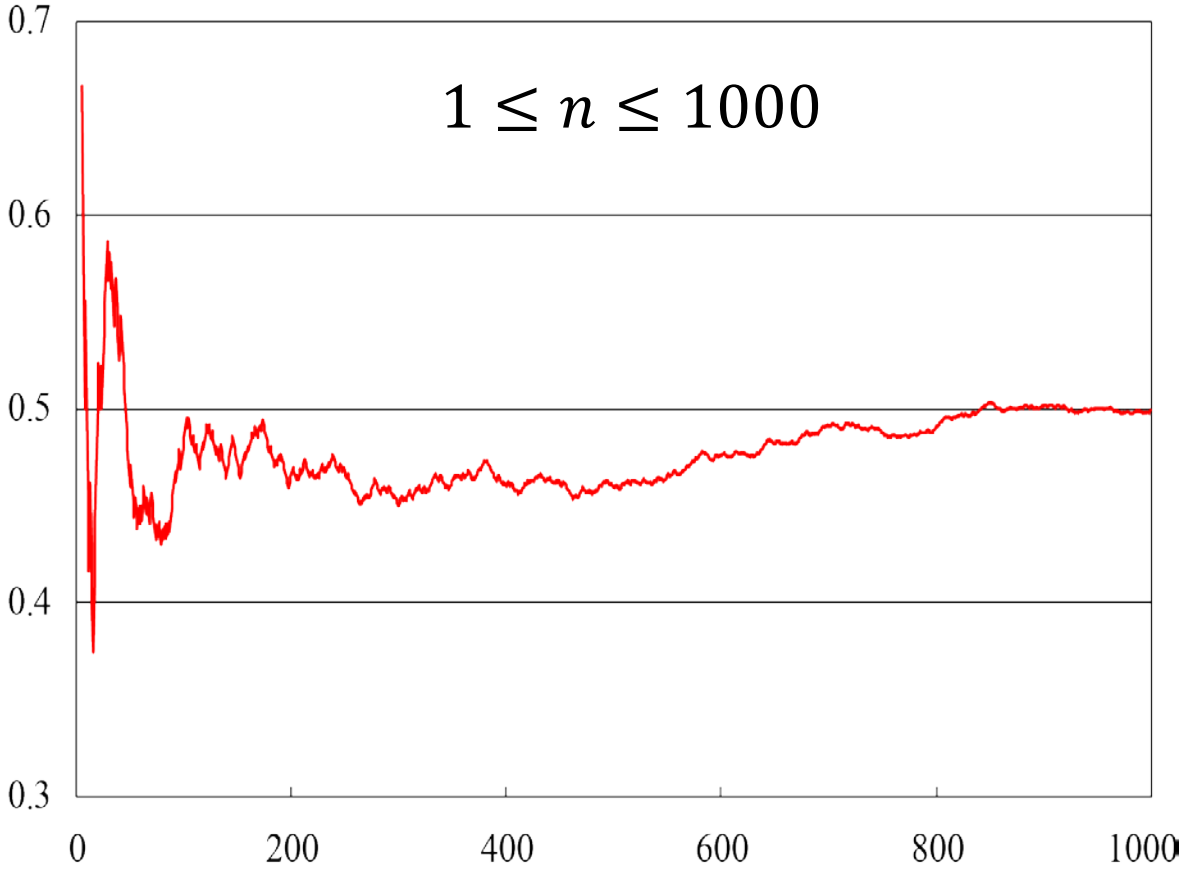
初めの n 回の内，表の回数の相対頻度

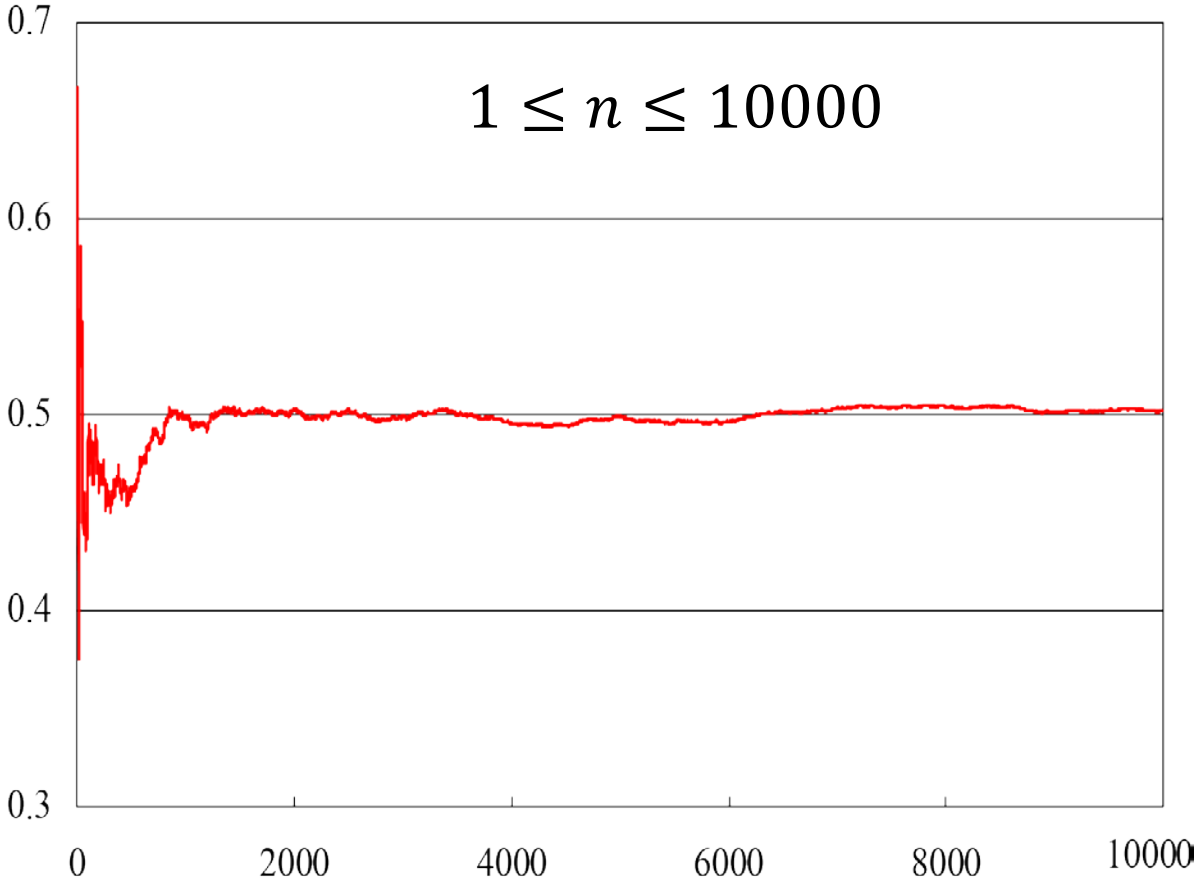
LLN

特に，確率 1 で

$$\lim_{n\to\infty} T_n = \frac{1}{2}$$

$$T_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$$





Lecture 6

標本抽出と正規分布

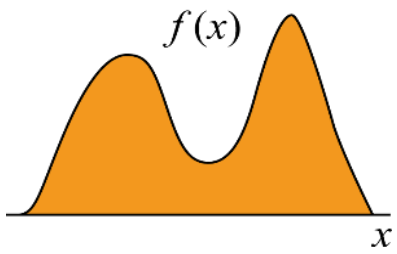
おわり

Lecture 7

点推定

推定量（estimator）

母集団



母集団分布

無作為
復元抽出

標本

$X_1, X_2, \ldots, X_n$

実際に知りたいのは，

母数 θ ＝母集団の統計量

平均値，分散，そのほか

θ を標本 $X_1, X_2, \ldots, X_n$ を用いて

何らかの合理的な計算で求める．

$$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$$

$\hat{\theta}$ を母数 θ の推定量という

推定量：確率変数
推定値：データから計算される数値

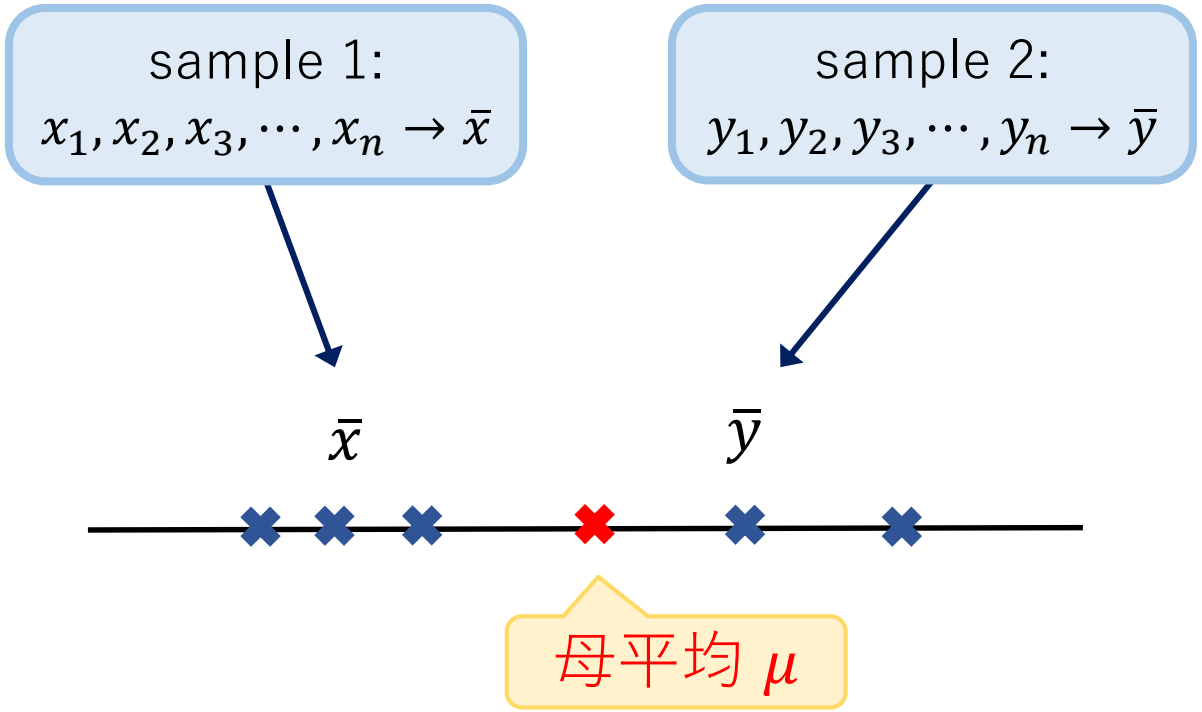
母平均の推定

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

標本
$X_1, X_2, \cdots, X_n$

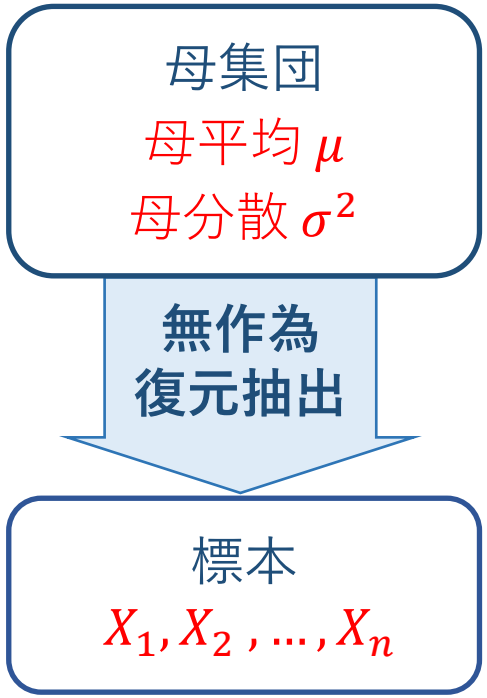
- 標本 $X_1, X_2, \cdots, X_n$ は母集団分布に従う独立同分布(iid)な確率変数列
- 標本平均 $\bar{X}$ は確率変数である



標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

母平均の推定



- 標本 $X_1, X_2, \cdots, X_n$ は母集団分布に従う独立同分布(iid)な確率変数列
- 標本平均 $\bar{X}$ は確率変数である

定理（再録）

母平均 μ，母分散 σ^2 の母集団から取り出された n 個の無作為標本の標本平均 $\bar{X}$ の平均値と分散は

$$E[\bar{X}] = \mu, \qquad V[\bar{X}] = \frac{\sigma^2}{n}$$

不偏性：$\bar{X}$ は μ の推定量として合理的である

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

不偏推定量

母集団
母数 θ

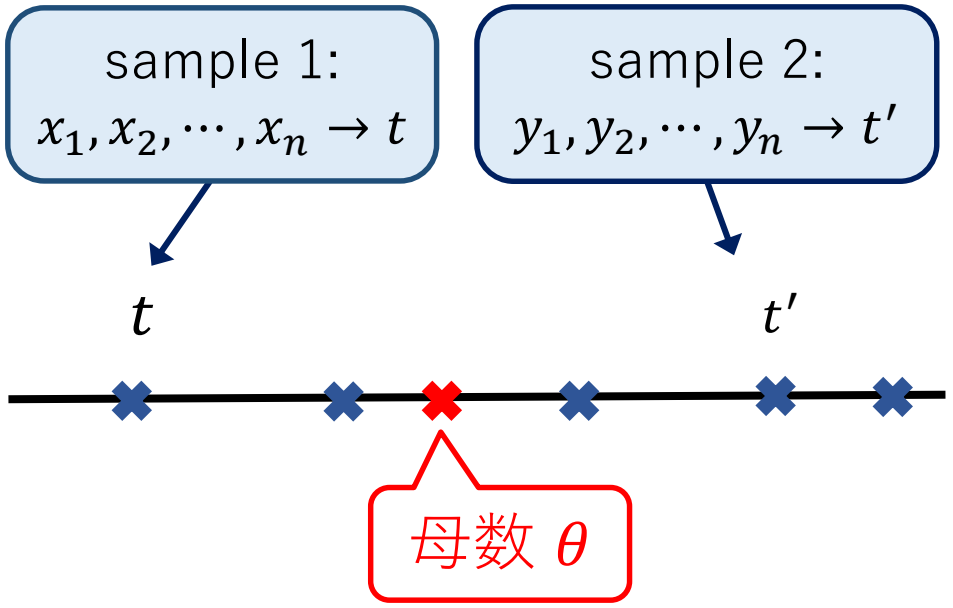
無作為
復元抽出

標本
$X_1, X_2, \ldots, X_n$

推定量

$$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$$
$$= T$$

➢ 推定量 $\hat{\theta} = T$ は確率変数である



定 義

推定量 T が $\boldsymbol{E[T] = \theta}$ を満たすとき，
その推定量を不偏推定量という．

例 題 7.1　母平均 μ，母分散 σ^2 の母集団から取り出された無作為標本を $X_1, X_2, \ldots, X_n$ とする.

（1）標本平均 $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ は母平均 μ の不偏推定量である.

（2）加重平均 $T = \sum_{k=1}^{n} c_k X_k$ も母平均 μ の不偏推定量である.

ただし，$c_k \geq 0$ かつ $\sum c_k = 1$

例 題 7.1　母平均 μ, 母分散 σ^2 の母集団から取り出された無作為標本を $X_1, X_2, \ldots, X_n$ とする.

（1）標本平均 $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ は母平均 μ の不偏推定量である.

（1）$E[\bar{X}] = \mu$ を示せばよい. 簡単である:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

例 題 7.1　母平均 μ，母分散 σ^2 の母集団から取り出された無作為標本を $X_1, X_2, \ldots, X_n$ とする.

（1）標本平均 $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ は母平均 μ の不偏推定量である.

（2）加重平均 $T = \sum_{k=1}^{n} c_k X_k$ も母平均 μ の不偏推定量である.

ただし，$c_k \geq 0$ かつ $\sum c_k = 1$　　※ 等加重 $c_k = \frac{1}{n}$ の場合が標本平均

（2）平均値の線形性を用いるだけ.

$$E[T] = E\left[\sum_{k=1}^{n} c_k X_k\right] = \sum_{k=1}^{n} c_k\, E[X_k] = \sum_{k=1}^{n} c_k\, \mu = \mu \sum_{k=1}^{n} c_k = \mu$$

例 題 7.2　$X_1, X_2, \ldots, X_n$ は独立で同一分布に従う観測で, その平均は $E[X_i] = \mu$, 分散は $V[X_i] = \sigma^2$ であるとする. 定数 $c_1, c_2, \cdots, c_n$ を係数とする観測の線形関数 $T = c_1X_1 + \cdots + c_nX_n$ を考える.

（1）T が μ の不偏推定量であるために, 定数 $c_1, \cdots, c_n$ の満たすべき条件を求めよ.

（2）線形不偏推定量の中で, 分散を最小にする定数 $c_1, \cdots, c_n$ を求めよ.

【演習 10分】

［ヒント］

(1) [平均値の線形性] 確率変数 X, Y と 定数 a, b に対して

$$E[aX + bY] = aE[X] + bE[Y]$$

(2) [分散の加法性] 独立な確率変数 X, Y と 定数 a, b に対して

$$V[aX + bY] = a^2\,V[X] + b^2\,V[Y]$$

例 題 7.2　$X_1, X_2, \ldots, X_n$ は独立で同一分布に従う観測で, その平均は $E[X_i] = \mu$, 分散は $V[X_i] = \sigma^2$ であるとする. 定数 $c_1, c_2, \cdots, c_n$ を係数とする観測の線形関数 $T = c_1X_1 + \cdots + c_nX_n$ を考える.

（1）T が μ の不偏推定量であるために, 定数 $c_1, \cdots, c_n$ の満たすべき条件を求めよ.

（2）線形不偏推定量の中で, 分散を最小にする定数 $c_1, \cdots, c_n$ を求めよ.

（1）　$E[T] = E[c_1X_1 + \cdots + c_nX_n]$

$= c_1E[X_1] + \cdots + c_nE[X_n]$

$= c_1\mu + \cdots + c_n\mu$

$= (c_1 + \cdots + c_n)\mu$

$E[T] = \mu$ となるのは

$c_1 + \cdots + c_n = 1$

のときである.

※ このとき T を加重平均という

例 題 7.2　$X_1, X_2, \ldots, X_n$ は独立で同一分布に従う観測で, その平均は $E[X_i] = \mu$, 分散は $V[X_i] = \sigma^2$ であるとする. 定数 $c_1, c_2, \cdots, c_n$ を係数とする観測の線形関数 $T = c_1X_1 + \cdots + c_nX_n$ を考える.

（1）T が μ の不偏推定量であるために, 定数 $c_1, \cdots, c_n$ の満たすべき条件を求めよ.

（2）線形不偏推定量の中で, 分散を最小にする定数 $c_1, \cdots, c_n$ を求めよ.

（2）　$V[T] = V[c_1X_1 + \cdots + c_nX_n]$

$= V[c_1X_1] + \cdots + V[c_nX_n]$

$= c_1^2V[X_1] + \cdots + c_n^2V[X_n]$

$= (c_1^2 + \cdots + c_n^2)\sigma^2$

※ $c_1 + \cdots + c_n = 1$ のもとで

$c_1^2 + \cdots + c_n^2$

の最小値を考える.

※　$c_1+\cdots+c_n=1$ のもとで $c_1^2+\cdots+c_n^2$ の最小値を考える.

$$c_1^2+\cdots+c_n^2=\left(c_1-\frac{1}{n}\right)^2+\cdots+\left(c_n-\frac{1}{n}\right)^2+2\frac{1}{n}(c_1+\cdots+c_n)-n\left(\frac{1}{n}\right)^2$$

$$=\left(c_1-\frac{1}{n}\right)^2+\cdots+\left(c_n-\frac{1}{n}\right)^2+\boxed{\frac{2}{n}-\frac{1}{n}}=\frac{1}{n}$$

したがって，$c_1^2+\cdots+c_n^2$ は

$$c_1=\cdots=c_n=\frac{1}{n}\quad(*)$$

算術平均
（普通の平均）

のとき最小値 $\frac{1}{n}$ をとる.

つまり，$T=c_1X_1+\cdots+c_nX_n$ の分散は（＊）のとき 最小になる.

不偏推定量の比較

母集団
母平均 μ
母分散 σ^2

→ 無作為復元抽出 →

標本
$X_1, X_2, \cdots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

加重平均

$$T = \sum_{k=1}^{n} c_k X_k$$

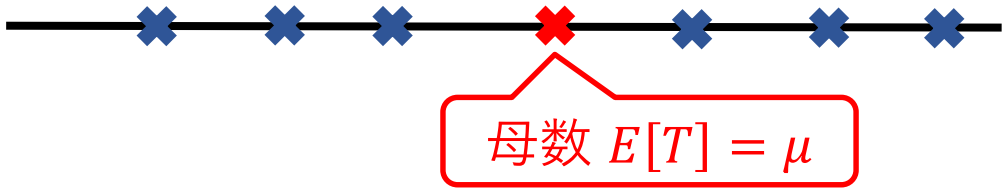
- 標本平均 $\bar{X}$，加重平均 T ともに母平均 μ の不偏推定量である．

$$E[\bar{X}] = E[T] = \mu$$

- $\bar{X}$ の方が揺らぎが小さい

$$V[\bar{X}] \leq V[T]$$

$\Rightarrow \bar{X}$ の方が推定量として優れている



不偏推定量の比較

母集団

母数 θ

無作為復元抽出

標本

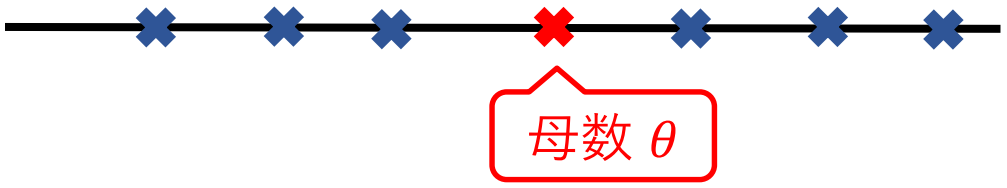
$X_1, X_2, \ldots, X_n$

2つの不偏推定量

$$T_1 = T_1(X_1, X_2, \ldots, X_n)$$

$$T_2 = T_2(X_1, X_2, \ldots, X_n)$$

- 不偏性から $E[T_1] = E[T_2] = \theta$



- 分散：　$V[T_1] = E[(T_1 - \theta)^2]$

$V[T_2] = E[(T_2 - \theta)^2]$

定 義

$V[T_1] \leq V[T_2]$のとき，T_1 は T_2 より有効（efficient）であるという．

標本分散と不偏分散

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

標本分散

$$S^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - \bar{X})^2$$

不偏分散

$$U^2 = \frac{1}{n-1}\sum_{k=1}^{n} (X_k - \bar{X})^2$$

定 理　母平均 μ，母分散 σ^2 の母集団から取り出された n 個の無作為標本を $X_1, X_2, \cdots, X_n$ とするとき，

(1) 標本分散　$S^2 = \dfrac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$　の平均値は　$E[S^2] = \dfrac{n-1}{n}\sigma^2$

(2) 不偏分散　$U^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$　の平均値は　$E[U^2] = \sigma^2$

つまり，不偏分散は母分散の不偏推定量である．

証明　計算でわかる（教科書を見よ）

例 題 7.3　（ドイツ戦車問題）

1番から順に通し番号がついている戦車が街を巡回している．目撃された番号が X_1, X_2, X_3 であったとき，それらの最大値を $M = \max\{X_1, X_2, X_3\}$ とおく．

このとき，$\widehat{N} = \frac{4}{3}M - 1$ は戦車の総数 N の不偏推定量である．

例 題 7.3　（ドイツ戦車問題）

1番から順に通し番号がついている戦車が街を巡回している．目撃された番号が X_1, X_2, X_3 であったとき，それらの最大値を $M = \max\{X_1, X_2, X_3\}$ とおく．

このとき，$\widehat{N} = \frac{4}{3}M - 1$ は戦車の総数 N の不偏推定量である．

まず，M の確率分布を求めて，$E[M]$ を計算しよう．

$$P(M = k) = \frac{\binom{k-1}{2}}{\binom{N}{3}} = \frac{3!\,(k-1)(k-2)}{2!\,N(N-1)(N-2)} = \frac{3(k-1)(k-2)}{N(N-1)(N-2)}$$

$$E[M] = \sum_{k=1}^{N} kP(M = k) = \frac{3}{N(N-1)(N-2)} \sum_{k=1}^{N} k(k-1)(k-2)$$

※　$E[M]=\dfrac{3}{N(N-1)(N-2)}\displaystyle\sum_{k=1}^{N}k(k-1)(k-2)$　の計算

【計算のトリック】

$$\sum_{k=1}^{N}k(k-1)(k-2)=\frac{1}{4}\sum_{k=1}^{N}\{(k+1)k(k-1)(k-2)-k(k-1)(k-2)(k-3)\}$$

$$=\frac{1}{4}(N+1)N(N-1)(N-2)$$

したがって,

$$E[M]=\frac{3}{N(N-1)(N-2)}\times\frac{1}{4}(N+1)N(N-1)(N-2)=\frac{3}{4}(N+1)$$

N について解くと,　$E\left[\dfrac{4}{3}M-1\right]=N.$　　よって, $\widehat{N}=\frac{4}{3}M-1$ は N の不偏推定量.

例 題 7.3（ドイツ戦車問題）

目撃された番号が $X_1, X_2, \cdots, X_n$ であったとき，それらの最大値を

$$M = \max\{X_1, X_2, \cdots, X_n\}$$

とすれば，$\widehat{N} = \left(1 + \frac{1}{n}\right) M - 1$ は戦車の総数 N の不偏推定量である．

$n = 3$ の場合に倣って確認できる

例 題 7.3（ドイツ戦車問題）

目撃された番号が $X_1, X_2, \cdots, X_n$ であったとき，それらの最大値を

$$M = \max\{X_1, X_2, \cdots, X_n\}$$

とすれば，$\hat{N} = \left(1 + \frac{1}{n}\right) M - 1$ は戦車の総数 N の不偏推定量である．

Month	Statistical estimate	Conventional estimate*)	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

*) 連合軍情報本部の伝統手法

Lecture 7

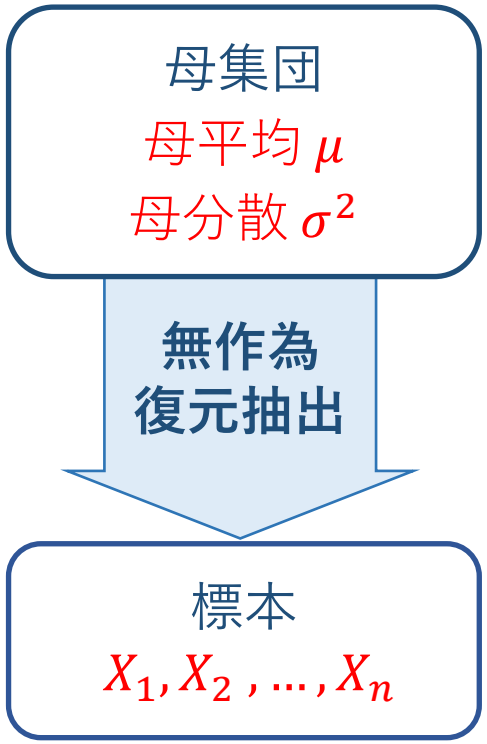
点推定

おわり

Lecture 8

区間推定

母平均の点推定（復習）



母平均 μ の点推定量として

標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

が最も基本的である

理由

（1）［不偏性］　$E[\bar{X}] = \mu$

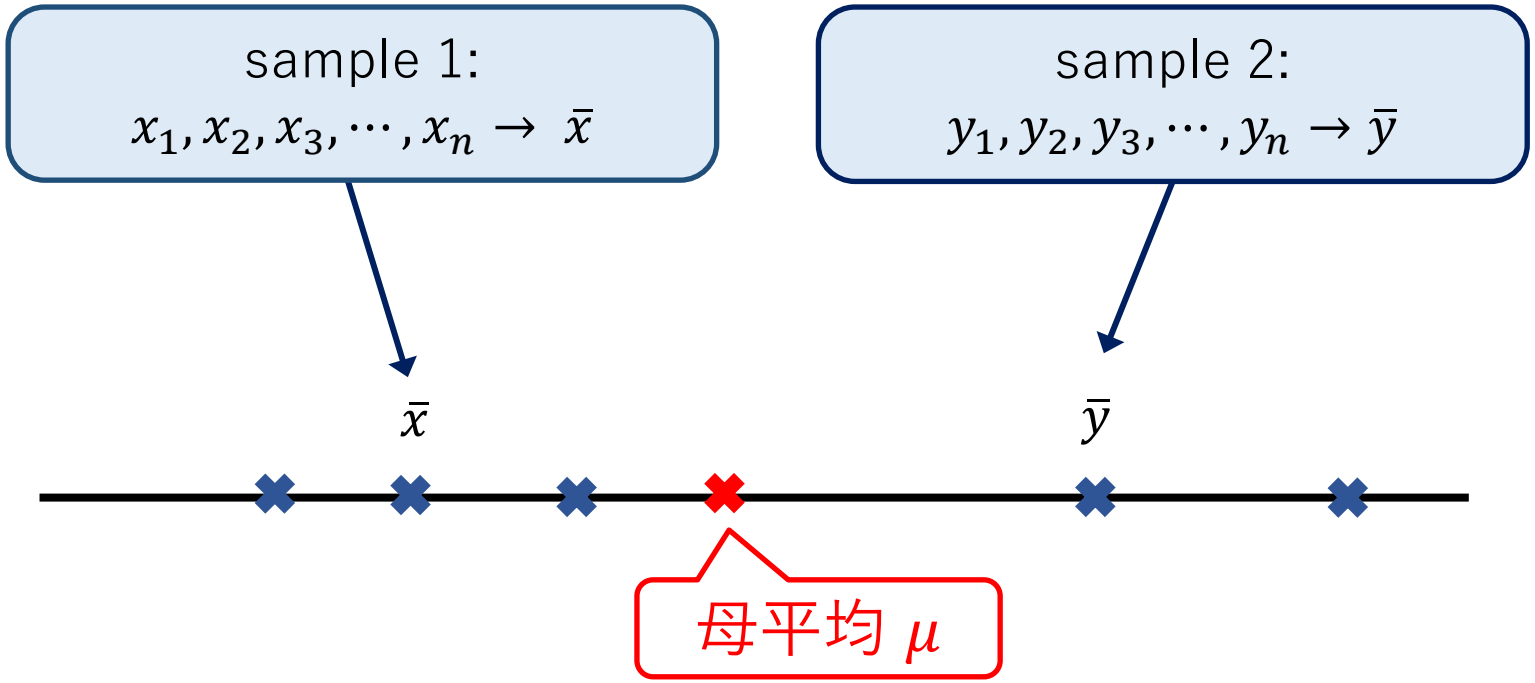
（2）［一致性］　$P\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1$

大数の法則

点推定の問題点

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

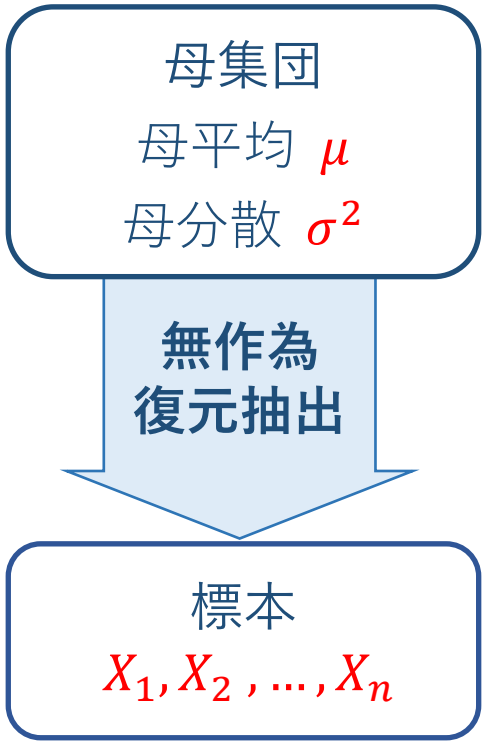


1回の標本調査で得られる標本平均 $\bar{x}$ が
母平均 μ に近いかどうか全く不明である.

→ 信頼性の問題

※ 得られた標本平均と母平均の差を確率的に評価

標本平均の分布（復習）



標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

基本：標本平均の分布は

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

（1）正規母集団なら厳密に成り立つ.
（2）一般の母集団なら，n が大きいとき近似的に成り立つ.

以降, 標準化して扱う

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

母平均の区間推定

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

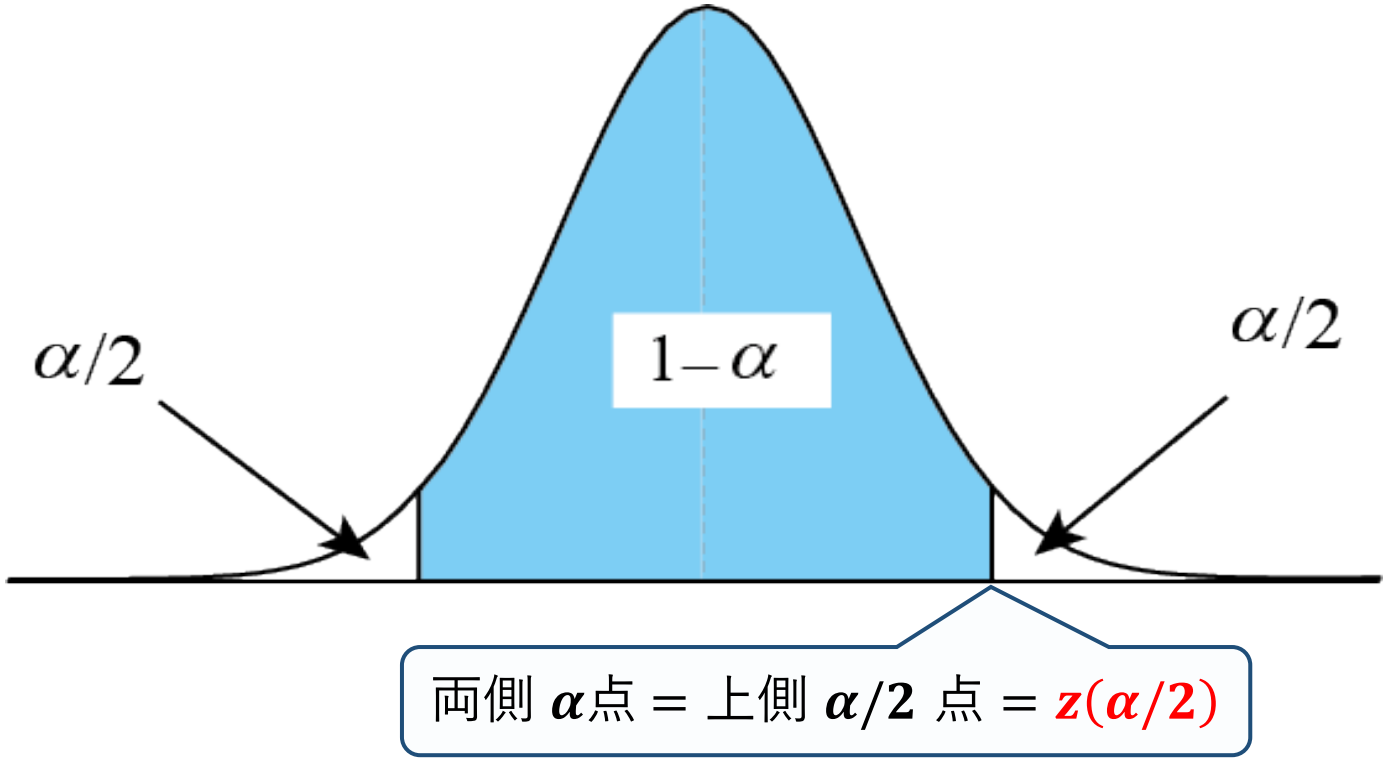
標本
$X_1, X_2, \ldots, X_n$

標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

標準化　$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$Z \sim N(0,1)$ の意味

$0 < \alpha < 1$

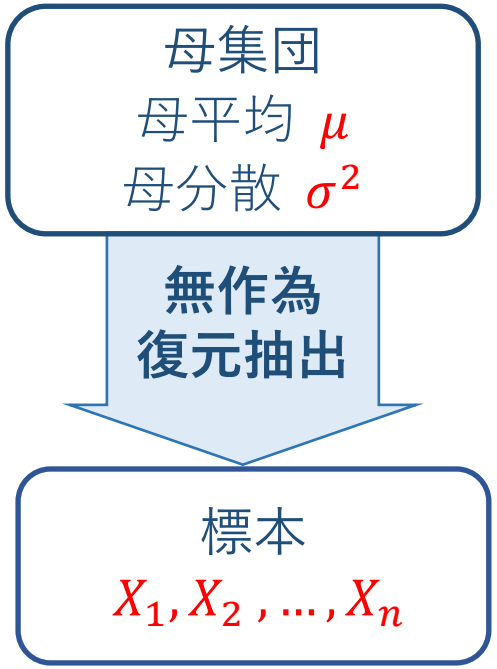


両側 $\boldsymbol{\alpha}$点 = 上側 $\boldsymbol{\alpha/2}$ 点 = $\boldsymbol{z(\alpha/2)}$

$$P\left(|Z| \leq z(\alpha/2)\right) = 1 - \alpha$$

※ 文献によっては
$\boldsymbol{z(\alpha)}$ =**両側 $\boldsymbol{\alpha}$点**

母平均の区間推定



標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

標準化　$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$P(|Z| \leq z(\alpha/2)) = 1 - \alpha$ を変形する

$$|Z| \leq z(\alpha/2)$$

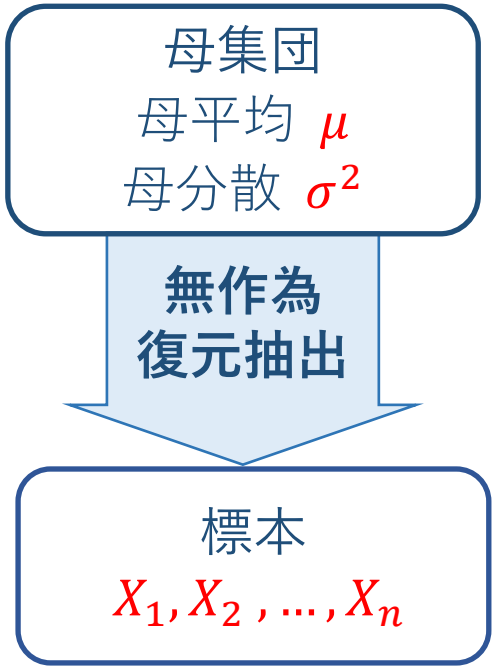
$$\Leftrightarrow -z(\alpha/2) \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z(\alpha/2)$$

$$\Leftrightarrow -z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

母平均 μ は区間 $\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ の中に確率 $1-\alpha$ で見つかる

母平均の区間推定



標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

標準化　$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

母平均 μ は区間 $\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ の中に確率 $1-\alpha$ で見つかる

$$P\left(\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

定義：信頼区間

母平均 μ に対する信頼係数 $1-\alpha$ の信頼区間

$$\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

例 題 8.1

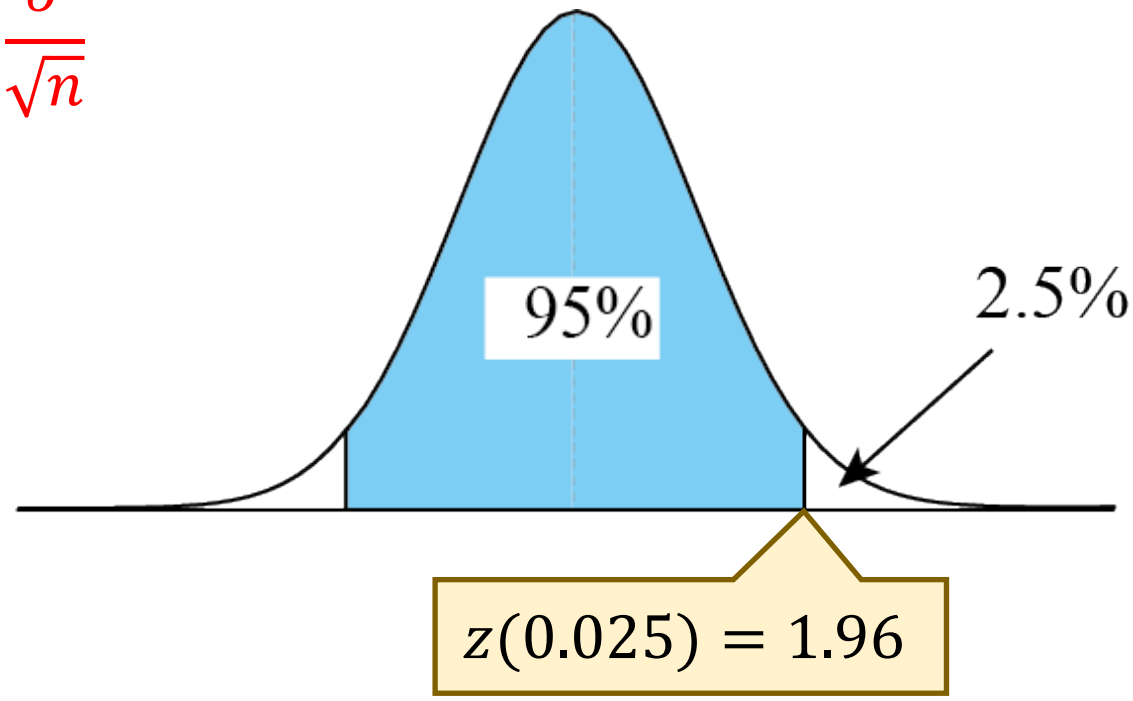
母分散が $\sigma^2 = 5^2$ である正規母集団から 10 個の無作為標本を抽出して標本平均 $\bar{x} = 12.8$ を得た. 母平均の 95% 信頼区間を求めよ.

例 題 8.1

母分散が $\sigma^2 = 5^2$ である正規母集団から 10 個の無作為標本を抽出して標本平均 $\bar{x} = 12.8$ を得た. 母平均の 95% 信頼区間を求めよ.

信頼係数 $1-\alpha$ の信頼区間は　$\bar{X} \pm z(\alpha/2)\dfrac{\sigma}{\sqrt{n}}$

信頼係数 95% ⇨ $\alpha = 0.05$

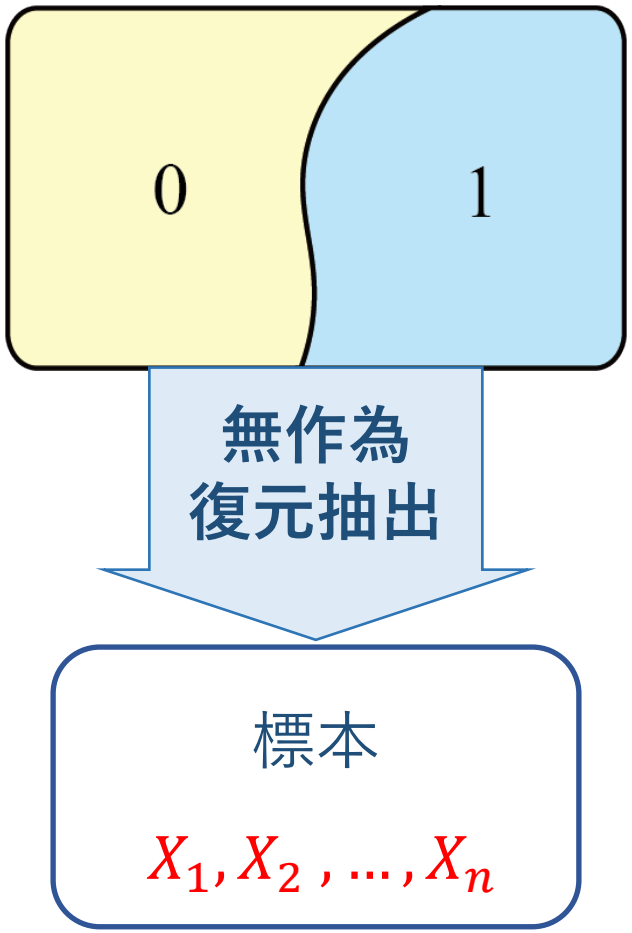


求めるべき信頼区間は

$$12.8 \pm 1.96 \times \frac{5}{\sqrt{10}} = 12.8 \pm 3.1$$

上側 2.5% 点 = 両側 5% 点

二項母集団の母比率の区間推定



- 二項母集団とは，0 と 1 からなる母集団
- 1 の割合を母比率 p という

母集団分布について

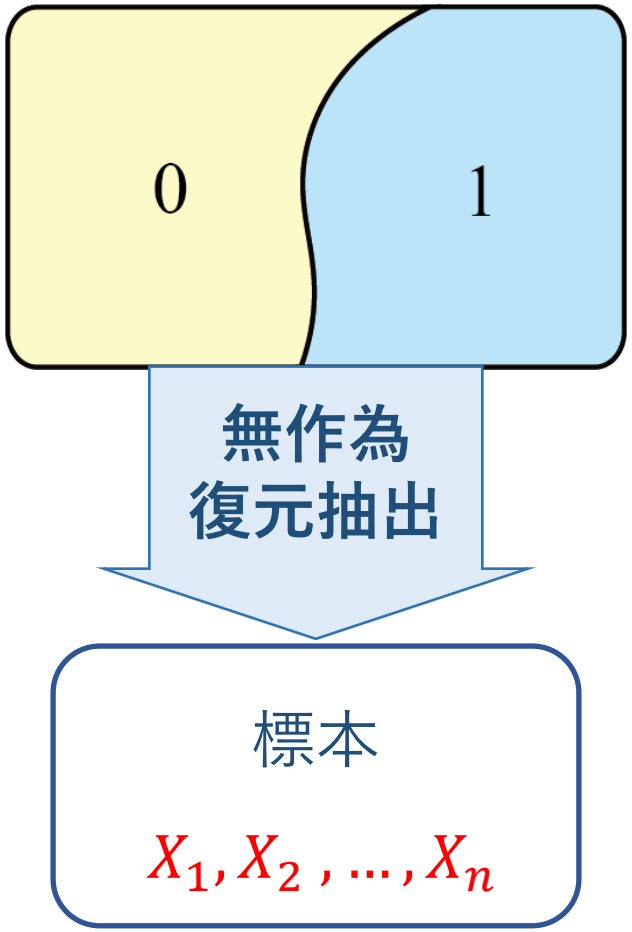
母平均　$\mu = p$

母分散　$\sigma^2 = p(1-p)$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \hat{p}$$　　標本比率という

二項母集団の母比率の区間推定



母平均 ＝ 母比率　$\mu = p$

母分散　$\sigma^2 = p(1-p)$

標本平均 ＝ 標本比率　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \hat{p}$

【一般論】　$\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$

【大数の法則】$\sigma^2 = p(1-p) \approx \hat{p}(1-\hat{p})$

母比率 p に対する信頼係数 $1-\alpha$ の信頼区間

$$\hat{p} \pm z(\alpha/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例 題 8.2　(視聴率調査)

ビデオリサーチ社：関東地区（ドラマ）

順位	番組タイトル	放送局	視聴率
1	連続テレビ小説・なつぞら （4/10）	NHK総合	23.1 %
2	特捜9 （4/10）	テレビ朝日	15.2 %
2	木曜ドラマ・緊急取調室 （4/11）	テレビ朝日	15.2 %
4	ラジエーションハウス・放射線科の診断レポート （4/8）	フジテレビ	12.7 %
5	日曜プライム「ドラマスペシャル アガサ・クリスティ　予告殺人」（4/14）	テレビ朝日	11.5 %
6	金曜ドラマ・インハンド （4/12）	TBS	11.3 %
7	白衣の戦士！ （4/10）	日本テレビ	10.3 %
8	緊急取調室（4/11）	テレビ朝日	9.9 %
9	いだてん～東京オリムピック噺～ （4/14）	NHK総合	9.6 %
10	特捜9[再] （4/10）	テレビ朝日	9.5 %

※ 関東地区の視聴率調査は 600 世帯を対象にしている.

順位	番組タイトル	放送局	視聴率
1	連続テレビ小説・なつぞら （4/10）	NHK総合	23.1 %
2	特捜9 （4/10）	テレビ朝日	15.2 %
・・・・・			
9	いだてん～東京オリムピック噺～ （4/14）	NHK総合	9.6 %
10	特捜9[再] （4/10）	テレビ朝日	9.5 %

95%信頼区間

$$\hat{p} \pm z(\alpha/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例

$$0.231 \pm 1.96 \times \sqrt{\frac{0.231(1-0.231)}{600}} = 0.231 \pm 0.034$$

$$0.095 \pm 1.96 \times \sqrt{\frac{0.095(1-0.095)}{600}} = 0.095 \pm 0.023$$

ここまでのまとめ

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

標本
$X_1, X_2, \ldots, X_n$

標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

標準化　$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

信頼区間

母平均 μ に対する信頼係数 $1-\alpha$ の信頼区間

$$\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

σ^2 が既知でないと使えない

二項母集団 $\sigma^2 = p(1-p) \approx \hat{p}(1-\hat{p})$

※ σ^2 が未知のとき, σ^2 も標本から推定する

⇨ 不偏分散を用いる　$U^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$

定 理 （t-変換） 重要

$X_1, X_2, \cdots, X_n$：正規母集団 $N(\mu, \sigma^2)$ から取り出された n 個の無作為標本

標本平均： $\bar{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$　　不偏分散： $U^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$

このとき， $T = \dfrac{\bar{X} - \mu}{U/\sqrt{n}}$ は自由度 $n-1$ の t-分布 t_{n-1} に従う．

証明の流れ

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \qquad Y = \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - \bar{X})^2 = \frac{(n-1)U^2}{\sigma^2} \sim \chi^2_{n-1}$$

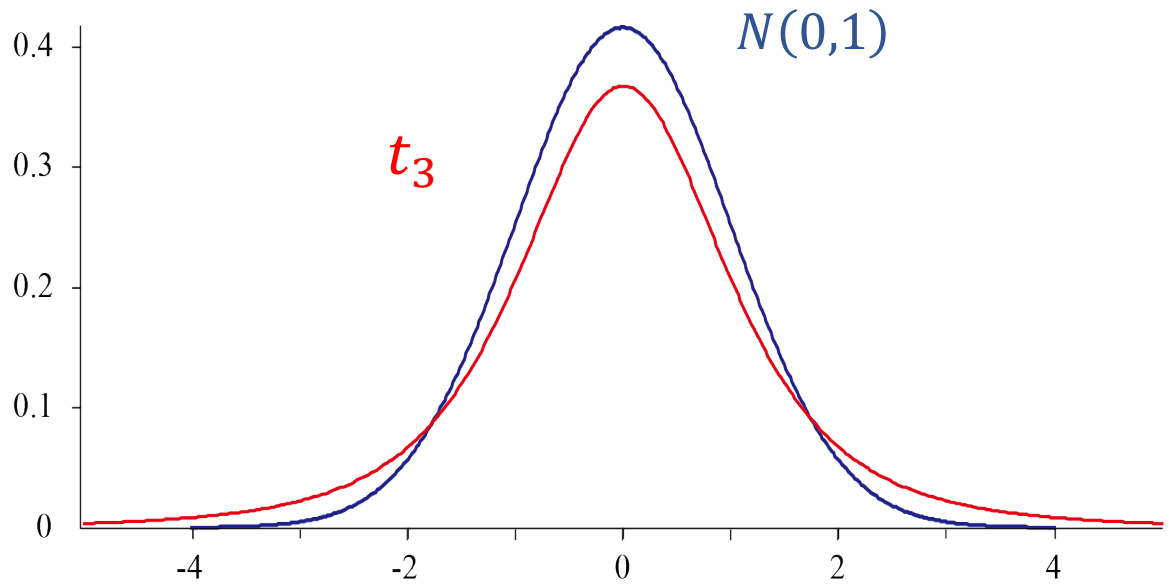
Z と Y は独立なので, t-分布 t_{n-1} の定義から　$\dfrac{Z}{\sqrt{\dfrac{Y}{n-1}}} \sim t_{n-1}$

$$\frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{U} = \frac{\bar{X} - \mu}{U/\sqrt{n}}$$　なので　$\dfrac{\bar{X} - \mu}{U/\sqrt{n}} \sim t_{n-1}$

t 分布

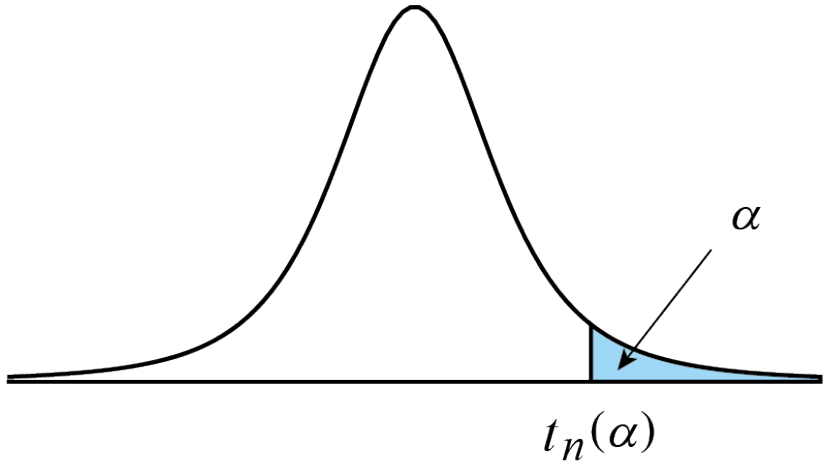
自由度 n の t_n - 分布の密度関数

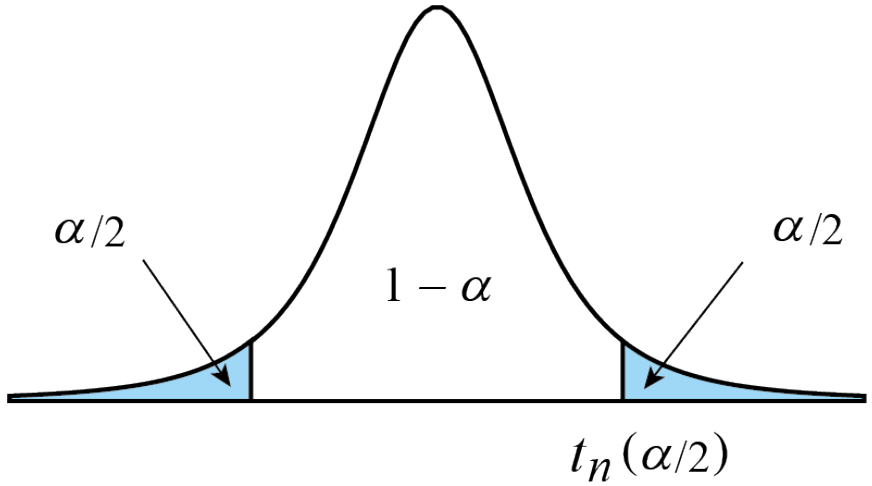
$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$



- $N(0,1)$ に比べて，すそ野が厚い．
- 自由度 $n \to \infty$ で t_n-分布は標準正規分布 $N(0,1)$ に一致する．
- 実用上，$n \geq 30$ で標準正規分布 $N(0,1)$ で代用．
- 確率変数 X, Y が独立であり，$X \sim N(0,1)$，$Y \sim \chi_n^2$ であれば，$T = \dfrac{X}{\sqrt{Y/n}} \sim t_n$

上側 α 点　$t_n(\alpha)$



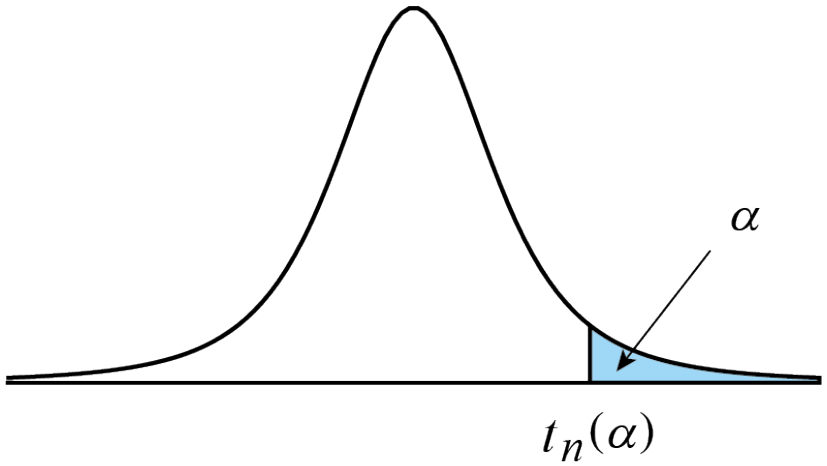


※ 文献によっては　$t_n(\alpha)$ = **両側 α点**

$$P\left(|T| \geq t_n(\alpha/2)\right) = \alpha$$

$$\Leftrightarrow P\left(|T| \leq t_n(\alpha/2)\right) = 1 - \alpha$$

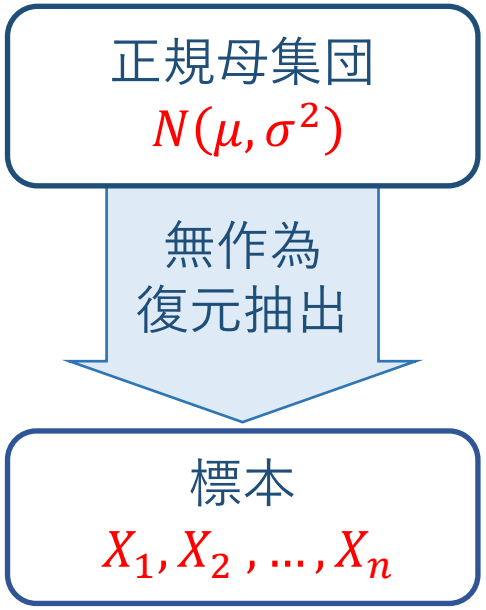
t 分布の上側 α 点



例

$t_7(0.05) = 1.895$

n \ α	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.001
1	1.00000	1.37638	1.96261	3.07768	6.31375	12.70620	31.82052	63.65674	318.30884
2	0.81650	1.06066	1.38621	1.88562	2.91999	4.30265	6.96456	9.92484	22.32712
3	0.76489	0.97847	1.24978	1.63774	2.35336	3.18245	4.54070	5.84091	10.21453
4	0.74070	0.94096	1.18957	1.53321	2.13185	2.77645	3.74695	4.60409	7.17318
5	0.72669	0.91954	1.15577	1.47588	2.01505	2.57058	3.36493	4.03214	5.89343
6	0.71756	0.90570	1.13416	1.43976	1.94318	2.44691	3.14267	3.70743	5.20763
7	0.71114	0.89603	1.11916	1.41492	1.89458	2.36462	2.99795	3.49948	4.78529
8	0.70639	0.88889	1.10815	1.39682	1.85955	2.30600	2.89646	3.35539	4.50079
9	0.70272	0.88340	1.09972	1.38303	1.83311	2.26216	2.82144	3.24984	4.29681
10	0.69981	0.87906	1.09306	1.37218	1.81246	2.22814	2.76377	3.16927	4.14370
11	0.69745	0.87553	1.08767	1.36343	1.79588	2.20099	2.71808	3.10581	4.02470
12	0.69548	0.87261	1.08321	1.35622	1.78229	2.17881	2.68100	3.05454	3.92963
13	0.69383	0.87015	1.07947	1.35017	1.77093	2.16037	2.65031	3.01228	3.85198
14	0.69242	0.86805	1.07628	1.34503	1.76131	2.14479	2.62449	2.97684	3.78739
15	0.69120	0.86624	1.07353	1.34061	1.75305	2.13145	2.60248	2.94671	3.73283
16	0.69013	0.86467	1.07114	1.33676	1.74588	2.11991	2.58349	2.92078	3.68615
17	0.68920	0.86328	1.06903	1.33338	1.73961	2.10982	2.56693	2.89823	3.64577
18	0.68836	0.86205	1.06717	1.33039	1.73406	2.10092	2.55238	2.87844	3.61048
19	0.68762	0.86095	1.06551	1.32773	1.72913	2.09302	2.53948	2.86093	3.57940
20	0.68695	0.85996	1.06402	1.32534	1.72472	2.08596	2.52798	2.84534	3.55181
21	0.68635	0.85907	1.06267	1.32319	1.72074	2.07961	2.51765	2.83136	3.52715
22	0.68581	0.85827	1.06145	1.32124	1.71714	2.07387	2.50832	2.81876	3.50499
23	0.68531	0.85753	1.06034	1.31946	1.71387	2.06866	2.49987	2.80734	3.48496
24	0.68485	0.85686	1.05932	1.31784	1.71088	2.06390	2.49216	2.79694	3.46678
25	0.68443	0.85624	1.05838	1.31635	1.70814	2.05954	2.48511	2.78744	3.45019
26	0.68404	0.85567	1.05752	1.31497	1.70562	2.05553	2.47863	2.77871	3.43500
27	0.68368	0.85514	1.05673	1.31370	1.70329	2.05183	2.47266	2.77068	3.42103
28	0.68335	0.85465	1.05599	1.31253	1.70113	2.04841	2.46714	2.76326	3.40816
29	0.68304	0.85419	1.05530	1.31143	1.69913	2.04523	2.46202	2.75639	3.39624
30	0.68276	0.85377	1.05466	1.31042	1.69726	2.04227	2.45726	2.75000	3.38518
35	0.68156	0.85201	1.05202	1.30621	1.68957	2.03011	2.43772	2.72381	3.34005
40	0.68067	0.85070	1.05005	1.30308	1.68385	2.02108	2.42326	2.70446	3.30688
45	0.67998	0.84968	1.04852	1.30065	1.67943	2.01410	2.41212	2.68959	3.28148
50	0.67943	0.84887	1.04729	1.29871	1.67591	2.00856	2.40327	2.67779	3.26141
∞	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32635	2.57583	3.09023



標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

不偏分散

$$U^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

- 母分散 σ^2 が既知なら

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

が使えるが，未知なら σ^2 の代わりに U^2 で置き換えて，

$$T = \frac{\bar{X} - \mu}{U/\sqrt{n}} \sim t_{n-1}$$

自由度 $n-1$ の t-分布

正規母集団
$N(\mu, \sigma^2)$

無作為
復元抽出

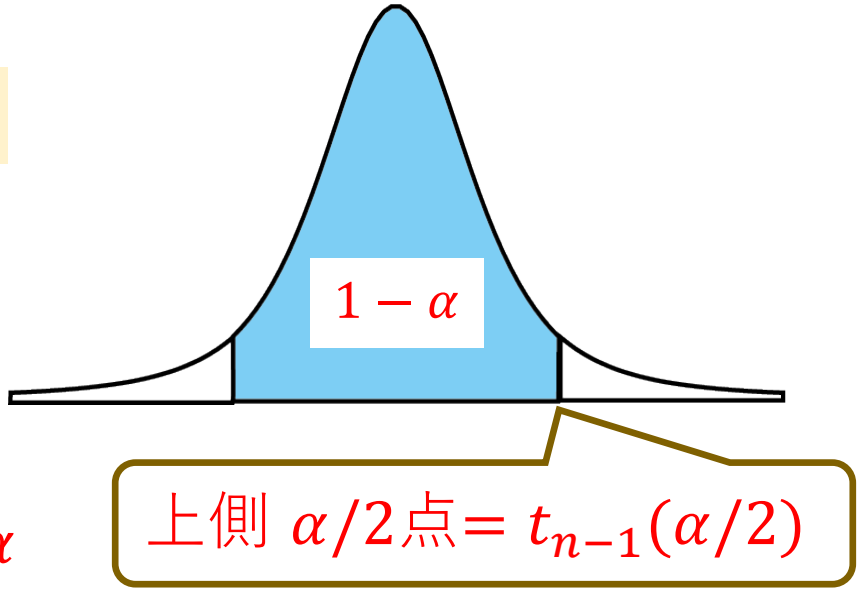
標本
$X_1, X_2, \ldots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$$T = \frac{\bar{X} - \mu}{U/\sqrt{n}} \sim t_{n-1}$$

自由度 $n-1$ の t-分布



$$P\left(|T| \le t_{n-1}(\alpha/2)\right) = 1-\alpha$$

$$\bar{X} - t_{n-1}(\alpha/2)\frac{U}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1}(\alpha/2)\frac{U}{\sqrt{n}}$$

信頼区間

母平均 μ に対する信頼係数 $1-\alpha$ の信頼区間

$$\bar{X} \pm t_{n-1}(\alpha/2)\frac{U}{\sqrt{n}}$$

まとめ

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

標本
$X_1, X_2, \ldots, X_n$

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

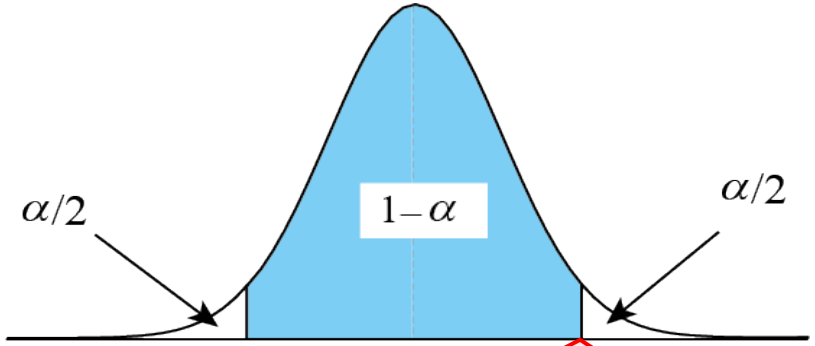
$$U^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

母平均 μ に対する 信頼係数 $1-\alpha$ の信頼区間	
σ^2が既知	σ^2が未知
正規母集団または, 一般の母集団で n が大きい (二項母集団も可)	正規母集団
$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\dfrac{\bar{X}-\mu}{U/\sqrt{n}} \sim t_{n-1}$
$\bar{X} \pm z(\alpha/2)\dfrac{\sigma}{\sqrt{n}}$	$\bar{X} \pm t_{n-1}(\alpha/2)\dfrac{U}{\sqrt{n}}$

信頼係数と信頼区間の幅

$$\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

$$\bar{X} \pm t_{n-1}(\alpha/2)\frac{U}{\sqrt{n}}$$



上側 $\boldsymbol{\alpha/2}$点
$= \boldsymbol{z(\alpha/2)}$ または $\boldsymbol{t_{n-1}(\alpha/2)}$

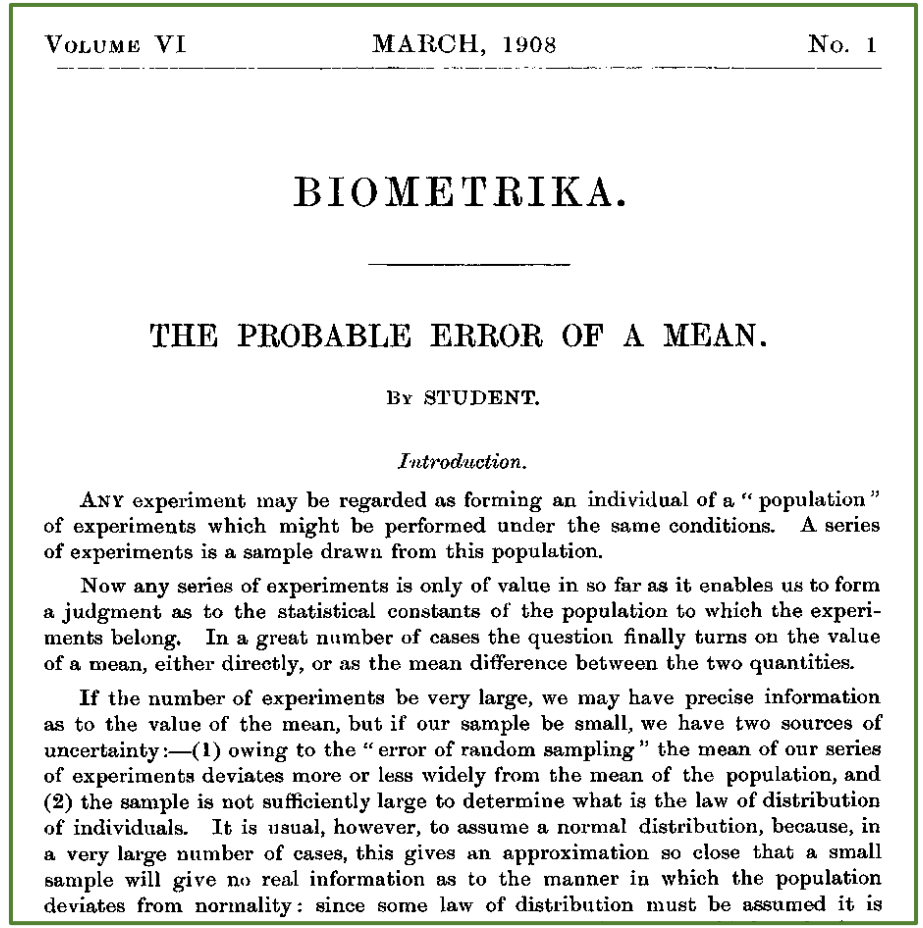
信頼係数 $1-\alpha$	0	小		大	1
α	1	大		小	0
信頼区間の幅	0	小		大	無限大 ∞

点推定

何も言わない

William Sealy Gosset (1876-1937)

1899年 ギネスビール社
　　ダブリン醸造所に就職
1906-07年 カール・ピアソンに学ぶ
1908年 有名な論文
　　スチューデントの t 分布
1935年 新設のロンドン醸造所に転勤



VOLUME VI　MARCH, 1908　No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is

- 小標本の問題を扱う
　（フィッシャーが高く評価）
- 当時の主流「大標本主義」
　（ゴルトン、ピアソンら）

例 題 8.3

正規母集団から, 無作為標本を抽出して次のような 24 個のデータを得た. 母平均の 95% 信頼区間を求めよ.

35.9	43.9	51.2	35.3	36.7	49.4	39.5	59.6
43.8	32.9	36.0	43.0	41.9	44.6	47.2	56.2
45.6	47.7	38.1	51.8	42.3	46.6	35.5	32.4

例 題 8.3

正規母集団から, 無作為標本を抽出して次のような 24 個のデータを得た. 母平均の 95% 信頼区間を求めよ.

35.9	43.9	51.2	35.3	36.7	49.4	39.5	59.6
43.8	32.9	36.0	43.0	41.9	44.6	47.2	56.2
45.6	47.7	38.1	51.8	42.3	46.6	35.5	32.4

信頼係数 $1-\alpha$ の信頼区間　$\bar{X} \pm t_{n-1}(\alpha/2)\dfrac{U}{\sqrt{n}}$

$$\bar{x} = \frac{1}{24}\sum x_i = 43.19 \qquad u^2 = \frac{1}{23}\sum (x_i - \bar{x})^2 = 54.231 = 7.36^2$$

95% 信頼区間　$43.19 \pm t_{23}(0.025) \times \dfrac{7.36}{\sqrt{24}} = 43.19 \pm 2.069 \times \dfrac{7.36}{\sqrt{24}} = 43.19 \pm 4.22$

例 題 8.4

ある製品の検査の所要時間は正規分布に従うといわれている. 大きさ 10 の無作為標本について, 次のデータを得た. 母平均の 95 % 信頼区間と 90% 信頼区間を求めよ.

12.4　13.5　12.7　14.1　13.8　14.1　12.0　12.8　13.1　15.4

【演習10分】

例 題 8.4　ある製品の検査の所要時間は正規分布に従うといわれている. 大きさ 10 の無作為標本について, 次のデータを得た. 母平均の 95 % 信頼区間と 90% 信頼区間を求めよ.

12.4　13.5　12.7　14.1　13.8　14.1　12.0　12.8　13.1　15.4

$$\bar{x} = \frac{1}{10}\sum x_i = 13.39 \qquad \sum (x_i - \bar{x})^2 = 9.049 \qquad u^2 = \frac{1}{9}\sum (x_i - \bar{x})^2 = 1.0054 = 1.0027^2$$

信頼係数 $1-\alpha$ の信頼区間は, $\bar{X} \pm t_{n-1}(\alpha/2)\,\frac{U}{\sqrt{n}}$

95 % 信頼区間

$$\bar{x} \pm t_9(0.025) \times \frac{1.0027}{\sqrt{10}} = 13.39 \pm 2.262 \times 0.317 = 13.39 \pm 0.717$$

90 % 信頼区間

$$\bar{x} \pm t_9(0.05) \times \frac{1.0027}{\sqrt{10}} = 13.39 \pm 1.833 \times 0.317 = 13.39 \pm 0.581$$

例 題 8.5　世論調査により, ある候補者の支持率を信頼度 95% で推定したいとき, 信頼区間の幅が 0.05 以下になるようにするには標本数をいくらとらなければいけないか.

【演習10分】

例 題 8.5　世論調査により, ある候補者の支持率を信頼度 95% で推定したいとき, 信頼区間の幅が 0.05 以下になるようにするには標本数をいくらとらなければいけないか.

信頼係数 $1-\alpha$ の信頼区間　$\hat{p} \pm z(\alpha/2)\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

95% 信頼区間のためには $z(0.025) = 1.96$

信頼区間の幅　$2z(\alpha/2)\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = 2 \times 1.96\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \leq 0.05$

$0 \leq \hat{p} \leq 1$ なので, $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$

$2 \times 1.96\sqrt{\frac{1/4}{n}} \leq 0.05$ を解いて　$\sqrt{n} \geq \dfrac{1.96}{0.05}$　よって, $n \geq 1537$

Lecture 8

区間推定

おわり

Lecture 9

母平均の検定

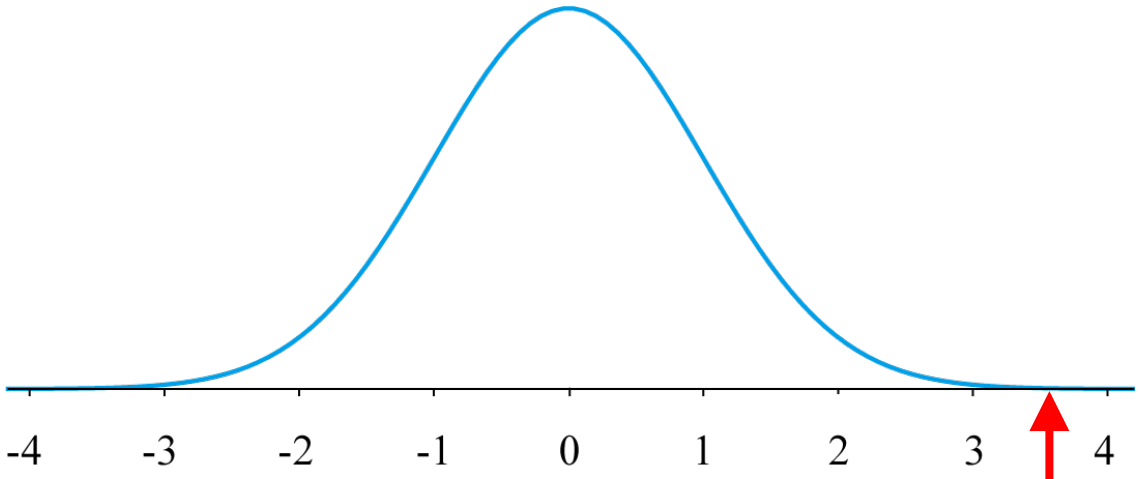
典型例　コインを 100 回投げて, 表が 67 回出た. コインは公平といえるか?

X ：100回投げて表の出る回数

$$X \sim B\left(100, \frac{1}{2}\right) \approx N(50, 5^2)$$

正規分布近似が便利

$$Z = \frac{X - 50}{5} \sim N(0,1)$$



実現値　$z = \dfrac{67 - 50}{5} = 3.4$

判断：かなり稀？

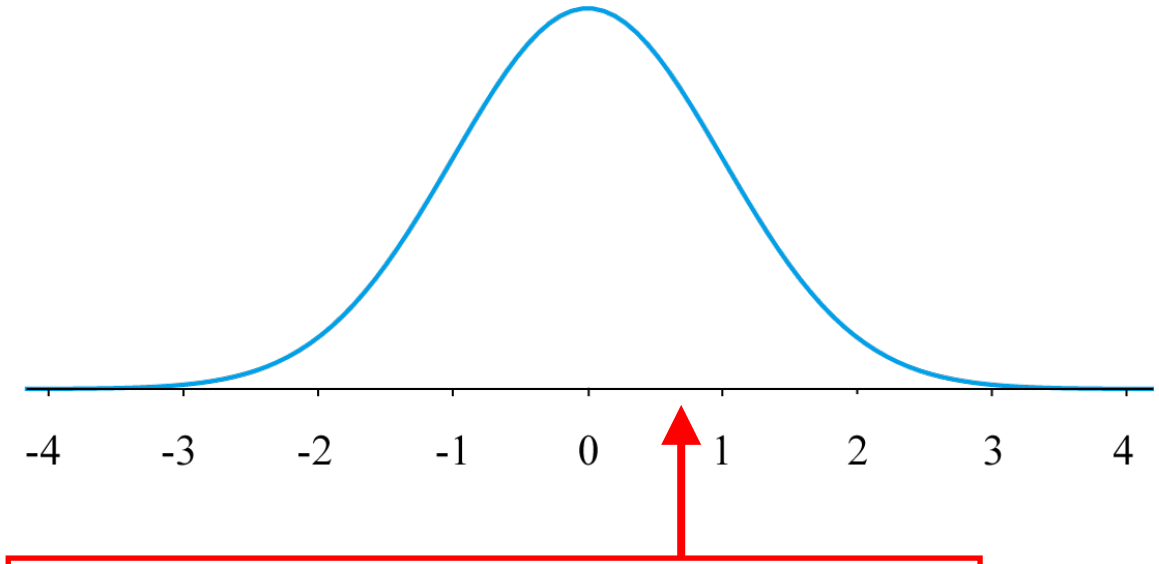
典型例　コインを 100 回投げて, 表が 54 回出た. コインは公平といえるか?

X ：100回投げて表の出る回数

$$X \sim B\left(100, \frac{1}{2}\right) \approx N(50, 5^2)$$

正規分布近似が便利

$$Z = \frac{X - 50}{5} \sim N(0,1)$$

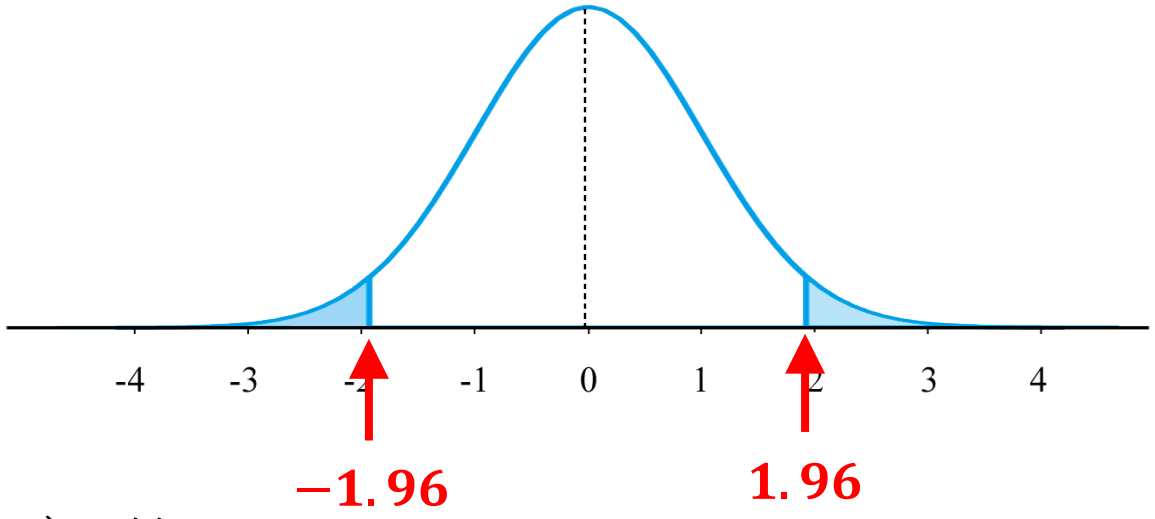


実現値　$z = \dfrac{54 - 50}{5} = 0.8$

判断：ふつうに起こる？

仮説検定の考え方

コインを 100 回投げて,



- 表が 54 回出た.

$$z = \frac{54 - 50}{5} = 0.8$$

⇒ 偶然の揺らぎの範囲

- 表が 67 回出た.

$$z = \frac{67 - 50}{5} = 3.4$$

⇒ 稀なことが起こった

合理的、客観的な判断のためには、どのくらい小さな確率を稀とするかを決める

たとえば、$\alpha = 0.05$

仮説検定の手順

(1) 母数に関する**帰無仮説**と**対立仮説**を決める.

H_0　　H_1

(2) 関連する確率変数 T **(検定統計量)**を選び, H_0 の下で, この確率変数の分布を調べる

(3) **有意水準** $0 < \alpha < 1$ と**棄却域** W を決める.

(4) 標本から T の**実現値** t を計算する.

- $t \in W$ ⇒実現値は有意水準 α で**有意**である
  ⇒ H_0 を**棄却する**⇒ H_1 を採択する.
- $t \notin W$ ⇒実現値は有意水準 α で有意でない
  ⇒ H_0 を**棄却できない**
  (⇒ H_0 を採択する)

例　コインを 100 回投げて, 表が 67 回出た. コインは公平といえるか?

p ：表の出る確率

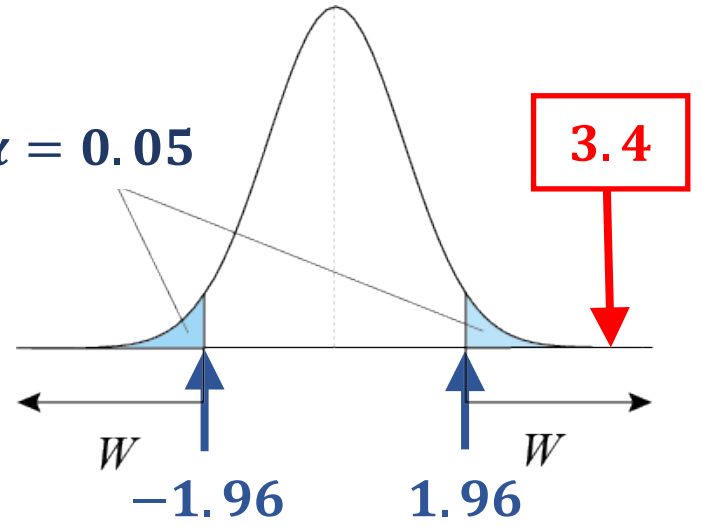
$H_0: p = \frac{1}{2}$　　$H_1: p \neq \frac{1}{2}$　　$\alpha = 0.05$

X ：表の回数 $\sim B(100, 0.5) \approx N(50, 5^2)$

$$Z = \frac{X - 50}{5} \sim N(0,1)$$

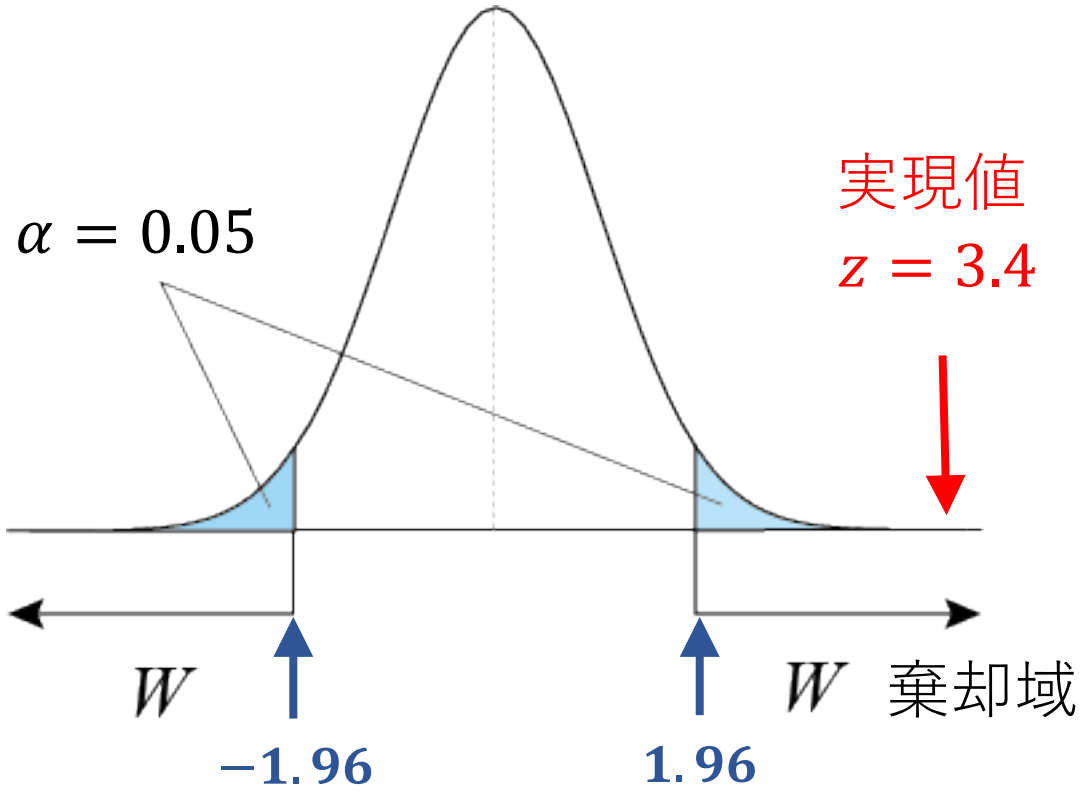
実現値

$z = 3.4$



結論　H_0 を棄却する

有意水準の意味

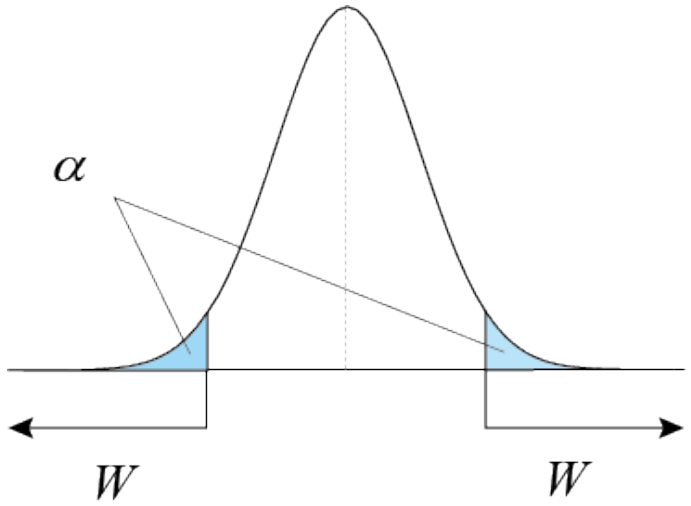


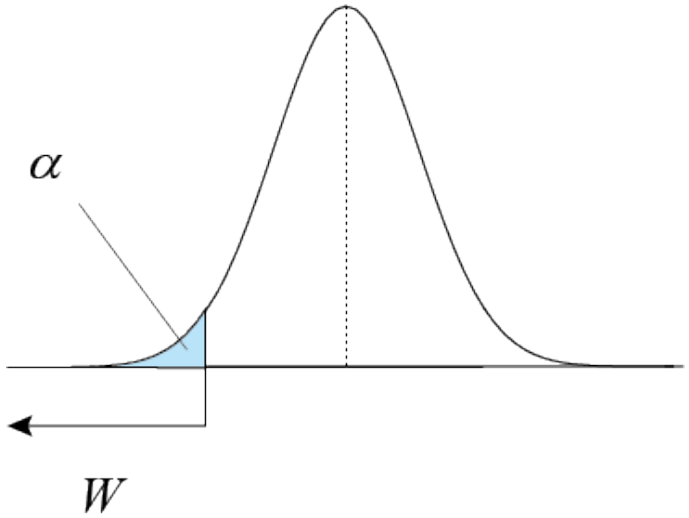
結論：H_0 を棄却する

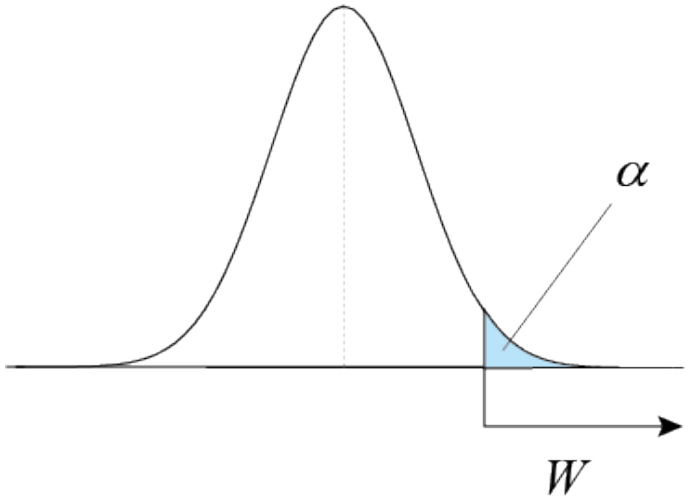
有意水準 α は，帰無仮説 H_0 が正しいのに，帰無仮説を棄却してしまって，検定の結論を間違える誤り確率

【注目】有意水準 α は文脈に応じて自分で設定する

棄却域の設定：両側検定と片側検定

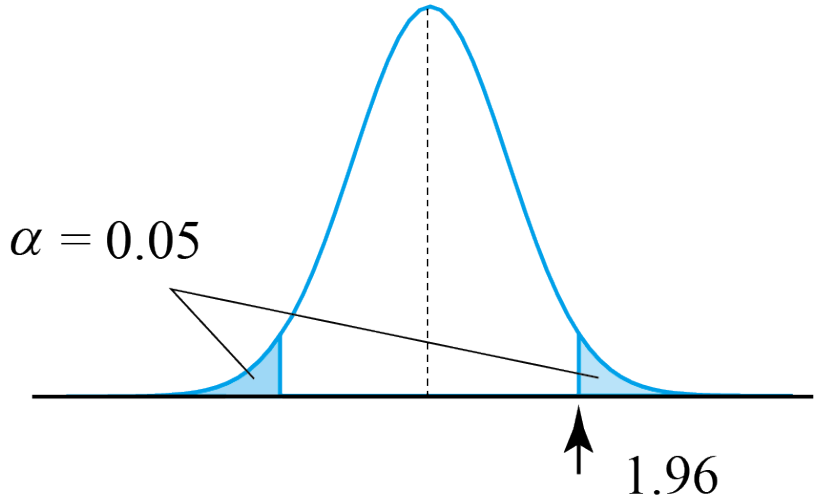


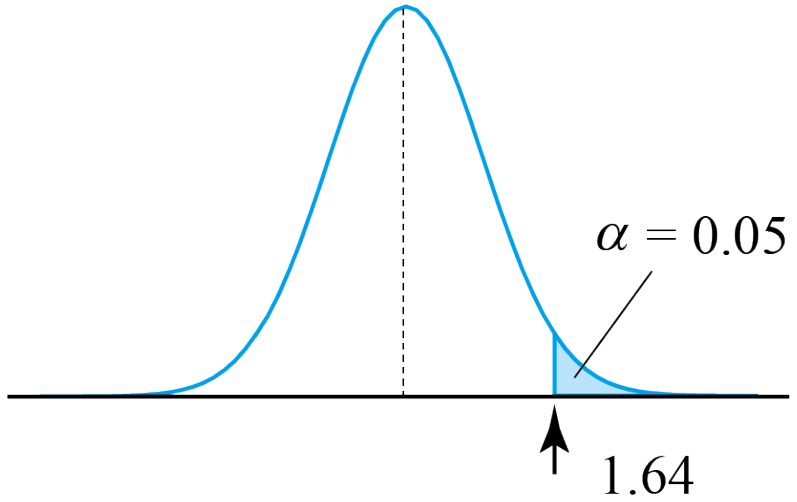




- 使い分けは文脈による
- 数理統計学の範疇ではない

両側 α 点と上側 α 点： $N(0,1)$ の場合





1.96 ＝ 両側 5 % 点 ＝ 上側 2.5 % 点

1.64 ＝ 上側 5 % 点 ＝ 両側 10 % 点

両側	α	0.3173	0.1000	0.0500	0.0455	0.0100	0.0027	0.0010
上側	$\alpha/2$	0.1587	0.0500	0.0250	0.0228	0.0050	0.0013	0.0005
	z	1.000	1.645	1.960	2.00	2.576	3.000	3.290

母平均の検定

母集団

母平均　μ

母分散　σ^2

無作為
復元抽出

標本

$X_1, X_2, \cdots, X_n$

基本的な問題

母平均を$\mu = m_0$ とみなしてよいか？

帰無仮説と対立仮説

H_0: $\mu = m_0$

H_1: $\mu \neq m_0$（両側検定）

H_1: $\mu > m_0$　または　H_1: $\mu < m_0$（片側検定）

有意水準　α

$\alpha = 0.05$,　$\alpha = 0.01$ など

標本平均の分布（復習）

母集団

母平均 μ

母分散 σ^2

無作為
復元抽出

標本（大きさ：n）

標本平均：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

不偏分散：$U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

母集団	基本定理	使う確率分布
正規母集団	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	標準正規分布
正規母集団	$\frac{\bar{X}-\mu}{U/\sqrt{n}} \sim t_{n-1}$	自由度 $n-1$ の t-分布
一般の母集団 n：大きい	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 近似的に成立	標準正規分布

例題 9.1

ある中学校で 1 年生 44 名に集団式知能検査を実施したところ，偏差値の平均は 52.4 であった. この学校の 1 年生は平均的な生徒といえるか.ただし, 全国における知能検査の偏差値は $N(50, 10^2)$ に従うことが知られている.

例題 9.1

母集団
母平均 μ
母分散 $\sigma^2 = 10^2$

標本
大きさ $n = 44$

標本平均の実現値
$\bar{x} = 52.4$

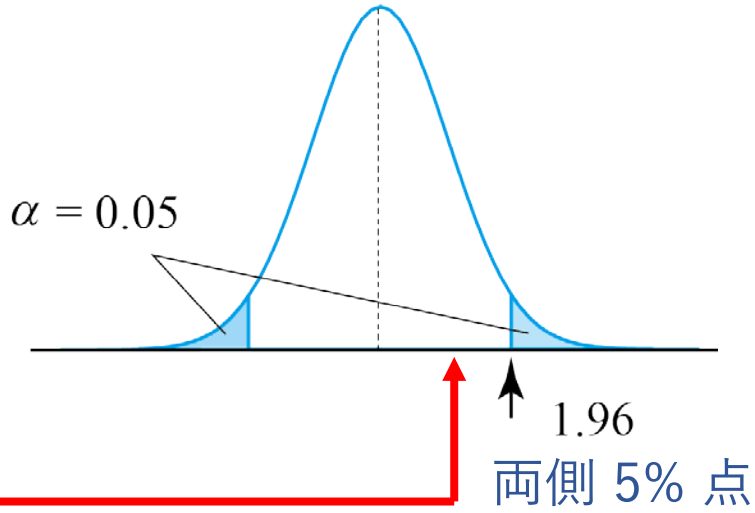
帰無仮説と対立仮説　$H_0: \mu = 50$　$H_1: \mu \neq 50$

有意水準　$\alpha = 0.05$

検定統計量　H_0 の下で,

$$\bar{X} \sim N\left(50, \frac{10^2}{44}\right) = N(50, 1.51^2)$$

$$Z = \frac{\bar{X} - 50}{1.51} \sim N(0,1)$$



実現値　$z = \dfrac{52.4 - 50}{1.51} = 1.59$

結論　有意水準 5% の両側検定で H_0 は棄却されない.
（有意でない）

例題 9.2

あるメーカーの電化製品の寿命は, カタログによると平均 $\mu = 1200$ 時間, 標準偏差 $\sigma = 150$ 時間と書かれている. $n = 10$ 個のサンプルについてテストしたとき, 平均寿命が $\bar{x} = 1100$ 時間であった. カタログは偽りといえるか .

例題 9.2

母集団
母平均 μ
母分散 $\sigma^2 = 150^2$

標本
大きさ $n = 10$

標本平均の実現値
$\bar{x} = 1100$

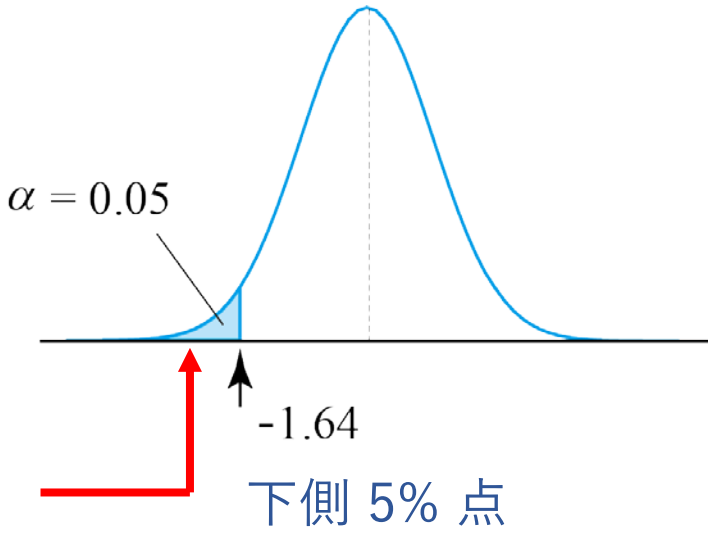
帰無仮説と対立仮説　$H_0: \mu = 1200$　$H_1: \mu < 1200$

有意水準　$\alpha = 0.05$

検定統計量　H_0 の下で,

$$\bar{X} \sim N\left(1200, \frac{150^2}{10}\right) = N(1200, 47.4^2)$$

$$Z = \frac{\bar{X} - 1200}{47.4} \sim N(0,1)$$



実現値　$z = \dfrac{1100 - 1200}{47.4} = -2.11$

結論　有意水準 5% の片側検定で H_0 は棄却される.
（有意である）

P 値

- 伝統的な仮説検定
  有意水準 α を示して H_0 の棄却・採択を述べる.

- P値を示す
  棄却・採択の判断はせず, 実現値が帰無仮説 H_0 の下で, どのくらい外れているかを数量的に示す.

定 義

P値 = 実現値 x を含めて, それ以上起こりにくい実現値が出現する確率

例題 9.3

公平なコインかどうか確認のため 100回振ったところ表が 64 回出た.

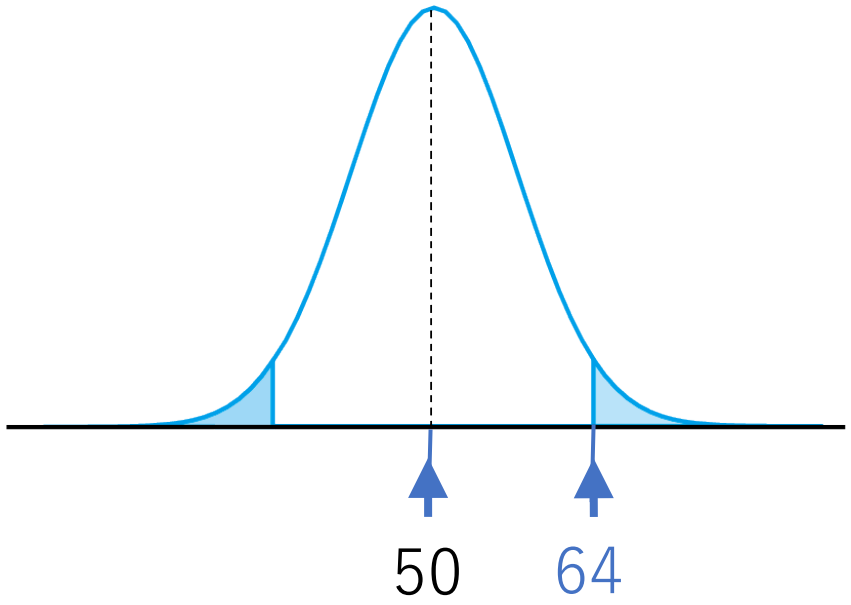
例題 9.3

公平なコインかどうか確認のため 100回振ったところ表が 64 回出た.

統計的有意性と P 値に関する ASA 声明
http://biometrics.gr.jp/news/all/ASA.pdf

$H_0: p = 0.5 \quad H_1: p \neq 0.5$

$X \sim B(100, 0.5) = N(50, 5^2)$



$$P = 2P(X \geq 64)$$

$$= 2P\left(\frac{X-50}{5} \geq \frac{64-50}{5}\right)$$

$$= 2P(Z \geq 2.8)$$

$$= 0.0052$$

今起きた現象の「稀さ」を表す確率.
これをどう判断するかはお任せするね.

例題 9.4

ある県の統計によると, 満 6 歳児の平均身長は 108.6 (cm) であるという. 同県のある小学校の 6 歳児 27 名について身長を調べ たところ, 平均 $\bar{x} = 109.7$ (cm), 不偏分散 $u^2 = 4.06^2$ (cm^2) であった. この結果から, 同校児童の身長は県平均に比べて高いといえるか.

例題 9.4

ある県の統計によると, 満 6 歳児の平均身長は 108.6 (cm) であるという. 同県のある小学校の 6 歳児 27 名について身長を調べ たところ, 平均 $\bar{x} = 109.7$ (cm), 不偏分散 $u^2 = 4.06^2$ (cm^2) であった. この結果から, 同校児童の身長は県平均に比べて高いといえるか.

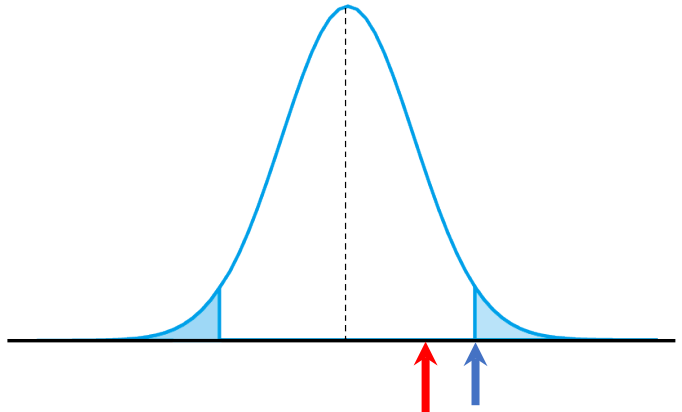
帰無仮説と対立仮説　$H_0: \mu = 108.6$　$H_1: \mu \neq 108.6$

有意水準　$\alpha = 0.05$

検定統計量

母分散未知

$$T = \frac{\bar{X} - \mu}{U/\sqrt{n}} = \frac{\bar{X} - 108.6}{U/\sqrt{27}} \sim t_{26}$$



$t_{26}(0.025) = 2.056$

両側 5% 点

実現値　$t = \dfrac{109.7 - 108.6}{4.06/\sqrt{27}} = 1.408$

結論　有意水準 5% の両側検定で H_0 は棄却されない.

例題 9.5

ある溶液に含まれる物質の濃度 (%) を測定して次のデータを得た.

12.6　13.4　14.1　12.4　11.2　12.5　10.9　11.8　11.6　13.1

真の濃度を μ として，仮説 $H_0: \mu = 12$ を検定せよ.

【練習（10分）】

例題 9.5　ある溶液に含まれる物質の濃度 (%) を測定して次のデータを得た.

12.6　13.4　14.1　12.4　11.2　12.5　10.9　11.8　11.6　13.1

真の濃度を μ として，仮説 $H_0: \mu = 12$ を検定せよ.

帰無仮説と対立仮説　$H_0: \mu = 12$　　$H_1: \mu \neq 12$

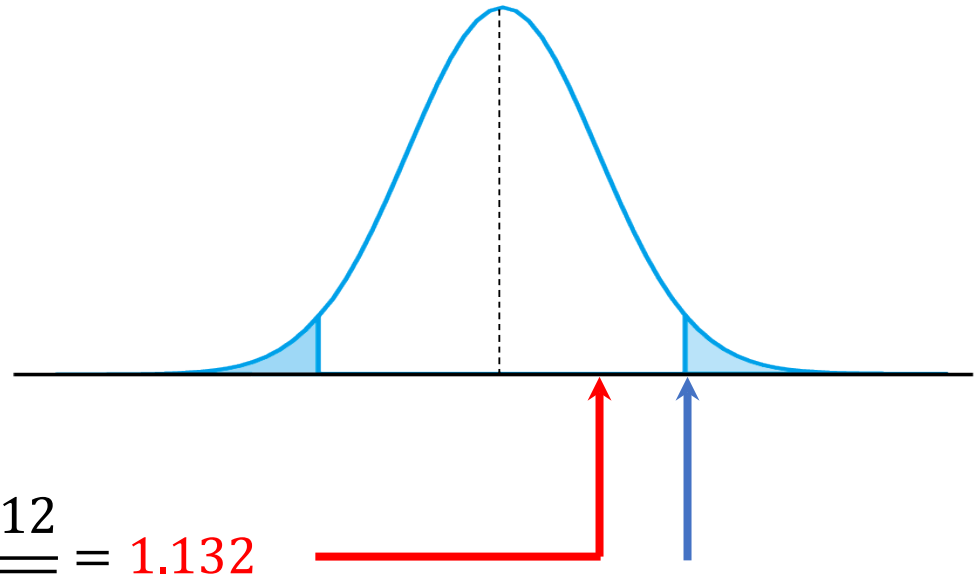
有意水準　$\alpha = 0.05$

検定統計量

母分散未知

$$T = \frac{\bar{X} - \mu}{U/\sqrt{n}} = \frac{\bar{X} - 12}{U/\sqrt{10}} \sim t_9$$

実現値　$\bar{x} = 12.36$　$u^2 = 1.0116 = 1.006^2$　$t = \dfrac{12.36 - 12}{1.006/\sqrt{10}} = 1.132$

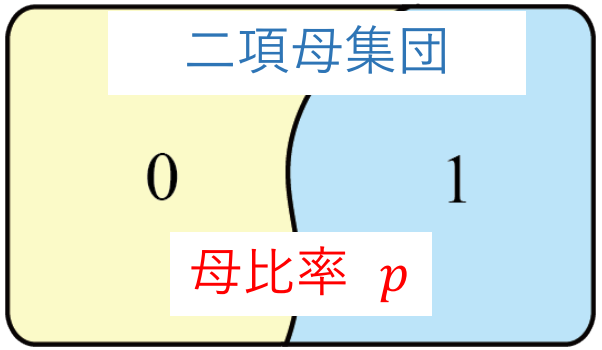


$t_9(0.025) = 2.262$

両側 5% 点

結論　有意水準 5% の両側検定で H_0 は棄却されない.

二項母集団の母比率の検定



無作為
復元抽出

標本

大きさ：n

標本比率：$\hat{p} = \frac{1}{n}\sum_{k=1}^{n} X_k$

帰無仮説と対立仮説　$H_0: p = p_0$　$H_1: p \neq p_0$（両側検定）

$H_1: p > p_0$　または　$H_1: p < p_0$　（片側検定）

検定統計量

$$n\hat{p} \sim B(n, p_0) \approx N(np_0, np_0(1-p_0))$$

$$\hat{p} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

標準化

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

例題 9.6

あるクラスの平均出席率は 0.90 であるといわれている. ある日の欠席者は, 160 人中 25 人であった. この日は通常の出席ではないといえるか. 有意水準は $\alpha = 0.01$ とせよ.

例題 9.6

あるクラスの平均出席率は 0.90 であるといわれている. ある日の欠席者は, 160 人中 25 人であった. この日は通常の出席ではないといえるか. 有意水準は $\alpha = 0.01$ とせよ.

帰無仮説と対立仮説　$H_0: p = 0.9$　$H_1: p \neq 0.9$（両側検定）

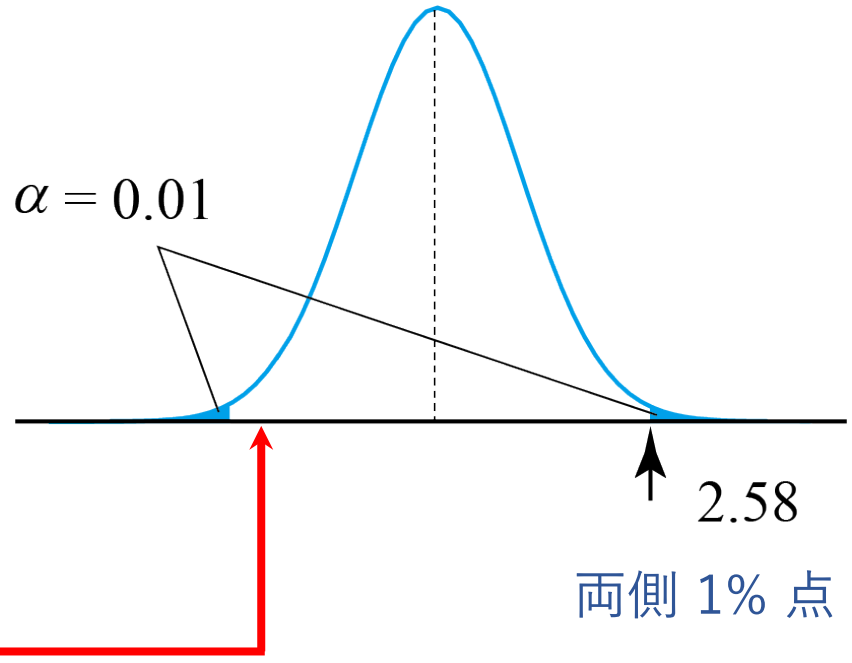
有意水準　$\alpha = 0.01$

検定統計量　$n\hat{p} \sim B(160, 0.9) \approx N(144, 3.79^2)$

$$\hat{p} \sim N\left(\frac{144}{160}, \frac{3.79^2}{160^2}\right) = N(0.9, 0.0237^2)$$

$$Z = \frac{\hat{p} - 0.9}{0.0237} \sim N(0,1)$$

実現値　$$z = \frac{0.8438 - 0.9}{0.0237} = -2.37$$



結論　有意水準 1% の両側検定で H_0 は棄却されない.

例題 9.7

ある意見項目に対する賛成率を 30% は欲しいと思われていた．実際に，調査では 80 人中 23 人の賛成を得た．賛成率の目標を達成したと考えてよいか.

【練習（10分）】

例題 9.7

ある意見項目に対する賛成率を 30% は欲しいと思われていた．実際に，調査では 80 人中 23 人の賛成を得た．賛成率の目標を達成したと考えてよいか.

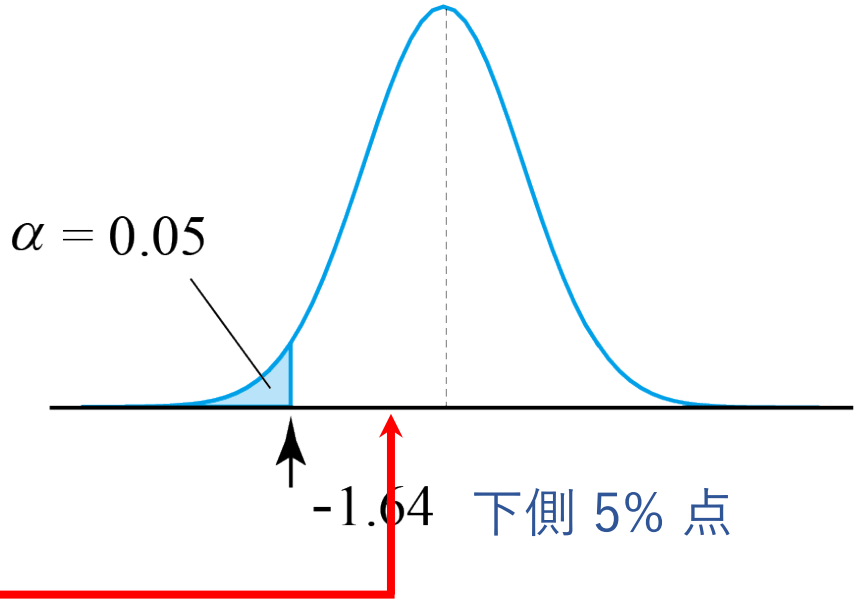
帰無仮説と対立仮説　$H_0: p = 0.3$　$H_1: p < 0.3$（片側検定）

有意水準　$\alpha = 0.05$

検定統計量　$n\hat{p} \sim B(80, 0.3) \approx N(24, 4.10^2)$

$$\hat{p} \sim N\left(\frac{24}{80}, \frac{4.10^2}{80^2}\right) = N(0.3, 0.0512^2)$$

$$Z = \frac{\hat{p} - 0.3}{0.0512} \sim N(0,1)$$



実現値　$z = \dfrac{0.2875 - 0.3}{0.0512} = -0.244$

結論　有意水準 5% の片側検定で H_0 は棄却されない.

2 種類の過誤

帰無仮説 H_0 をめぐって

採否\真偽	H_0 は真	H_0 は偽
H_0 を採択	○	第 2 種の誤り β
H_0 を棄却	第 1 種の誤り α	○

第 1 種の誤り
＝ 生産者危険
＝ あわて者の間違い

第 2 種の誤り
＝ 消費者危険
＝ ぼんやり者の間違い

第 1 種の誤り確率　α ＝ 有意水準　← 自分で設定する

第 2 種の誤り確率　β　← 要注意

第2種誤り確率 β は一般には不明

例　コインを100回投げて表が58回出た. コインは公平といえるか?

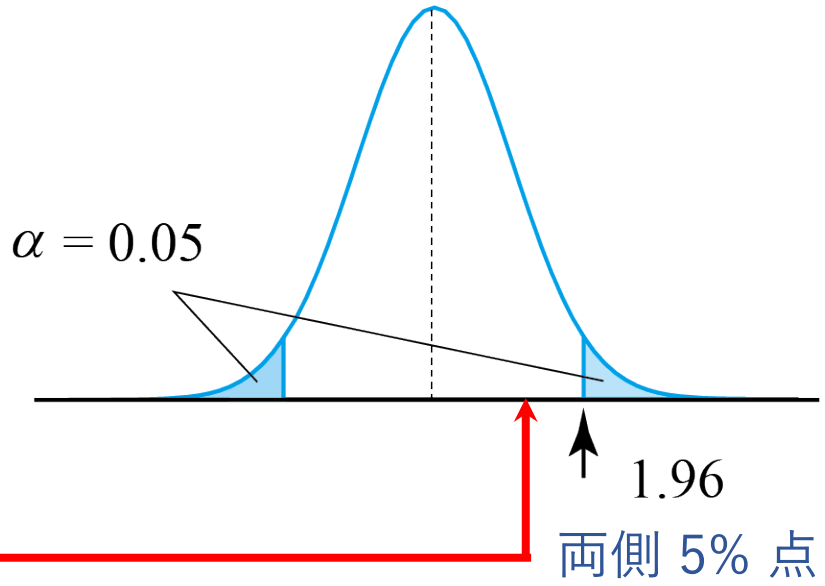
帰無仮説と対立仮説　$H_0: p = 0.5$　$H_1: p \neq 0.5$

有意水準　$\alpha = 0.05$

検定統計量　表の回数　$X \sim B(100, 0.5) \approx N(50, 5^2)$

$$Z = \frac{X - 50}{5} \sim N(0,1)$$

実現値　$z = \frac{58 - 50}{5} = 1.6$



結論　有意水準 $\alpha = 0.05$ の両側検定によって
H_0 は棄却されない＝採択される

この採択とした結論を誤る確率＝第2種誤り確率 β

β の評価は困難

「$H_0: p = 0.5$」でないとすると, 可能な p は無限にあって特定できないから

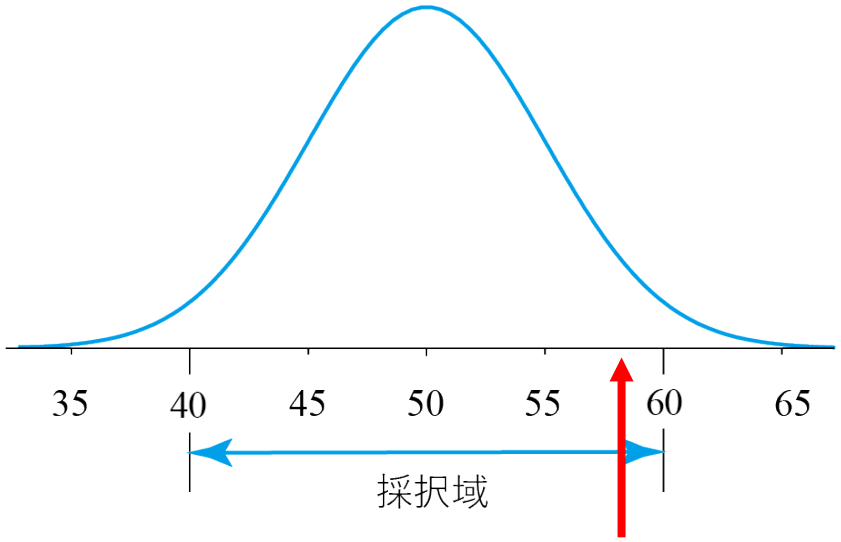
例　コインを100回投げて表が58回出た. コインは公平といえるか?

$H_0: p = 0.5$ を有意水準$\alpha = 0.05$ で両側検定

表の回数 $X \sim B(100, 0.5) \approx N(50, 5^2)$

標準化しないで考察しよう.

採択域 $50 \pm 1.96 \times 5 \approx 50 \pm 10$



実現値 58

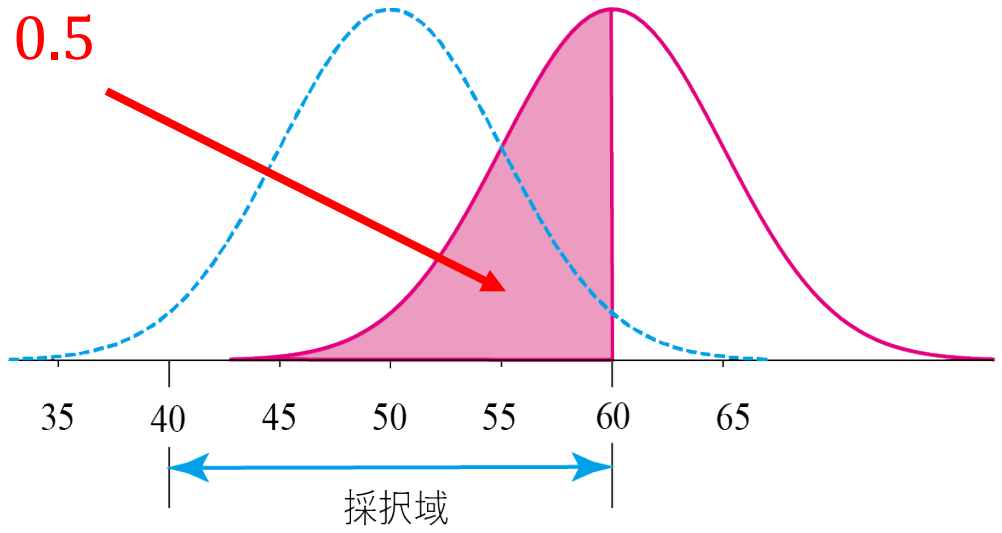
「$H_0: p = 0.5$」でないとすると, 可能な p は無限にあって特定できない

※ 仮に $p = 0.6$ としてみる

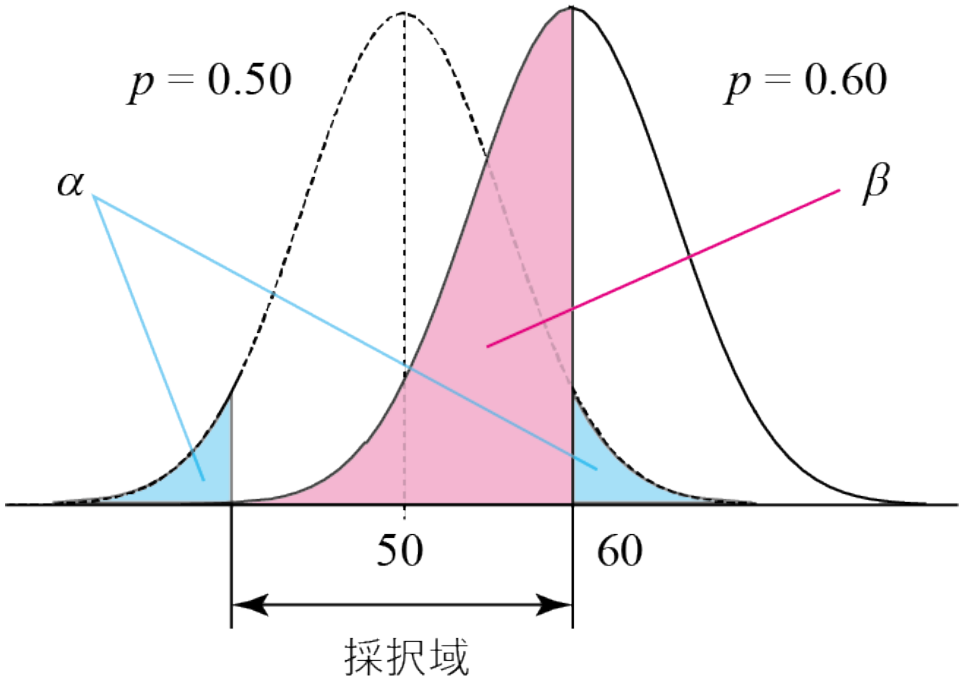
$p = 0.6$ の下で

表の回数 $Y \sim B(100,0.6) \approx N(60, 4.9^2)$

$\beta \approx 0.5$



α と β はトレードオフの関係



(1) α 小 ⇔ 採択域が大 ⇔ β 大

(2) α, β ともに小さくするためには,
標本数 n を大きくする.

(3)「H_0 を採択」という判断を誤る場合,
真の母数が H_0 で仮定した母数に近いほど β は大きい

- 「H_0 を採択する」は消極的な採択.
  はっきり否定するだけの状況ではないという意味.
  そこで「H_0 を棄却できない」ということが多い.

Lecture 9

母平均の検定

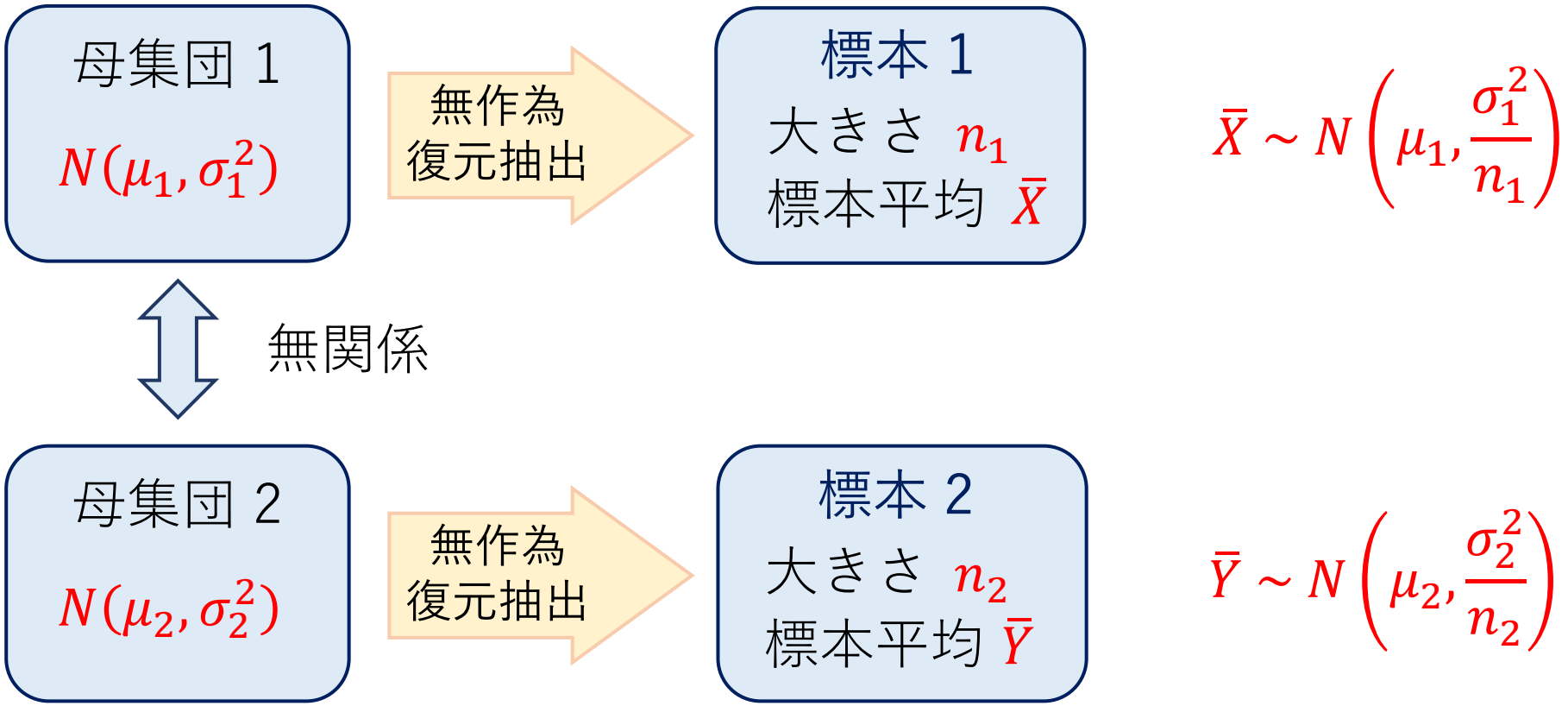
おわり

Lecture 10

母集団の比較

母平均の差の検定

※ 2つの母集団の母平均に差があるかを検定する



$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

確率変数の和（復習）

- 一般の確率変数 $X,\ Y$ と定数 a, b に対して

（1）[平均値の線形性] $E[aX+bY]=aE[X]+bE[Y]$

（2）分散は線形性をもたないが $V[aX]=a^2V[X]$

- 確率変数 $X,\ Y$ が独立ならば

（3）[平均値の乗法性] $E[XY]=E[X]E[Y]$

（4）[分散の加法性] $V[X+Y]=V[X]+V[Y]$

- 確率変数 $X\sim N(\mu_1,\sigma_1^2),\ Y\sim N(\mu_2,\sigma_2^2)$ が独立ならば

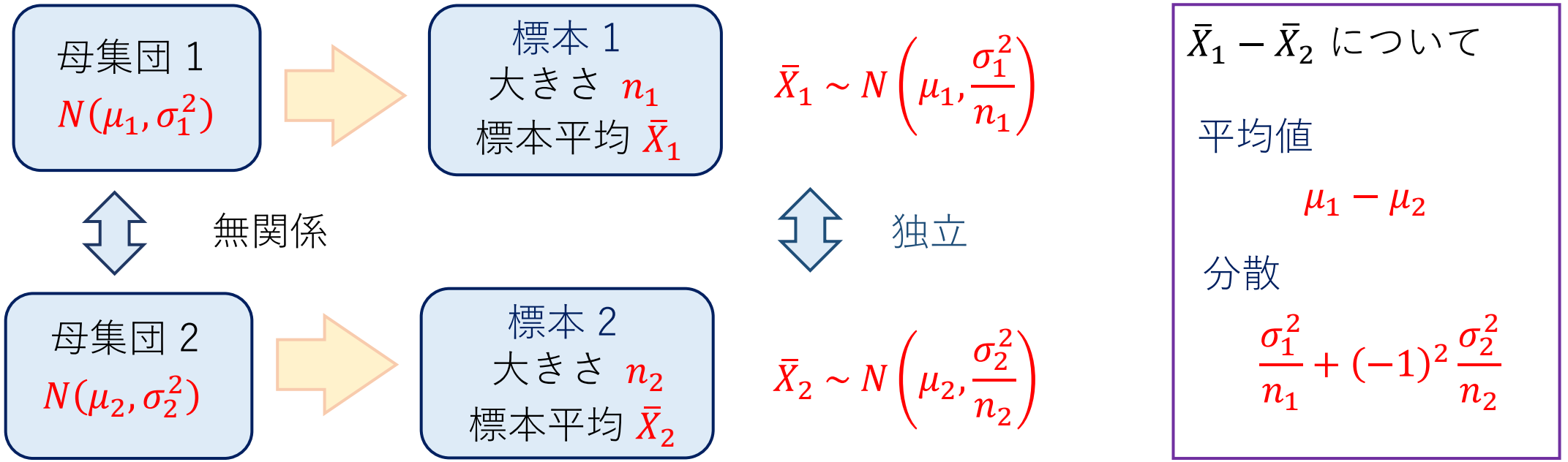
（5）線形結合 $aX+bY$ も正規分布に従い,

$$aX+bY\ \sim\ N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2)$$

定理　（標本平均の差の分布）

正規母集団 $N(\mu_1, \sigma_1^2)$ から取り出した n_1 個の無作為標本の標本平均を $\bar{X}_1$，別の正規母集団 $N(\mu_2, \sigma_2^2)$ から取り出した n_2 個の無作為標本の標本平均を $\bar{X}_2$ とするとき，

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2\ , \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$



$\bar{X}_1 - \bar{X}_2$ について

平均値

$$\mu_1 - \mu_2$$

分散

$$\frac{\sigma_1^2}{n_1} + (-1)^2 \frac{\sigma_2^2}{n_2}$$

母平均の差の推定と検定（母分散既知）

母集団 1
$N(\mu_1, \sigma_1^2)$

標本 1
大きさ n_1
標本平均 $\bar{X}_1$

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

無関係　　独立

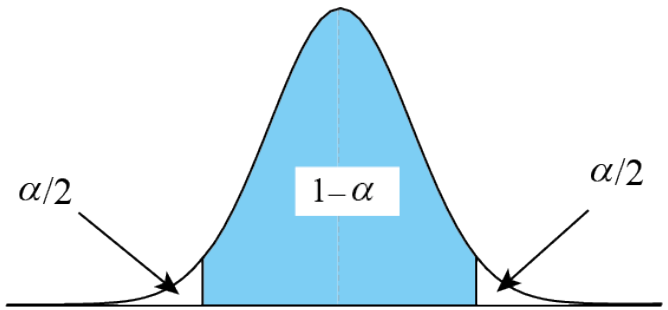
母集団 2
$N(\mu_2, \sigma_2^2)$

標本 2
大きさ n_2
標本平均 $\bar{X}_2$

$$\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2\ , \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$



両側 α点 ＝ 上側 $\alpha/2$点 ＝ $z(\alpha/2)$

$$P\left(-z(\alpha/2) \le Z \le z(\alpha/2)\right) = 1 - \alpha$$

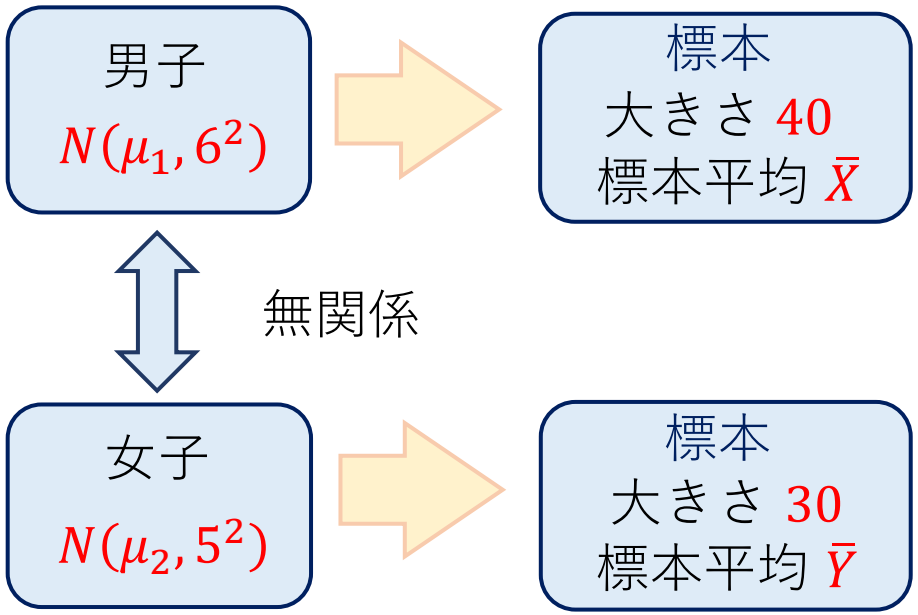
例題 10.1

ある大学の男子学生 40 人と女子学生 30 人の身長を調べたところそれぞれ 171.47, 157.22 (cm) であった. それぞれの分散は既知であって 36 , 25 (cm^2) であるとして, 男女学生の身長差の 95% 信頼区間を求めよ.

例題 10.1

ある大学の男子学生 40 人と女子学生 30 人の身長を調べたところそれぞれ 171.47, 157.22 (cm) であった. それぞれの分散は既知であって 36 , 25 (cm^2) であるとして, 男女学生の身長差の 95% 信頼区間を求めよ.

男子学生, 女子学生の身長は正規分布に従うものとして, それぞれの平均値をそれぞれ μ_1, μ_2 とおく.



$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

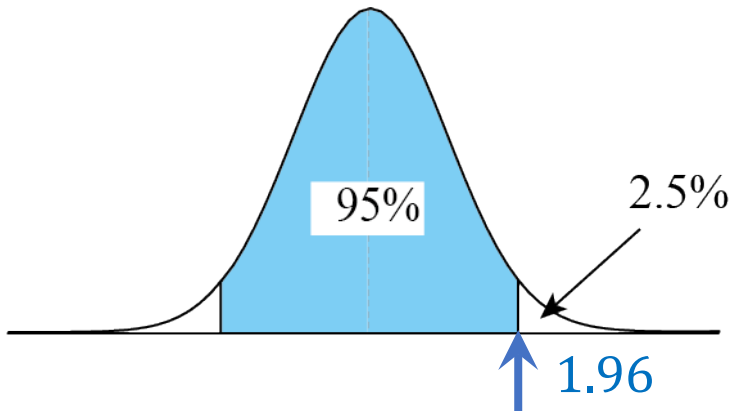
$$= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{6^2}{40} + \frac{5^2}{30}}}$$

$$= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{1.317} \sim N(0,1)$$

例題 10.1

ある大学の男子学生 40 人と女子学生 30 人の身長を調べたところそれぞれ 171.47, 157.22 (cm) であった. それぞれの分散は既知であって 36 , 25 (cm^2) であるとして, 男女学生の身長差の 95% 信頼区間を求めよ.

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{1.317} \sim N(0,1)$$



$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$-1.96 \leq Z \leq 1.96$$

$$\bar{X} - \bar{Y} - 1.96 \times 1.317 \leq \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} + 1.96 \times 1.317$$

実現値 $\bar{x} = 171.47, \quad \bar{y} = 157.22$

$\mu_1 - \mu_2$ に対する 95% 信頼区間

$$14.25 \pm 2.58$$

例題 10.2

ある学年で知能指数を測定し, 男女別に集計したところ次の結果が得られた. 男女差があるといえるか. ただし, 知能指数の分布は標準偏差 15 の正規分布に従うといわれている.

	平均	人数
男生徒	103	40
女生徒	101	35

男生徒, 女生徒の平均値をそれぞれ μ_1, μ_2 とおく. 分散は 15^2 を用いる.

帰無仮説と対立仮説　$H_0: \mu_1 = \mu_2$　　$H_1: \mu_1 \neq \mu_2$

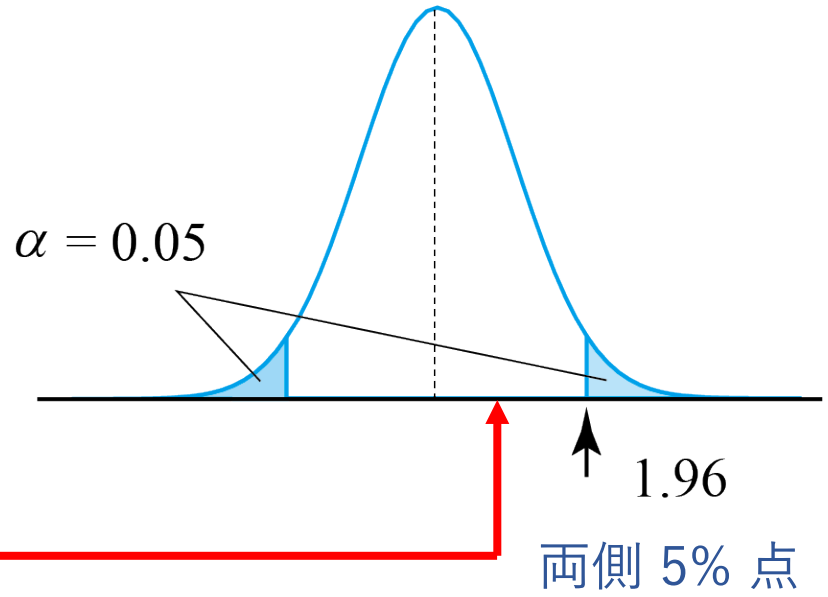
有意水準　$\alpha = 0.05$

検定統計量

$$\bar{X} \sim N\left(\mu_1, \frac{15^2}{40}\right) \qquad \bar{Y} \sim N\left(\mu_2, \frac{15^2}{35}\right)$$

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{15^2}{40} + \frac{15^2}{35}\right) = N(0, 3.47^2)$$

$$Z = \frac{\bar{X} - \bar{Y}}{3.47} \sim N(0,1)$$



実現値　$z = \dfrac{103 - 101}{3.47} = 0.576$

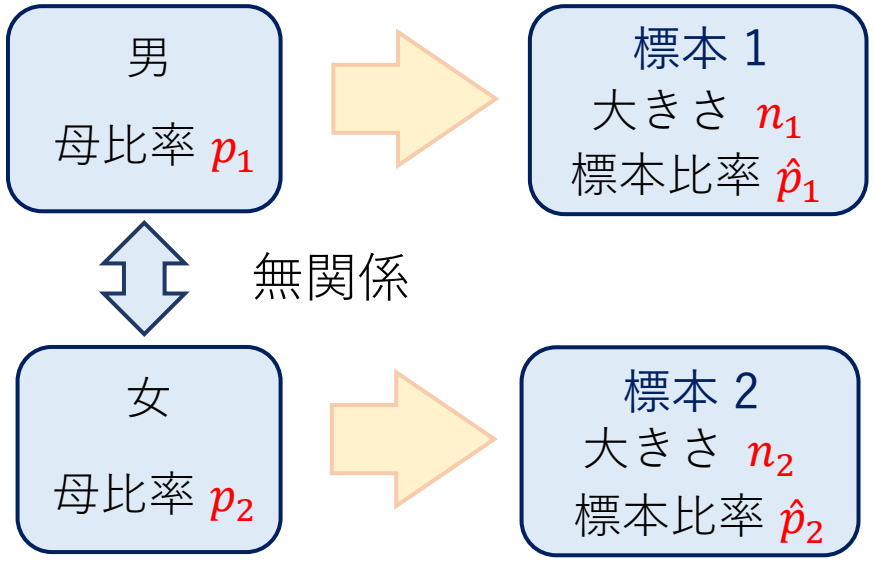
結論　有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却されない.

例題 10.3 （母比率の差）

ある意見項目に対する賛否を男女別に集計したところ，次の結果を得た．賛成者の比率に男女差があるといえるか．

	賛成	反対	計
男	58 (0.592)	40 (0.408)	98 (1.000)
女	28 (0.394)	43 (0.606)	71 (1.000)

男, 女の賛成の母比率をそれぞれ p_1, p_2 とおく.



$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$$

独立

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

$$\hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

稲垣他の本(pp.138-140)の記述は不適切

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　　$H_1: p_1 \neq p_2$

有意水準　$\alpha = 0.05$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) = N(0, 0.078^2)$

p を実現値から推定　$n_1 = 98$　$n_2 = 71$　$p = \frac{58+28}{98+71} = 0.509$

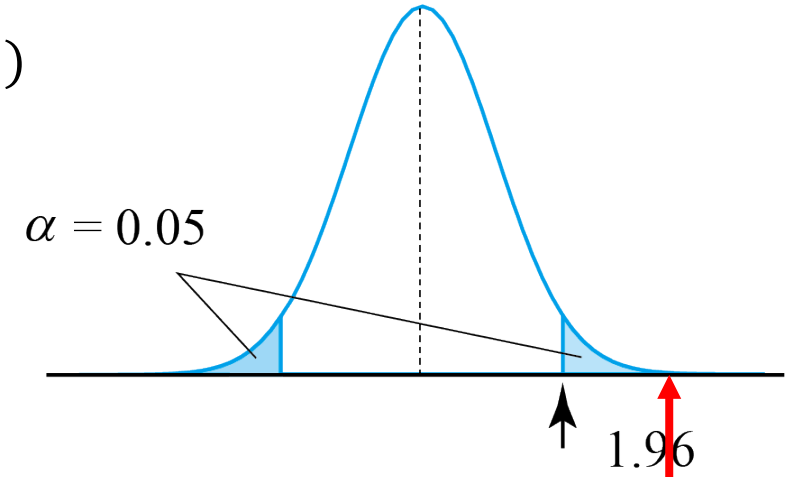
	賛成	反対	計
男	58 (0.592)	40 (0.408)	98 (1.000)
女	28 (0.394)	43 (0.606)	71 (1.000)

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　$H_1: p_1 \neq p_2$

有意水準　$\alpha = 0.05$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) = N(0, 0.078^2)$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{0.078} \sim N(0,1)$$



実現値　$\hat{p}_1 = \frac{58}{98} = 0.592$　$\hat{p}_2 = \frac{28}{71} = 0.394$

$$z = \frac{0.592 - 0.394}{0.078} = 2.54$$

結論　有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される.

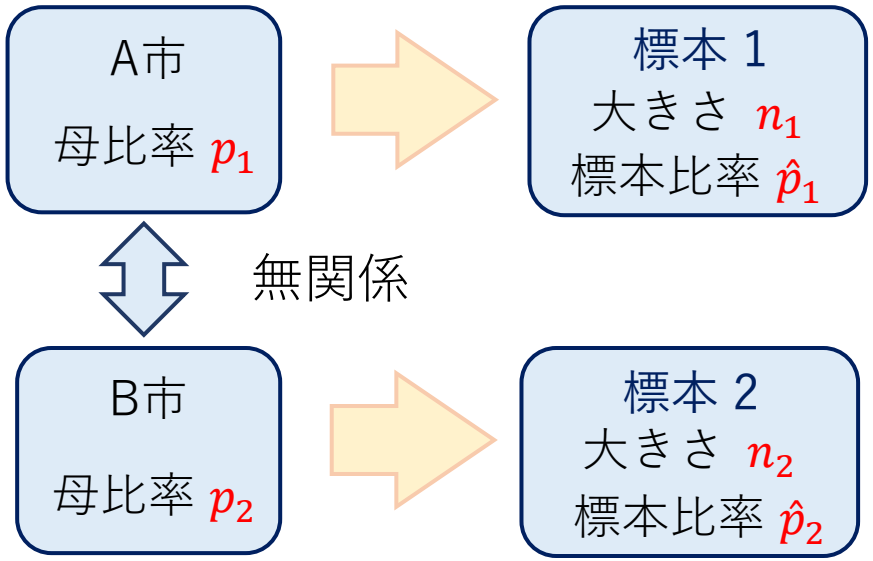
例題 10.4 （母比率の差）

ある疾患に対する免疫の有無を調べるため, A市とB市で標本調査をおこなって, 次の結果を得た. 免疫をもっ ている人の比率は両市で差があると考えられるか.

	免疫あり	免疫なし	計
A市	35 (0.35)	65 (0.65)	100 (1.00)
B市	60 (0.50)	60 (0.50)	120 (1.00)

【練習（10分）】

A市, B市の免疫ありの母比率をそれぞれ p_1, p_2 とおく.



$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$$

独立

$$\hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

	免疫あり	免疫なし	計
A市	35 (0.35)	65 (0.65)	100 (1.00)
B市	60 (0.50)	60 (0.50)	120 (1.00)

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　　$H_1: p_1 \neq p_2$

検定統計量　H_0 の下で,　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) = N(0, 0.067^2)$

p は観測値を用いて推定する　$p = \frac{35+60}{100+120} = 0.432$

A市, B市の免疫ありの母比率をそれぞれ p_1, p_2 とおく.

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　　$H_1: p_1 \neq p_2$

有意水準　$\alpha = 0.05$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

	免疫あり	免疫なし	計
A	35 (0.35)	65 (0.65)	100 (1.00)
B	60 (0.50)	60 (0.50)	120 (1.00)

$n_1 = 100$

$n_2 = 120$

$$p = \frac{35 + 60}{100 + 120} = 0.432$$

$$\hat{p}_1 - \hat{p}_2 \sim N(0, 0.067^2)$$

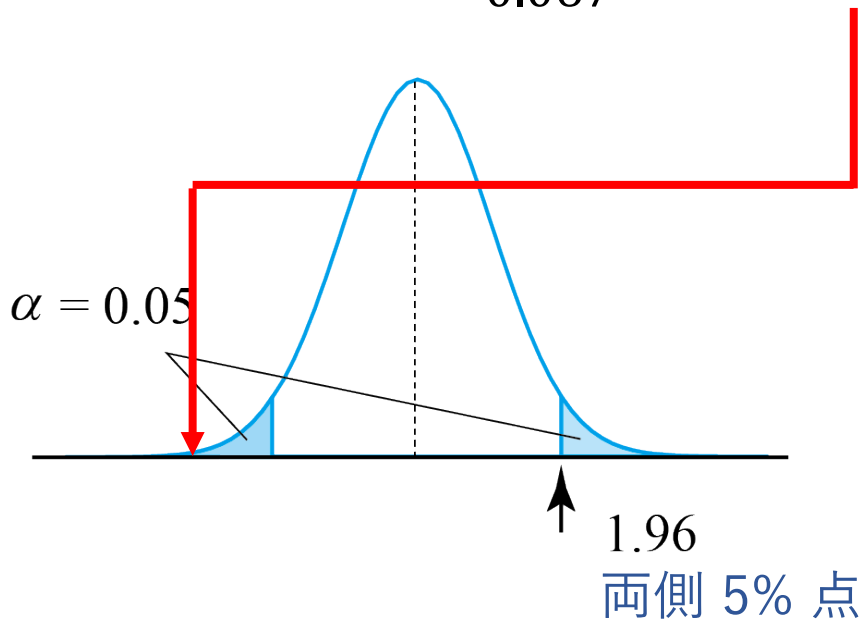
$$Z = \frac{\hat{p}_1 - \hat{p}_2}{0.067} \sim N(0, 1)$$

実現値

$$\hat{p}_1 = \frac{35}{100} = 0.35$$

$$\hat{p}_2 = \frac{60}{120} = 0.50$$

$$z = \frac{0.35 - 0.50}{0.067} = -2.239$$



両側 5% 点

結論　有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される.

例題 10.5

インフルエンザについて, 予防接種の有無と感染の有無のデータが次のようであった. 予防接種の効果はあるといえるか.

	感染	非感染
接種	18	67
非接種	45	65

【練習（10分）】

接種, 非接種群の感染率（母比率）をそれぞれ p_1, p_2 とおく.

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　　$H_1: p_1 \neq p_2$

有意水準　$\alpha = 0.01$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

	感染	非感染	合計
接種	18	67	85
非接種	45	65	110
合計	63	132	195

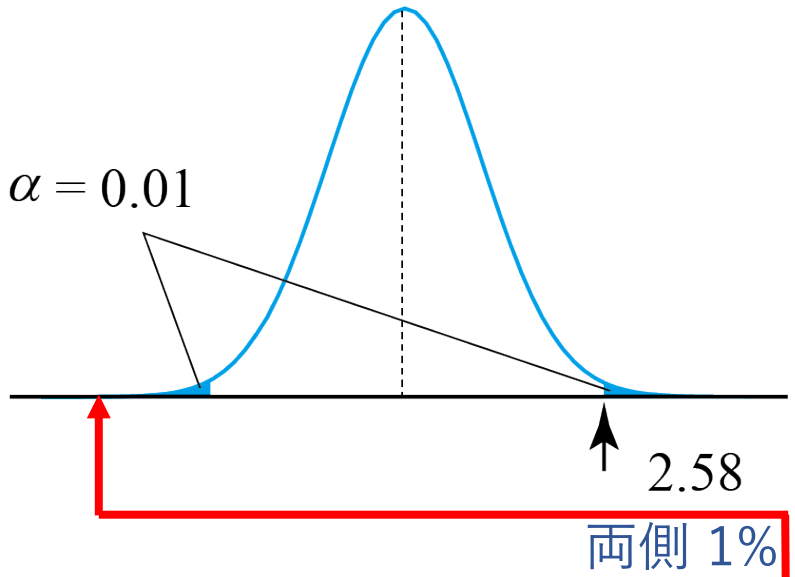
$n_1 = 85$

$n_2 = 110$

$p = \frac{18 + 45}{85 + 110} = 0.323$

$\hat{p}_1 - \hat{p}_2 \sim N(0, 0.0675^2)$

$Z = \frac{\hat{p}_1 - \hat{p}_2}{0.067} \sim N(0, 1)$



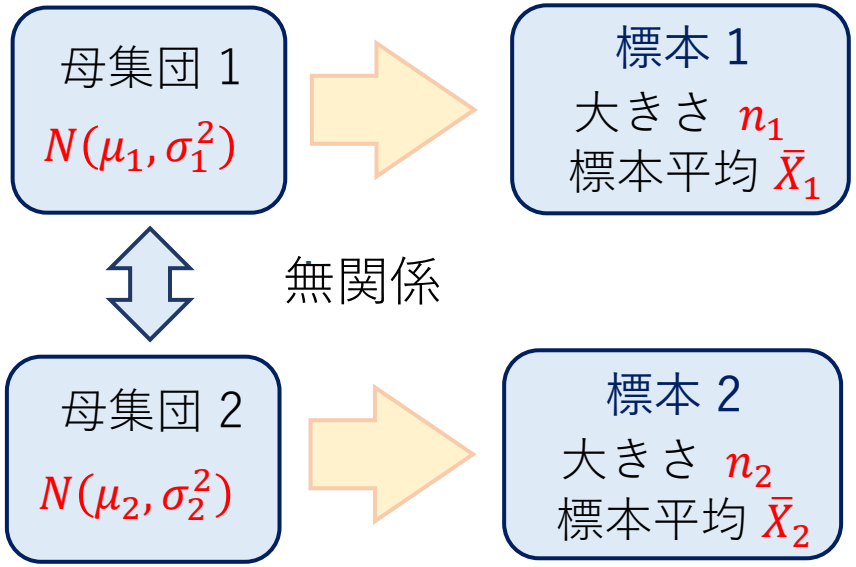
実現値　$\hat{p}_1 = \frac{18}{85} = 0.212$

$\hat{p}_2 = \frac{45}{110} = 0.409$

$z = \frac{0.212 - 0.409}{0.0675} = -2.91$

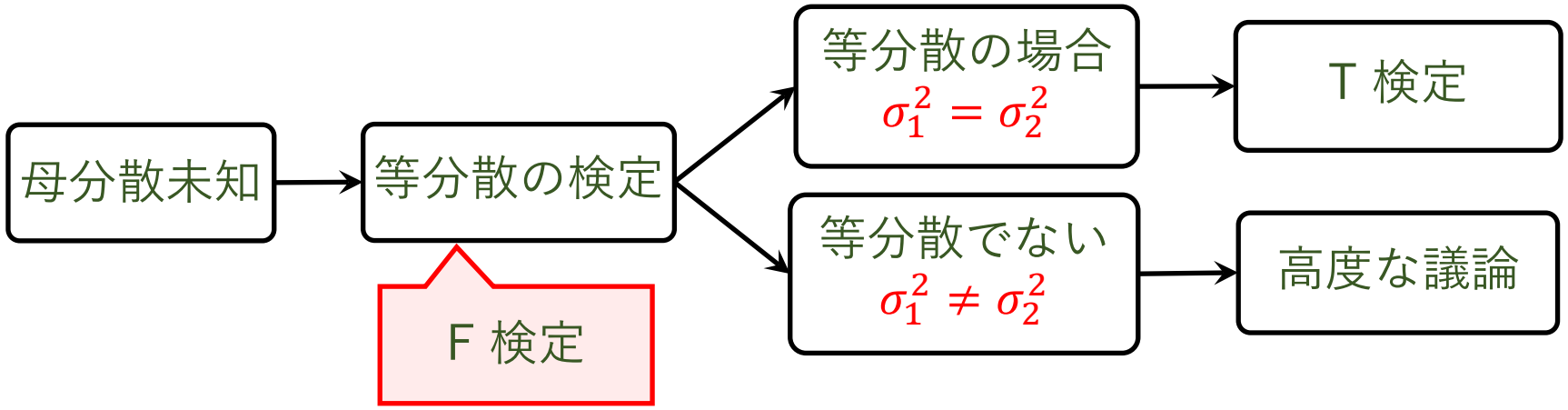
結論　有意水準 $\alpha = 0.01$ の両側検定によって H_0 は棄却される. （高度に有意である.）

母平均の差の推定と検定（母分散未知）



$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2\ , \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

この式は一般に正しいが，
母分散未知の場合は先に進めない．



Lecture 10

母集団の比較

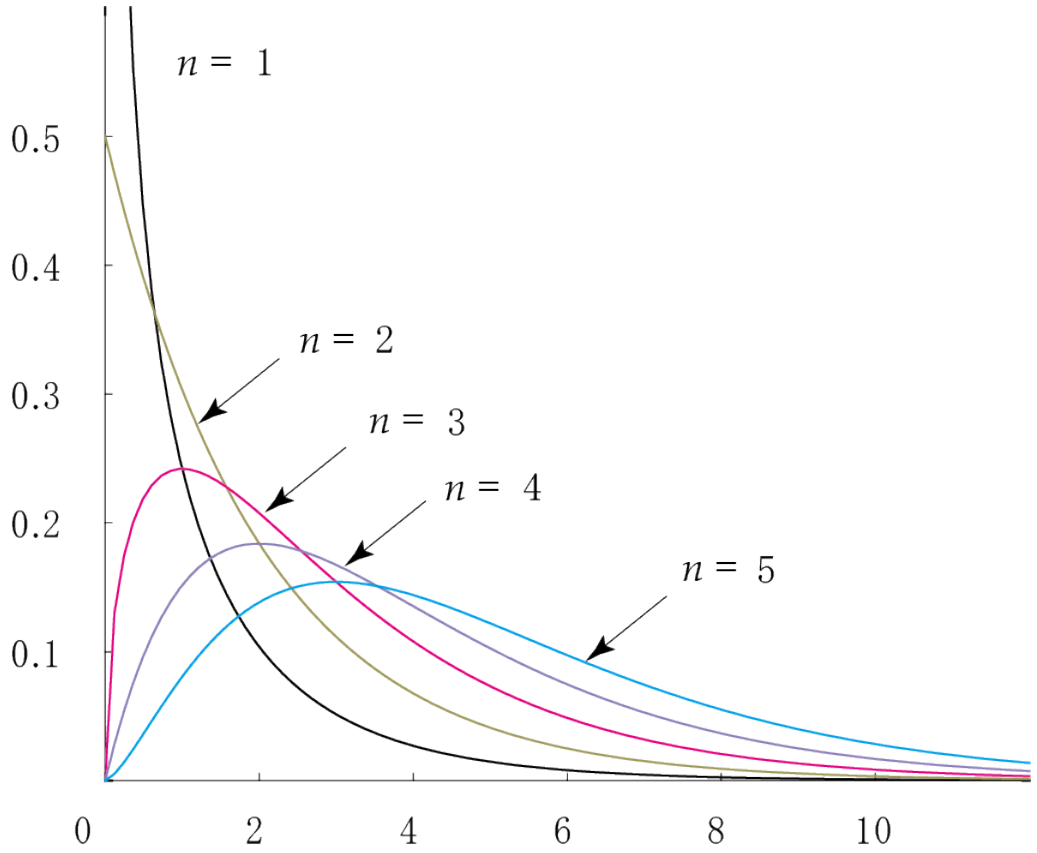
おわり

Lecture 11

カイ2乗検定

χ^2-分布

$$f_n(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \qquad (x \geq 0)$$



定理

$Z_1, Z_2, \cdots, Z_n$ が独立同分布な確率変数で，標準正規分布 $N(0,1)$ に従うとき，

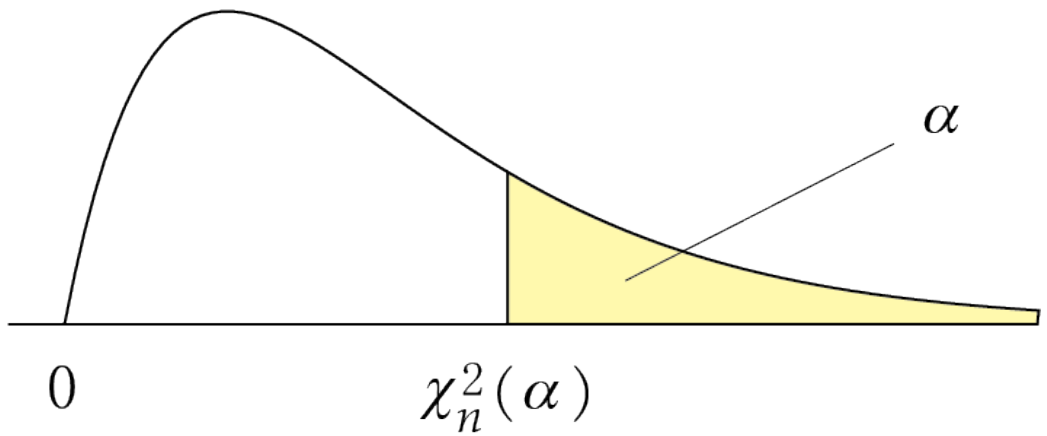
$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2 \qquad \text{（平方和）}$$

は自由度 n の χ^2-分布に従う

自由度 n の χ_n^2 -分布では，

平均値 $\mu = n$　分散 $\sigma^2 = 2n$

上側 α 点



$n\backslash\alpha$	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928

χ^2-値が大きい

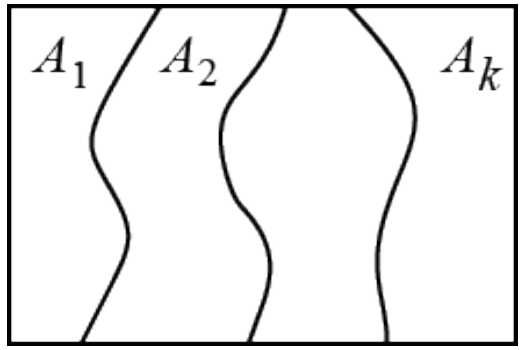
⇔ 想定からのずれが大きい

⇔ 有意差が認められる

例　$\chi_5^2(0.05) = 11.070$

分布の適合度検定

母集団



n 個の
無作為標本

属性	A_1	A_2	…	A_k	合計
観測度数	X_1	X_2	…	X_k	n
理論分布	p_1	p_2	…	p_k	1
理論度数	np_1	np_2	…	np_k	n

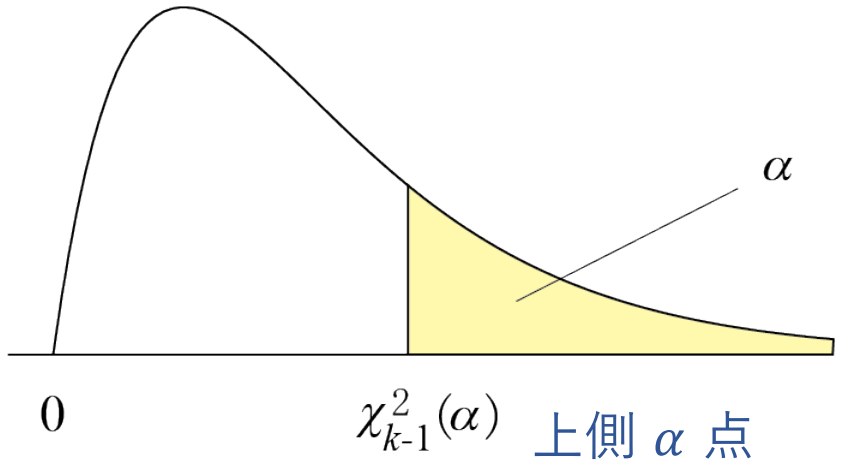
問題　観測度数をもとに，理論分布が妥当かどうかを検定する．

例（サイコロ振り）

目	1	2	3	4	5	6	合計
度数 X_i	24	18	16	22	23	17	120
理論度数 m_i	20	20	20	20	20	20	120

χ^2-検定

属性	A_1	A_2	…	A_k	合計
観測度数	X_1	X_2	…	X_k	n
理論分布	p_1	p_2	…	p_k	1
理論予想	m_1	m_2	…	m_k	n



定理　ピアソンの χ^2-値

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - m_i)^2}{m_i}$$

は, m_i が大きいとき(実用上, $m_i \geq 5$),

自由度 $k-1$ の χ^2-分布に近似的に従う.

検定の手順

1. データから χ^2-値を計算
2. 自由度 $k-1$ の χ^2_{k-1}-分布と比較
3. 上側 α 点を超えれば, 有意差を認める

例 題 11.1　サイコロを120回投げて出た目を記録した. このサイコロは公平と言えるだろうか?

目	1	2	3	4	5	6	合計
回数 X_i	24	18	16	22	23	17	120
理論予想 m_i	20	20	20	20	20	20	120

H_0：サイコロは公平である

H_1：サイコロは不公平である

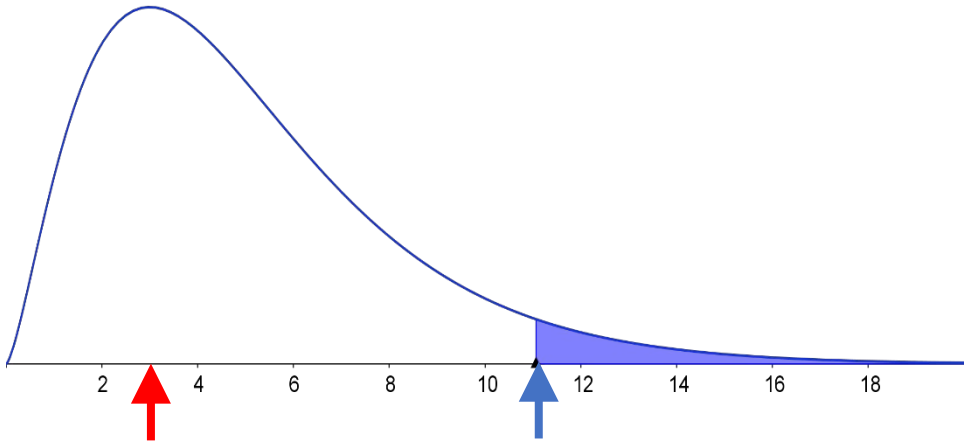
$\alpha = 0.05$　（有意水準 5%）

検定統計量

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i}$$

※ χ^2-値は自由度 5 の χ_5^2-分布に従う

実現値　$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 2.9$



2.9　　$\chi_5^2(0.05) = 11.07$　上側 5% 点

結論　有意水準 $\alpha = 0.05$ のカイ2乗検定によって，
H_0 は棄却できない（採択）

独立性の検定

2種類の属性 A, B に関するデータ ⇒ A, B の関連性（独立性）を問う

生まれ月 →

血液型 ↓

	B_1	B_2	…	B_s	合計
A_1	X_{11}	X_{12}	…	X_{1s}	$X_{1\cdot}$
A_2	X_{21}	X_{22}	…	X_{2s}	$X_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
A_r	X_{r1}	X_{r2}	…	X_{rs}	$X_{r\cdot}$
合計	$X_{\cdot 1}$	$X_{\cdot 2}$	…	$X_{\cdot s}$	n

定義　一般に，確率変数 A, B が独立とは

$$P(A = a, B = b) = P(A = a)P(B = b)$$

2つの属性 A, B が独立

⇕

$$\frac{X_{ij}}{n} = \frac{X_{i\cdot}}{n}\,\frac{X_{\cdot j}}{n}$$

ピアソンの χ^2- 値

	B_1	$\cdots$	B_j	$\cdots$	B_s	合計
A_1	X_{11}	$\cdots$	X_{1j}	$\cdots$	X_{1s}	$X_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	X_{i1}	$\cdots$	X_{ij}	$\cdots$	X_{is}	$X_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	X_{r1}	$\cdots$	X_{rj}	$\cdots$	X_{rs}	$X_{r\cdot}$
合計	$X_{\cdot 1}$	$\cdots$	$X_{\cdot j}$	$\cdots$	$X_{\cdot s}$	n

A と B の独立を仮定したとき,

$$P(A_i \cap B_j) = p_{ij} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}$$

$$\varepsilon_{ij} = \frac{X_{ij}}{n} - p_{ij} = \frac{X_{ij}}{n} - \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}$$

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\varepsilon_{ij}^2}{p_{ij}}$$

$$= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(nX_{ij} - X_{i\cdot}X_{\cdot j}\right)^2}{X_{i\cdot}X_{\cdot j}}$$

これが自由度 $(r-1)(s-1)$ のχ^2-分布に従う.

例 題 11.2 （予防接種の有効性）

実測データ

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

相対度数で表示

	発病有	発病無	合計
予防接種有	0.11	0.51	0.62
予防接種無	0.145	0.235	0.38
合計	0.255	0.745	1

実測データ

	発病有	発病無	合計
予防接種有	0.11	0.51	0.62
予防接種無	0.145	0.235	0.38
合計	0.255	0.745	1

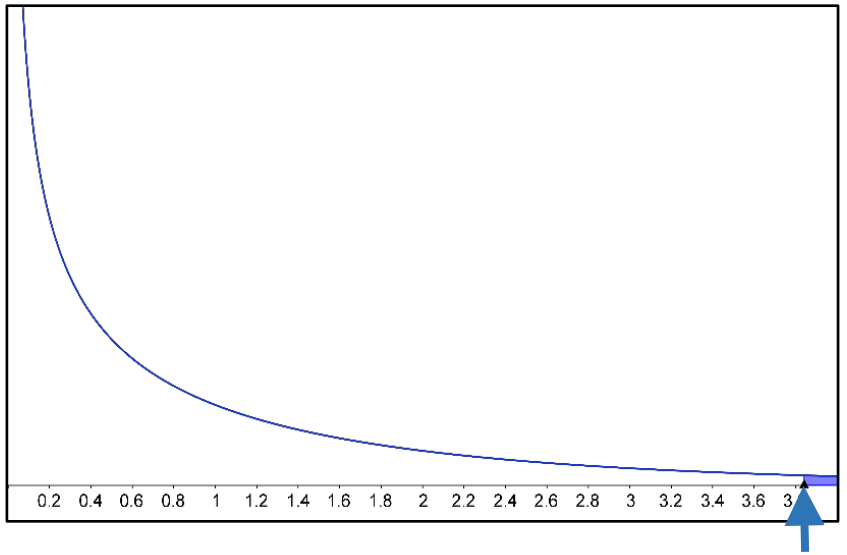
H_0（独立性）を仮定した理論値 p_{ij}

	発病有	発病無	合計
予防接種有	0.1581	0.4619	0.62
予防接種無	0.0969	0.2831	0.38
合計	0.255	0.745	1

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\varepsilon_{ij}^2}{p_{ij}}$$

$$= 200\left\{\frac{(0.11-0.1581)^2}{0.1581}+\cdots\right\}$$

$$= 10.338$$



$\chi_1^2(0.05) = 3.841$

$\chi_1^2(0.01) = 6.635$

結論　有意水準 $\alpha = 0.05$ または $\alpha = 0.01$ の
カイ2乗検定によって，H_0 は棄却される．

2 × 2 分割表

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

	属性あり	属性なし	合計
グループ1	a	b	n_1
グループ 2	c	d	n_2
合計			n

基本的な問題　グループ分けと属性の独立性　→ 独立性の検定

↕ 2× 2 の場合の特徴として同じことになる

各グループの母比率が一致する → 母比率の差の検定

例 題 11.2　（再考）

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

予防接種有, 予防接種無の発症率をそれぞれ p_1, p_2 とおく.

帰無仮説と対立仮説　$H_0: p_1 = p_2 = p$　$H_1: p_1 \neq p_2$

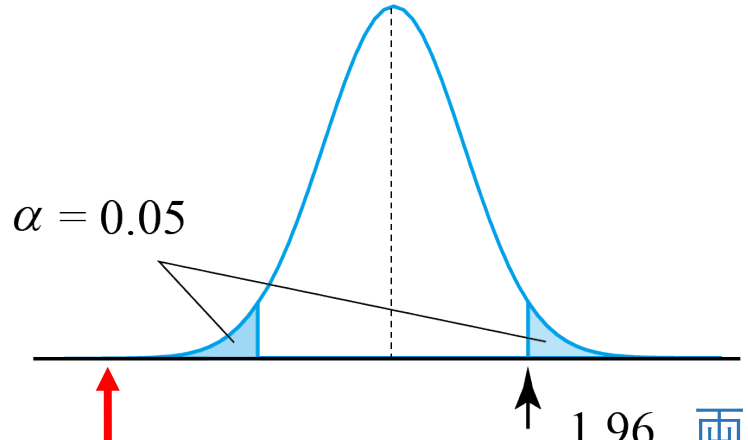
有意水準　$\alpha = 0.05$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

$p = \frac{22+29}{124+76} = 0.255$

$= N(0, 0.0635^2)$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{0.0635} \sim N(0,1)$$



両側 5% 点

実現値　$z = \frac{0.177 - 0.382}{0.0635} = -3.215$

結論　有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される.

例 題 11.2　（再考）

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

母比率の比較による検定

$z = -3.215 < -z(0.05) = -1.96$

有意水準 $\alpha = 0.05$ の両側検定によって
H_0 は棄却される.

ピアソンの χ^2- 値による検定

$\chi^2 = 10.338 > \chi_1^2(0.05) = 3.841$

有意水準 $\alpha = 0.05$ の両側検定に
よって H_0 は棄却される.

2 乗すると　$10.337 > 3.841$

一致している

2 × 2 分割表では, 厳密に

$$z^2 = \chi^2$$

したがって, 2 つの検定法は同値である.

例 題 11.3

ある疾患に対する免疫の有無を調べるため, A市とB市で標本調査を行って, 次の結果を得た. 免疫をもっ ている人の比率は両市で差があると考えられるか. カイ 2 乗検定を用いよ.

	免疫あり	免疫なし	計
A市	35 (0.35)	65 (0.65)	100 (1.00)
B市	60 (0.50)	60 (0.50)	120 (1.00)

【練習 10分】

例 題 11.3

	免疫あり	免疫なし	計
A市	35 (0.35)	65 (0.65)	100 (1.00)
B市	60 (0.50)	60 (0.50)	120 (1.00)

実測データ

	免疫あり	免疫なし	計
A市	0.159	0.295	0.454
B市	0.273	0.273	0.546
計	0.432	0.568	1

独立性を仮定した理論値 p_{ij}

	免疫あり	免疫なし	計
A市	0.196	0.258	0.454
B市	0.236	0.310	0.546
計	0.432	0.568	1

例 題 11.3

実測データ

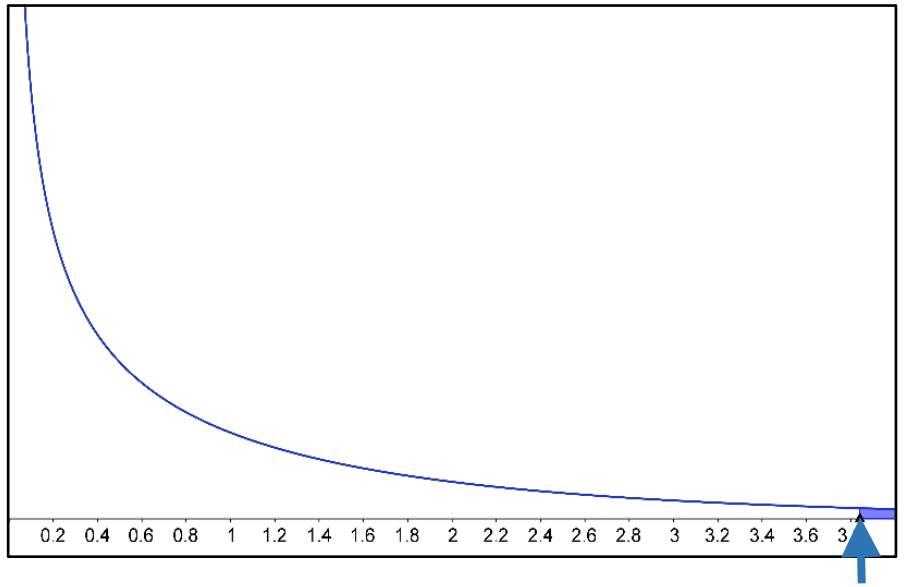
	免疫あり	免疫なし	計
A市	0.159	0.295	0.454
B市	0.273	0.273	0.546
計	0.432	0.568	1

H_0 (独立性)を仮定した理論値 p_{ij}

	免疫あり	免疫なし	計
A市	0.196	0.258	0.454
B市	0.236	0.310	0.546
計	0.432	0.568	1

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\varepsilon_{ij}^2}{p_{ij}}$$

$$= 220\left\{\frac{(0.159-0.196)^2}{0.196}+\cdots\right\} = 4.951$$



$\chi_1^2(0.05) = 3.841$

結論　有意水準 $\alpha = 0.05$ のカイ2乗検定によって H_0 は棄却される.

例 題 11.4　（サッカーのゴール数）1試合1チーム当たりのゴール数を調べた.

2013 年Jリーグ・ディビジョン 1 ・第34 節 18 チーム総当たり全306試合

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00

例 題 11.4　（サッカーのゴール数）1試合1チーム当たりのゴール数を調べた.

2013 年Jリーグ・ディビジョン 1 ・第34 節 18 チーム総当たり全306試合

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00

平均値 = 1.436

分散 = 1.367

⇒ ポアソン分布？

パラメータ $\lambda = 1.436$ のポアソン分布と比較

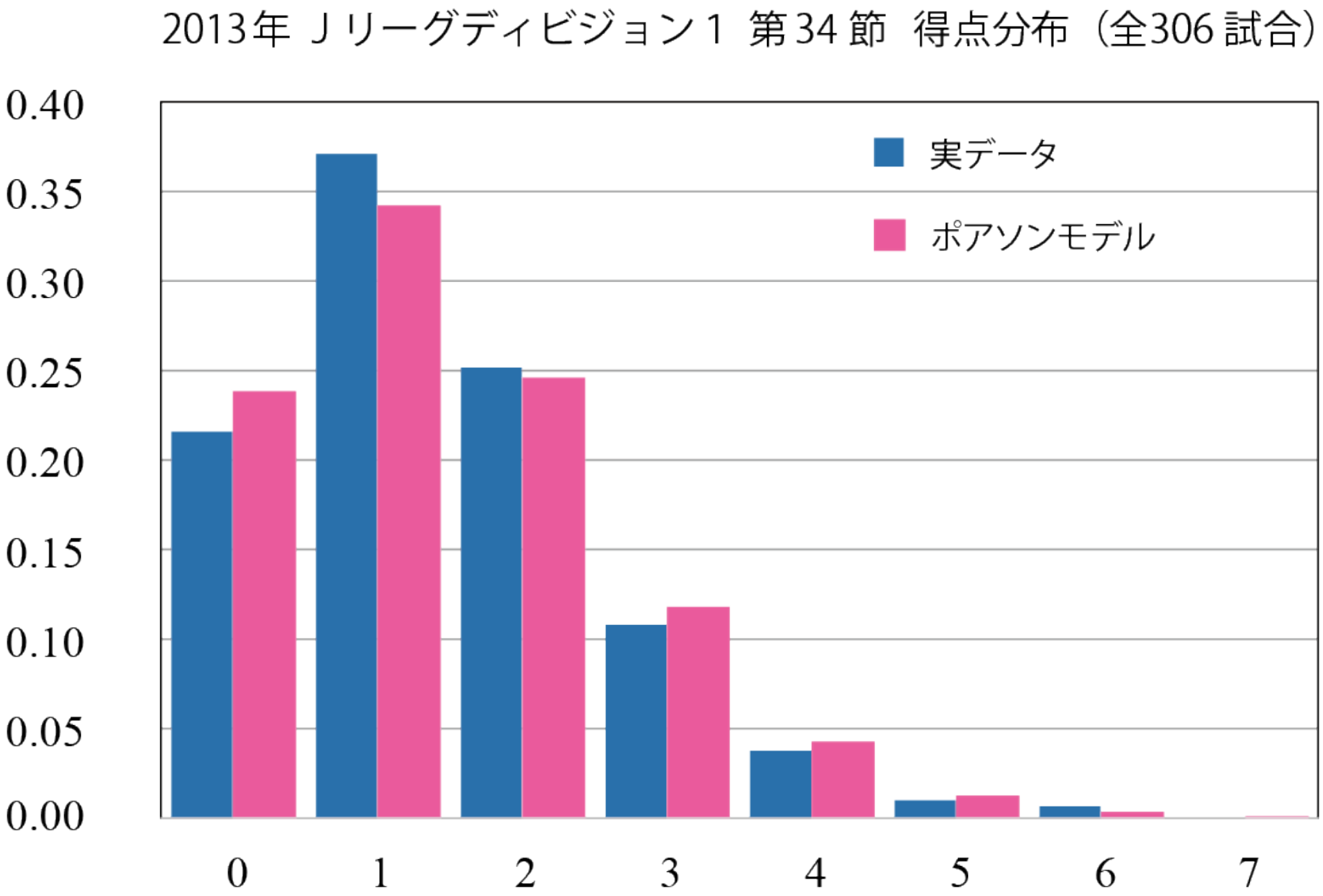
$$P(X = k) = \frac{1.436^k}{k!} e^{-1.436} \qquad k = 0,1,2,\cdots$$

例 題 11.4　パラメータ $\lambda = 1.436$ のポアソン分布と比較

$$P(X = k) = \frac{1.436^k}{k!} e^{-1.436} \qquad k = 0,1,2,\cdots$$

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00
ポアソン	0.2378	0.3416	0.2453	0.1174	0.0422	0.0121	0.0029	0.0006	0.9999
理論度数	145.54	209.04	150.12	71.87	25.81	7.41	1.77	0.36	611.92

例 題 11.4


2013年 Jリーグディビジョン1 第34節 得点分布（全306試合）
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
実データ
ポアソンモデル
0
1
2
3
4
5
6
7

例 題 11.4　χ^2- 検定

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
理論度数	145.54	209.04	150.12	71.87	25.81	7.41	1.77	0.36	611.92

⇩　理論度数 ≥ 5 となるように度数分布表を調整する

ゴール数	0	1	2	3	4	5以上	合計
試合数 X_i	132	227	154	66	23	10	612
理論度数 m_i	145.54	209.04	150.12	71.87	25.81	9.62	612

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 3.703$$

例 題 11.4　χ^2- 検定

ゴール数	0	1	2	3	4	5以上	合計
試合数 X_i	132	227	154	66	23	10	612
理論度数 m_i	145.54	209.04	150.12	71.87	25.81	9.62	612

H_0：ポアソン分布に従う

H_1：ポアソン分布に従わない

有意水準 $\alpha = 0.05$

検定統計量

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i}$$

※ ポアソン分布の特殊性から，χ^2-値は自由度

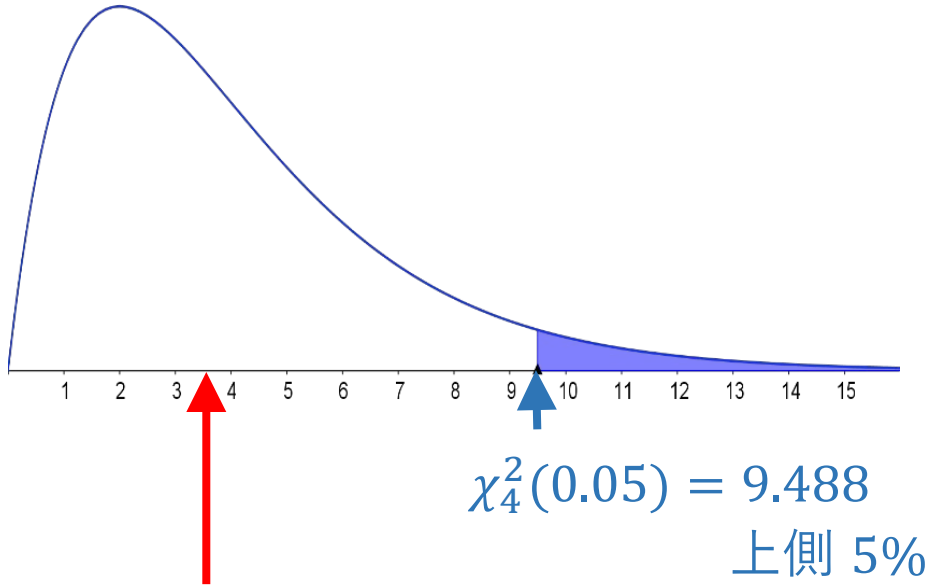
$6 - 1 - 1 = 4$ のカイ2乗分布 χ_4^2 -分布に従う．

例 題 11.4　χ^2- 検定

H_0：ポアソン分布に従う

H_1：ポアソン分布に従わない

有意水準 $\alpha = 0.05$



検定統計量

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i}$$

※ ポアソン分布の特殊性から，

χ^2-値は自由度 $6 - 1 - 1 = 4$

のカイ2乗分布 χ_4^2 -分布に従う．

実現値

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 3.703$$

結論

有意水準 $\alpha = 0.05$ のカイ2乗検定によって，
H_0 は棄却されない．

ちなみに，$P = 0.4467$