P 値について

森本 孝之（関西学院大学 理工学部 数理科学科）

2021 年度研究室紹介特設ページ用資料

2020 年 12 月 1 日

目次 (1)

目次（2）

目次（3）

P 値とは何だろうか（はじめに）1

ある薬剤を服用し血圧が下がったとき，この薬剤の効果で下がったのか?

- P 値（ピーチ）は，大雑把にいえば「血圧が下がったのは，薬剤の効果でありデータのバラツキのためではない」ことを知るためのツールである．
- 論文などでは小文字の p を用いて p-value，あるいは単に p と表されることが多いが，本講演では P 値と表記した．

★ P 値は，データにバラツキがある場合，薬剤の効果が無いと考えて，バラツキだけの原因で評価指標が観測された値以上の大きな値をとる確率

- この確率の値が 20 回中 1 回，つまり 5% 以下なら，バラツキだけの原因で評価指標がこのような大きな値をとる可能性は小さい．
- したがって，薬剤の効果があったと推論する．

P 値とは何だろうか（はじめに）2

P 値 $< \mathbf{0.05}$ のとき，統計的に薬剤の効果ありとする
P 値 $\geq \mathbf{0.05}$ のとき，薬剤の効果ありとは統計的にいえないとする

- ここで「統計的に薬剤の効果あり」はあくまで「統計的」な判定であり，「薬剤の効果があった」という医学的判定とは異なることに注意してほしい．
- 効果ありの「統計的」判定が得られれば，例えば薬剤製造販売の許認可を得る際に実施された臨床試験等の成績を調べ，他の種々の情報を確認した上で医学的判断が下される．
- 薬剤の効果ありの「統計的」判定が得られなければ，医学的判断の対象にはなりえない．
- 上で用いた用語「推論」はこのような意味を含んでいる．

P 値とは何だろうか（はじめに）3

- 一般に P 値は，データにバラツキがあるあらゆる科学の分野において統計的推論を行うためのツールとして頻繁に使われている．
- P 値は，医系の分野（看護・リハビリ・福祉・医学の分野など）の報告書や学術論文の中に溢れている．
- 日々データと接しデータと取り組むこれらの分野の人々は，データの不確実性にもろに直面し，不確実なデータから結論を導かざるを得ない帰納的推論の世界に身を置いている．
- 数理統計学の教科書では，P 値は有意確率という名前で，高々 1 行の数式で定義されるだけで，節を設けて P 値の解説をしている書籍は皆無に近い．
- 演繹的論理の世界では P 値は，それだけの話題でしかない．
- 医系分野の統計学の教科書でも，その影響を受けて，ページを費やして P 値の解説をしているテキストは多くない．

P 値とは何だろうか（はじめに）4

帰納的推論の世界 $\Longleftrightarrow$ 演繹的論理の世界

- このギャップが，P 値の様々な誤用と誤解をばらまいている．
- その規模は世界的であり 2016 年にはアメリカ合衆国統計協会が声明書を出し P 値の間違った理解や誤用に対して注意を喚起した．
- 2017 年度統計関連学会連合大会で日本計量生物学会，日本計算機学会の 2 学会が特別集会を設置して P 値の誤用や誤解に対して警鐘を鳴らした．
- 誤用とは気づかずに P 値が適用されている事例は結構多い．
- 「私の P 値の理解や使い方は正しいのか」と自問をしていただく必要がある．
- P 値を正しく理解して適用するためには，まずデータにはバラツキがあること，データのバラツキを確率法則として認識することが必要である．

予知能力の存在が示された!?　(1)

- ベム博士はコーネル大学の名誉教授であり，著名な社会心理学者です．
- 2011 年，JPSP という社会心理学の最も権威ある学術雑誌に，ベムは突然，「未来の予感」と題した論文を発表します．
- 奇妙でロマンチックなタイトルの「未来の予感」には，多くの実験と分析が掲載されていましたが，特に第 1 実験が注目すべきものでした．

ベムの予知能力実験（これを 1 試行とする）

- コンピュータ画面に，2 つのカーテンを並べて表示する．
- 「一方のカーテンの裏には異性のヌード写真が隠され，他方にはありません．写真のあると思うカーテンをクリックしてください」と被験者に教示する．
- ヌード写真があると感じるカーテンを被験者がクリックした後に，乱数によって写真を一方の裏にだけセットする．
- カーテンを開き，写真があるか否かを確認する．

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

予知能力の存在が示された!?　(2)

- 1560 試行中 829 試行でヌード写真の位置が的中した．
- ベムはこのデータを有意性検定と呼ばれる統計手法で分析し，統計的に有意であることを根拠に JPSP へ「未来の予感」の公刊を申請しました．
- 学術雑誌では，通常，その論文に学術価値があるか否かを判定するために，査読と呼ばれる審査を行います．
- 実験や調査を利用した多くの学術論文は，これまで統計学における有意性検定の結果が有意であることを根拠に公刊されてきました．
- 有意性検定では，通常 P 値と呼ばれる確率が 5% を下回ったときに「統計的に有意差あり」と判定します．
- 多くの研究分野において，分析目的に関係なく，適用対象を問わずに「P 値が 0.05 を切ったか否か」を価値判断の基準として利用してきました．
- ベムの予知能力実験の P 値は 1.3% であり，$\mathbf{P < 0.05}$ でした．

予知能力の存在が示された!? (3)

- JPSP の審査委員は困ってしまいました.
- 「予知能力が存在する」などというトンデモな論文を公刊したら, 権威ある学術雑誌の名誉に傷がつきます.
- さりとて $\mathbf{P < 0.05}$ という条件をクリアしているのにベムの論文だけを不採択にすれば, 公正で公平な扱いとはいえません.
- 迷った末に, 審査委員はベムの主張を認め, ついに「未来の予感」は JPSP から公刊されました.
- ベムの論文は大きな反響を呼び, 各方面から激しい批判を受けました.
- もちろん心理学者達は, ただちに追試実験を開始しました.
- しかし実験結果を誰も再現できませんでした.
- 統計的には有意でも, 心理学的には無意味な論文だったのです.

検定の目的（1）

- 「未来の予感」が権威ある社会心理学の学術誌に掲載された理由は，有意性検定が要求する $\mathbf{P < 0.05}$ という条件を満たしていたからです．
- ここではベムが用いた二項検定を例に，有意性検定の考え方を説明します．
- 統計学では，確率的に結果の変わる試みを試行といいます．
- 仮にヌード写真の位置を予知する試行を 10 回実施したとしましょう．
- 試行の結果，観察されうる状態（的中とハズレ）を事象といいます．
- たとえば，写真の位置を 7 回的中させたとします．
- 試行の回数を試行数 n と呼び，的中回数を成功数 x と呼びましょう．
- このとき成功数を試行数で割った値を標本比率といいます．

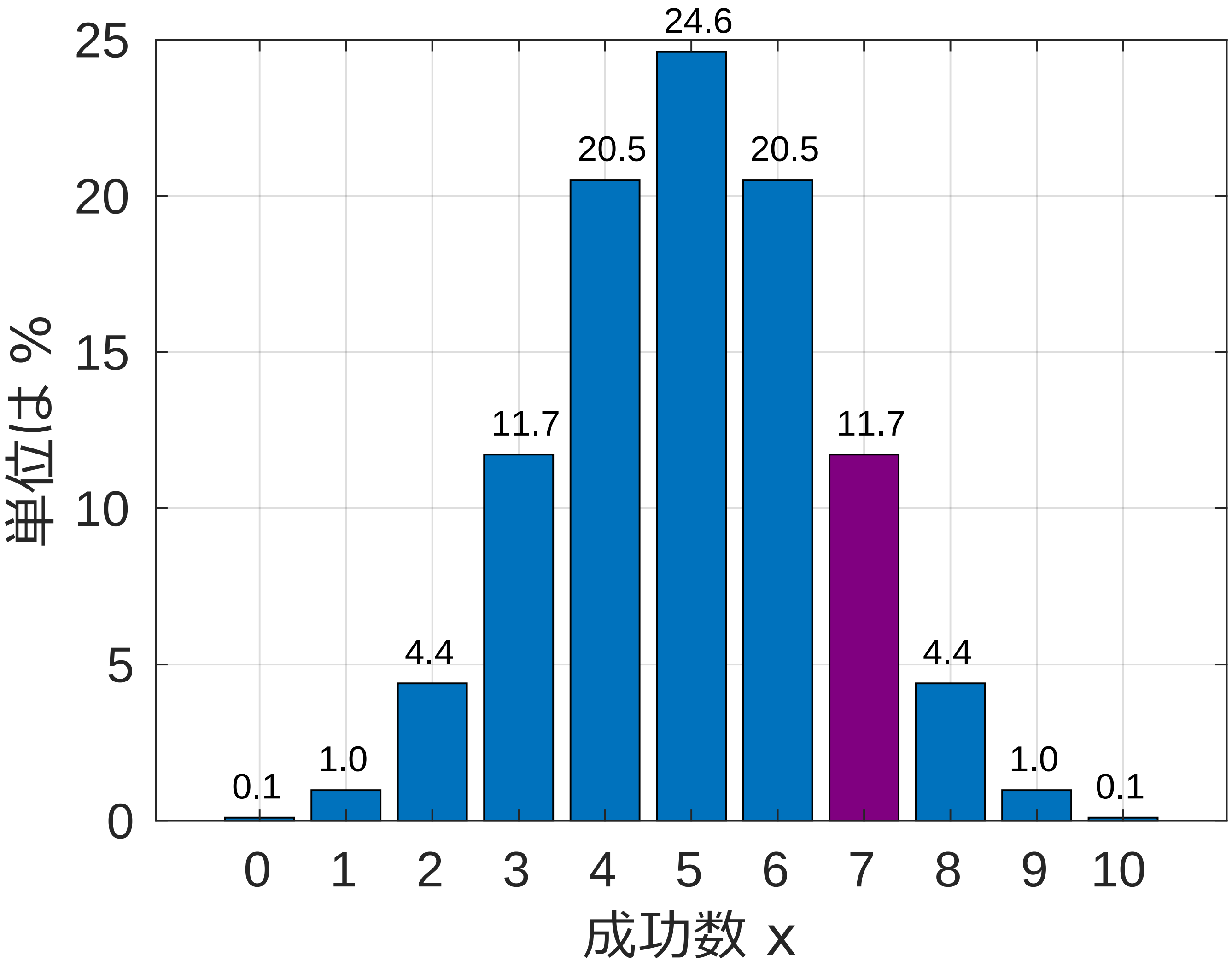
$$\text{標本比率} = \frac{\text{成功数}}{\text{試行数}} = \frac{7}{10} = 0.7 \tag{1}$$

検定の目的（2）

- 標本比率はデータから計算された比率です．
- 70% の確率でヌード写真の位置を的中できれば，それを予知能力だと考える人も，世の中にはほんの少しはいます．
- では，この標本比率は予知能力の存在を示しているでしょうか．
- いいえ，必ずしも示しているとは限らないのです．
- なぜならば，歪みのない正確なコインを投げても，10 回中 7 回表が出ることは珍しくないからです．
- コイントスのように，結果の事象が 2 つ（表か裏）のみの試行において，表が出る回数の確率は 2 項分布と呼ばれる確率分布で示されます．
- 確率分布とは，どの値がどのくらいの確率で観察されるかの様子です．
- 確率分布は単に分布と呼ばれることもあります．

検定の目的（3）

- 下図によればコイントスで 10 回中 7 回表が出る確率は 11.7% です．
- 追試実験をすれば，10 回に 1 回以上は観察できるのです．



検定の目的（4）

- 今回は偶々ヌードの位置を 7 回的中させたけれども，コイントスと同じで，次はそんなに当たらないかもしれません．
- 標本比率からの考察だけでは「十分高いから予知能力はある！」「いや偶然だ，コイントスと同じかもしれない」という水掛け論が始まってしまいます．
- 学術的に重要なことは現象の再現性，つまり，同じ状況下で無数の追試実験を行ったと想定したときの平均的な比率が，十分に大きいことが大切です．
- もしそれが示されれば，誰の目にも予知能力が存在することは明らかです．
- 平均的な比率は，目の前のデータから計算された標本比率とは異なりますから，区別するために母比率 π（パイ）と別の名前で呼びましょう．
- コイントスで表が出る事象の母比率は $\pi = 0.5$ です．
- 「超能力の存在を示す」ためには，少なくとも，ヌード写真の位置を的中させる母比率が $\pi = 0.5$ でないことを示す必要があります．

検定の手続き（1）

- 学術的興味の対象が，標本比率ではなく母比率 π にあるとしても，追試実験を無数に行うことは現実的には不可能です．
- そこで有意性検定の一種である二項検定の登場です．
- 二項検定は，1 回の実験データから（標本比率ではなく）母比率に関する一定の判断を示すための方法です．

有意性検定の流れ

Step1:　帰無仮説 H_0 と対立仮説 H_1 を設定
Step2:　H_0 を真として検定統計量を計算
Step3:　標本分布から P 値を計算
Step4:　帰無仮説 H_0 を棄却または採択

- この手続きは，二項検定に限らず，多くの有意性検定に共通しています．

検定の手続き（2）

- 前頁の「有意性検定の流れ」は専門用語ばかりなので，詳しく説明します．

Step 1

- 帰無仮説 H_0 と対立仮説 H_1 を設定します．

$$H_0 : \pi = 0.5 \tag{2}$$

- 帰無仮説は「ヌード写真の位置が当たる母比率は，コイントスと同じ 0.5 で，まったく偶然だ」という意味です．
- 分析者ベムが，無に帰したい（棄却したい）仮説です．
- 帰無仮説の否定は，予知能力存在の必要条件です．
- 対立仮説は帰無仮説の否定，つまり，$H_1 : \pi \neq 0.5$ とします．
- 対立仮説は，分析者が最終的に採択したい仮説です．

検定統計量

Step2

- H_0 を真として検定統計量を計算します．
- 統計量とはデータの関数です．

$$\text{統計量} = f(\text{データ}) \tag{3}$$

- 成功数 x や標本比率は，データから計算されますから，統計量です．
- 検定統計量とは，検定で使用する統計量のことです．
- 二項検定では成功数 x か標本比率を検定統計量として利用できます．
- (1) 式の例では，成功数は $x = 7$ で，標本比率は 0.7 でした．
- どちらを検定統計量として使っても，検定結果は同じです．

標本分布（1）

Step3

- 標本分布から P 値を計算します．
- 標本分布は統計量（たとえば成功数 x）の分布です．
- 帰無仮説（2）式が真であると仮定し，試行数 n の追試実験を無数に繰り返すことを想像します．
- さらにその実験ごとに成功数 x を調べたと想像してください．
- このときの成功数の分布を x の標本分布といいます．
- 結果は成功か失敗 2 通りなので，成功数 x の標本分布は 2 項分布です．
- 追試実験を無数に繰り返すことは実際にはしません．
- 検定は 1 回の実験データに適用されます．

標本分布（2）

- 標本分布は理論上の概念であり，実際には観測されません．
- 想像上の実在しない分布です．
- 標本分布は，標本の分布ではありませんから，注意してください．
- 2 項分布は一般的に次のように表記します．

$$2\text{ 項分布}_n(x|\pi) \tag{4}$$

- かっこの中は，エックス／ギブン／パイと読みます．
- 母比率 π が与えられた条件下での成功数 x の確率分布という意味です．
- 「**10** 回中 **7** 回表が出る確率は 11.7% である」という代わりに

$$2\text{ 項分布}_{n=10}(x=7|\pi=0.5)=0.117 \tag{5}$$

と表記することもできます．

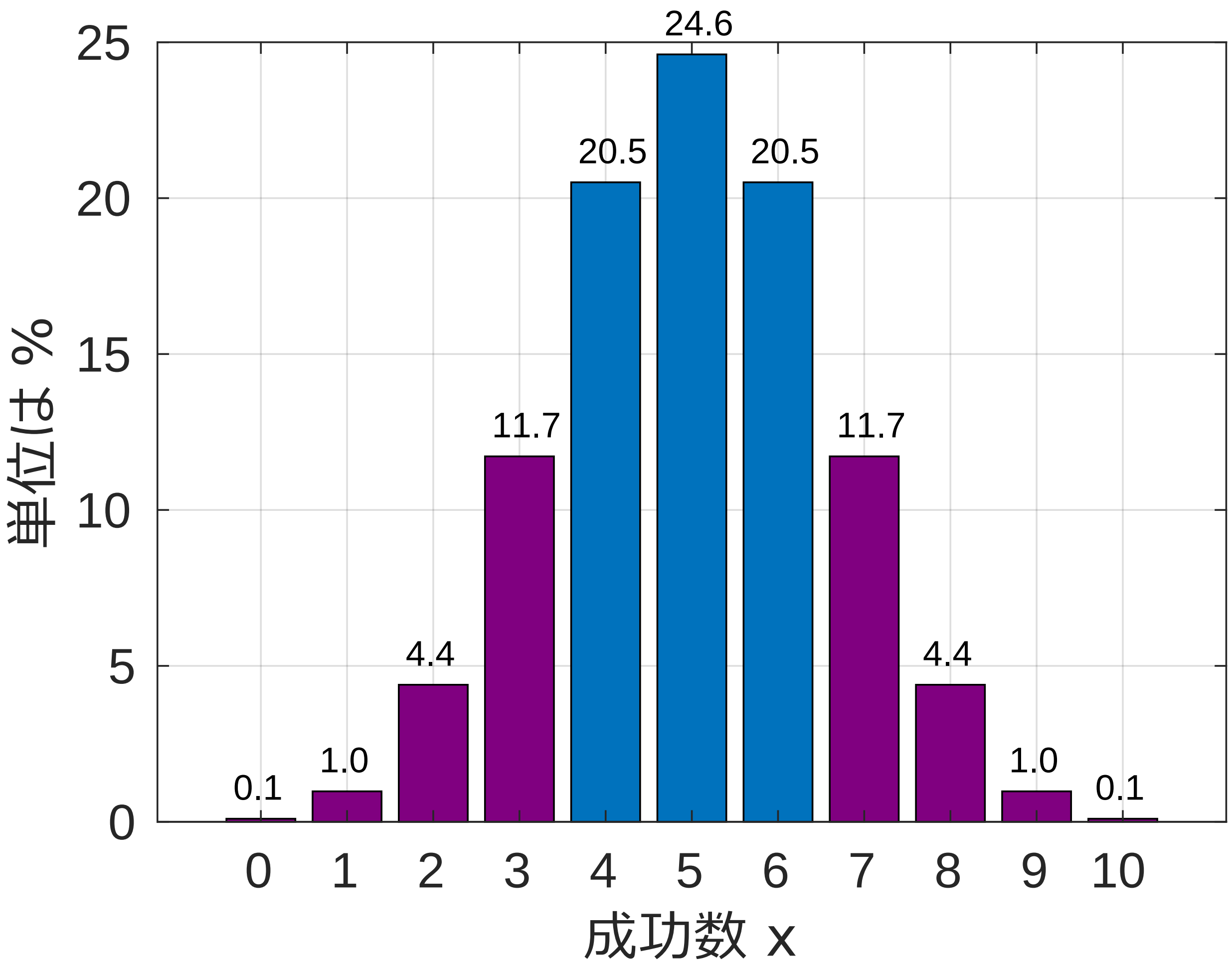
P 値（1）

- 検定統計量 x を計算したら，x の確率的な評価をします．
- そのときに計算されるのが P 値です．
- P 値とは，帰無仮説が真であるという仮定の下で，検定統計量が標本分布においてデータから計算された値以上に甚だしい値となる確率です．
- データから計算された検定統計量ピタリの値になる確率ではありませんから注意してください．
- くどいようですが，成功数 $\mathbf{7}$ そのものが観察される確率ではありません．
- 帰無仮説が真である状態で，観察された成功数 $\mathbf{7}$ 以上に甚だしい事象が生じる確率が P 値です．
- 成功数 $\mathbf{5}$ を基準に $\mathbf{7}$ 以上に甚だしい値は $\mathbf{3}$ 以下の領域も含みます．
- 具体的には $x = \{0, 1, 2, 3, 7, 8, 9, 10\}$ が観察される事象の確率が P 値です．

P 値 (2)

- 下図から該当する柱の確率を拾うと，$\mathbf{P = 0.344}$ となりました．

$$(11.7 + 4.4 + 1.0 + 0.1) \times 2 = 34.4\% \tag{6}$$



- なぜ P 値は, データから計算された検定統計量そのものが生じる確率ではないのでしょうか.
- それはその確率が高いのか低いのかを評価することが難しいからです.

P 値 (3)

- 試行数 **100** 回，的中数 **70** 回という実験データを考えてみましょう．
- 標本比率は変化せずに **0.7** のままです．
- しかし 2 項分布で確率を評価すると

$$2\ \text{項分布}_{n=100}(x = 70|\pi = 0.5) = 0.00002 \tag{7}$$

となり，ぴったり $x = 70$ が観察される確率は，たった **0.00002** です．

- 試行数 100 の場合は x は 0 から 100 まで 101 通りの可能性があるので，一つ一つは小さな確率になります．
- データから計算した検定統計量そのものが観察される確率は，n の大きさに強く依存してしまいます．
- このため「起きやすい事象なのか，起きにくい事象なのか」を判断するための確率としては，相応しくないのです．

P 値 (4)

- 言い換えるならば，試行結果の種類数が 11 と 101 のように元々違うのだから，標本比率 0.7 そのものが観察される確率は，評価の指標として適切ではないということです．

- 試行数 100 のときは，実験で観察された的中数 70 以上に甚だしい事象が生じる確率が P 値です．

- $n = 100$ の 2 項分布には 101 本の柱があり，該当する柱の確率を拾って合計を求めると，P 値は

$$\mathrm{P}(0 \leq x \leq 30 \quad \text{または} \quad 70 \leq x \leq 100) = 0.000079 \tag{8}$$

となります．

- ここで $\mathrm{P}()$ はかっこ内の事象が生じる確率を表す記号とします．

- 標本比率が 0.7 以上に甚だしい値が観察される確率は 0.000079 と言い換えても構いません．

P 値 (5)

- 試行数 10 の場合には，標本比率が 0.7 より甚だしくなる確率は 34.4% もありました．
- それに対して試行数 100 の場合には同じ事象が生起する確率は 0.0079% ときわめて小さくなります．
- ピタリの確率ばかりでなく，観測値以上に甚だしい確率を求めても，同じ標本比率なら，試行数 100 の場合の方が P 値は小さくなるということです．
- 言い換えるならば，コイントスを 10 回試みるくらいなら標本比率が 0.7 より甚だしくなる現象はよく起きます．
- それに対して 100 回も投げたら標本比率が 0.7 より甚だしくなる現象はまず起きないということです．
- 計算された検定統計量よりも甚だしくなる確率である P 値を用いると，このような実情を妥当に反映し確率の高低を正しく比較できるようになります．

有意水準（1）

Step4

- 帰無仮説 H_0 を棄却または採択します．

> 有意性検定では，P 値が小さくて起きにくい事実が観察されたら，帰無仮説が真であるという仮定が正しくて，かつ起きにくい値が観察されたのではなく，そもそも「帰無仮説は偽」であったのだろうと判断します．

- これを帰無仮説の棄却といい，この場合は対立仮説を採択します．
- ただし「起きにくい」という表現は，文学的かつ主観的です．
- そこで結果を見る前に，予め確率で定義しておきます．
- 起きにくさとして予め定めた確率を有意水準と言い，α と表記します．
- 有意水準には一般的に 0.05 が用いられます．

有意水準（2）

- なぜ有意水準は 5% なのかと問われても理由はありません．
- あらゆる研究分野において，分析目的に関係なく，適用対象を問わずに，また結果の重要度に係わらず，頻繁に利用されるのは $\alpha = 0.05$ です．
- 試行数 $n = 10$ の実験で，標本比率 0.7 の場合 $\mathrm{P} = 0.344 > 0.05$ です．
- 帰無仮説が真であるという仮定の下で起きにくい値は観察されていません．
- この場合「帰無仮説が偽」であるという証拠は得られなかったと判断します．
- これを帰無仮説の採択といい，この場合は対立仮説を棄却します．
- 要するに 10 回中 7 回も当たることは「コイントスでも起こり得る」という判断になります．
- $n = 100$ の実験で，標本比率 0.7 の場合は $\mathrm{P} = 0.000079 < 0.05$ です．

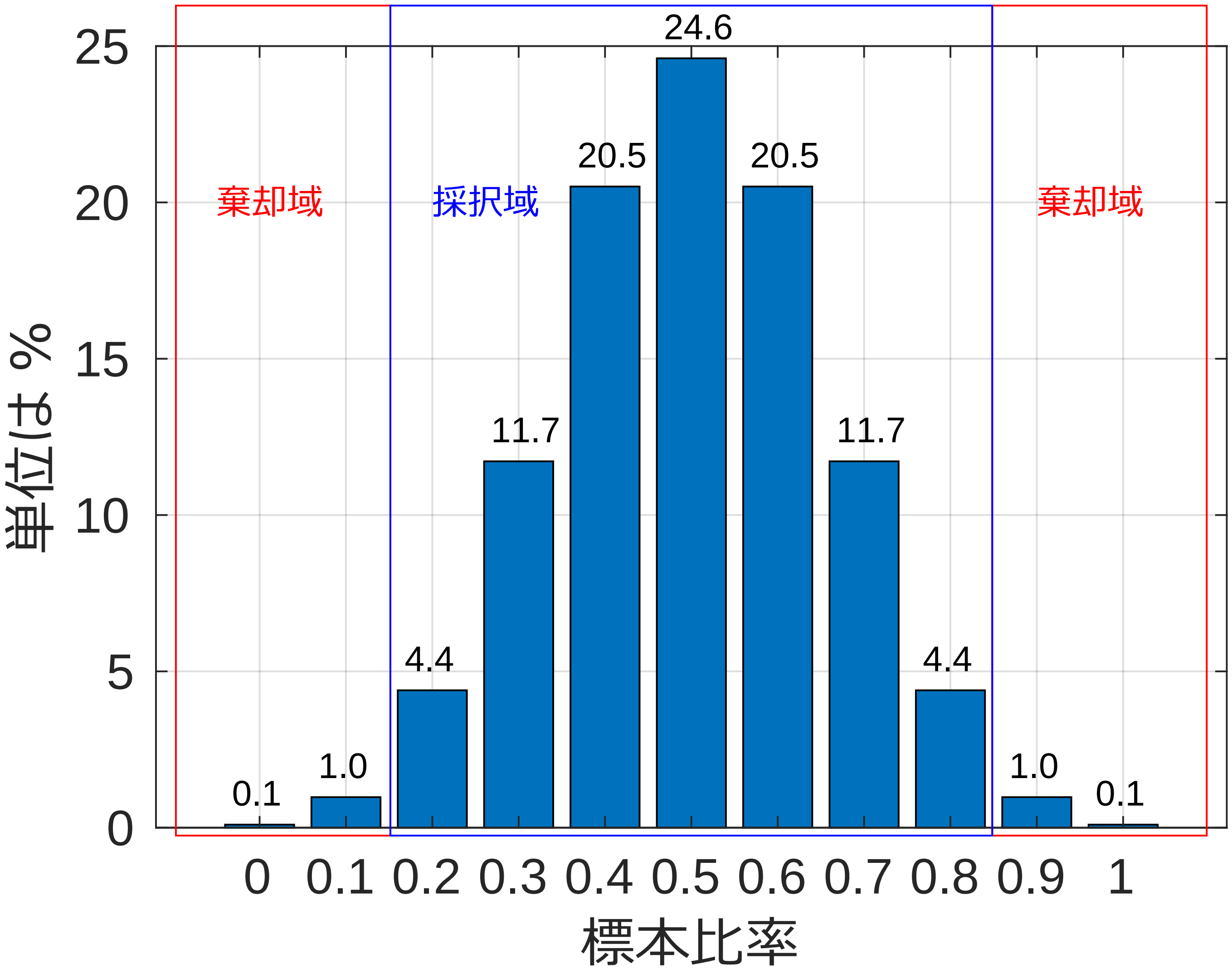
有意水準（3）

帰無仮説が真であるという仮定が正しくて，かつ起きにくい値が観察されたのではなく，そもそも「帰無仮説は偽」であったのだろうと判断します．

- 帰無仮説を棄却し対立仮説を採択します．
- この状態を「有意差がある」とか，単に「有意である」などといいます．
- 要するに「100 回中 70 回も当たるのは，コイントスのような試行とは違う」という判断です．
- 母比率 $\pi = 0.5$ という帰無仮説に対して，標本比率 0.7 は $n = 10$ では，5% 水準で統計的に有意ではありません．
- $n = 100$ では，5% 水準で統計的に有意です．

棄却域と採択域（1）

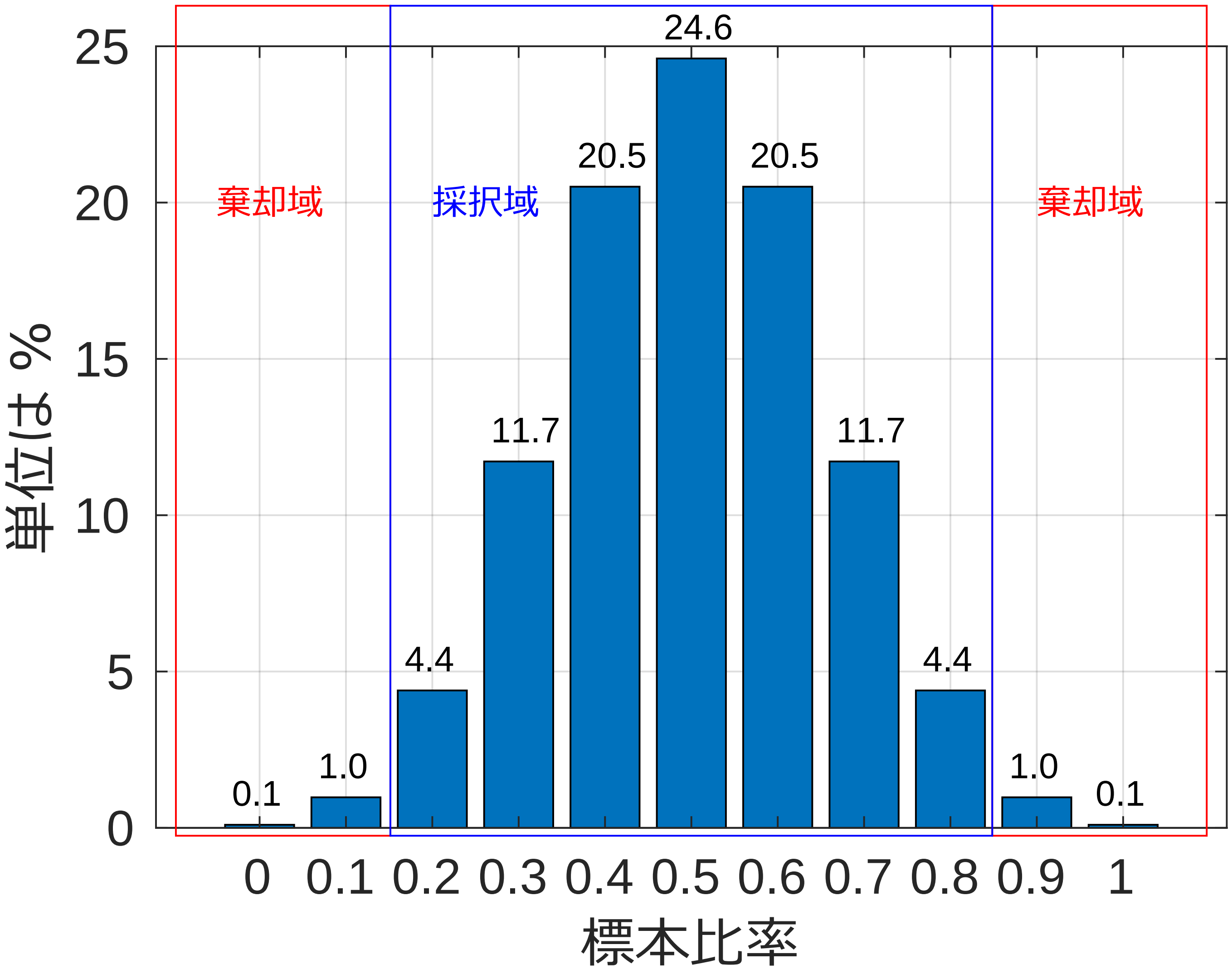
- 試行数 $n = 10$ の場合は，どこまでの甚だしい標本比率なら P 値は 5% 以内に収まるのでしょうか．



- 左図に $n = 10, \pi = 0.5$ のときの標本比率の標本分布を示しました．
- 後ほど試行数 $n = 100$ の場合と比較し易くするために，横軸を成功数 x から標本比率に変更してあります．

棄却域と採択域 (2)

- 標本比率が **1.0** または **0.0** の場合は，P 値は 0.2% (= **0.1 + 0.1**) だから，**P < 0.05** で有意です．

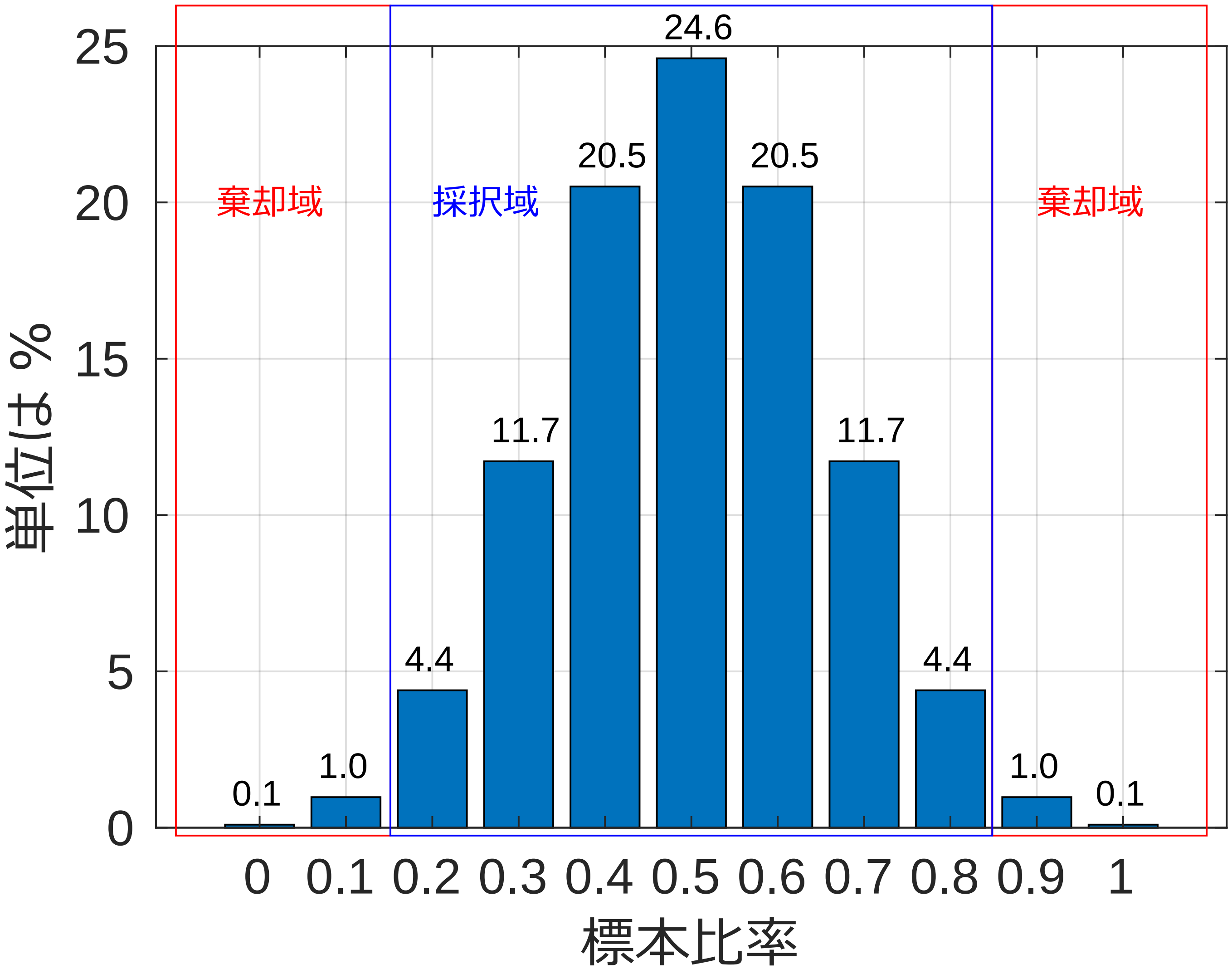


- 標本比率が **0.9** または **0.1** の場合は，P 値は 2.2% (= **(1.0 + 0.1) × 2**) だから，**P < 0.05** で有意です．

- 標本比率が **0.8** または **0.2** の場合は，P 値は 11.0% (= **(4.4 + 1.0 + 0.1) × 2**) だから，もう **P > 0.05** で有意ではありません．

棄却域と採択域（3）

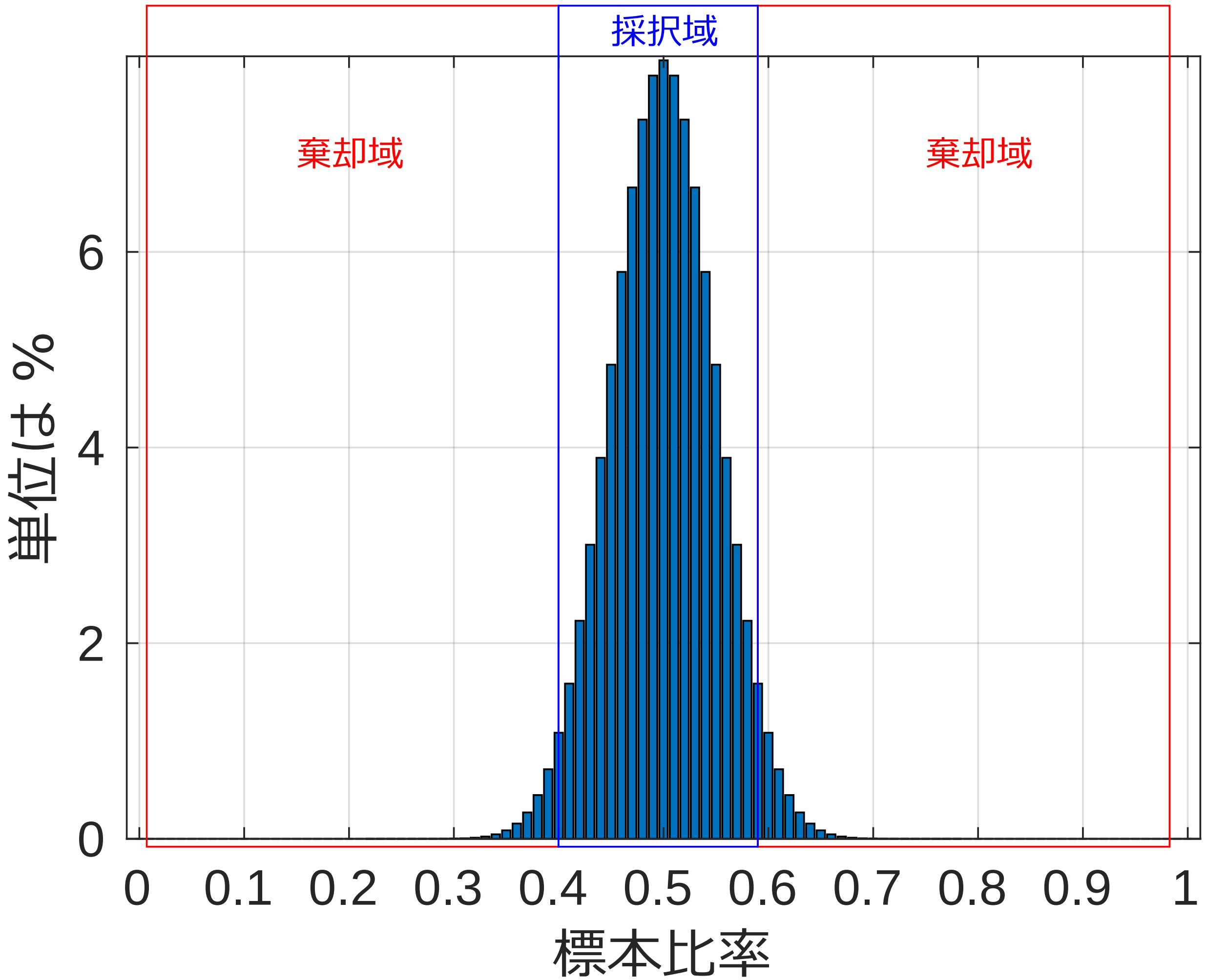
- 統計的に有意である場合には帰無仮説を棄却し，有意でない場合には帰無仮説を採択するのでした．



- 標本分布において，検定結果が有意になる検定統計量の領域を棄却域と呼び，**[0.0, 0.1]**, **[0.9, 1.0]** が棄却域です．
- 標本分布において，検定結果が有意にならない検定統計量の領域を採択域と呼び，**[0.2, 0.8]** が採択域です．

棄却域と採択域（4）

- 試行数 $n = 100$ のときの標本比率の標本分布を下図に示します．

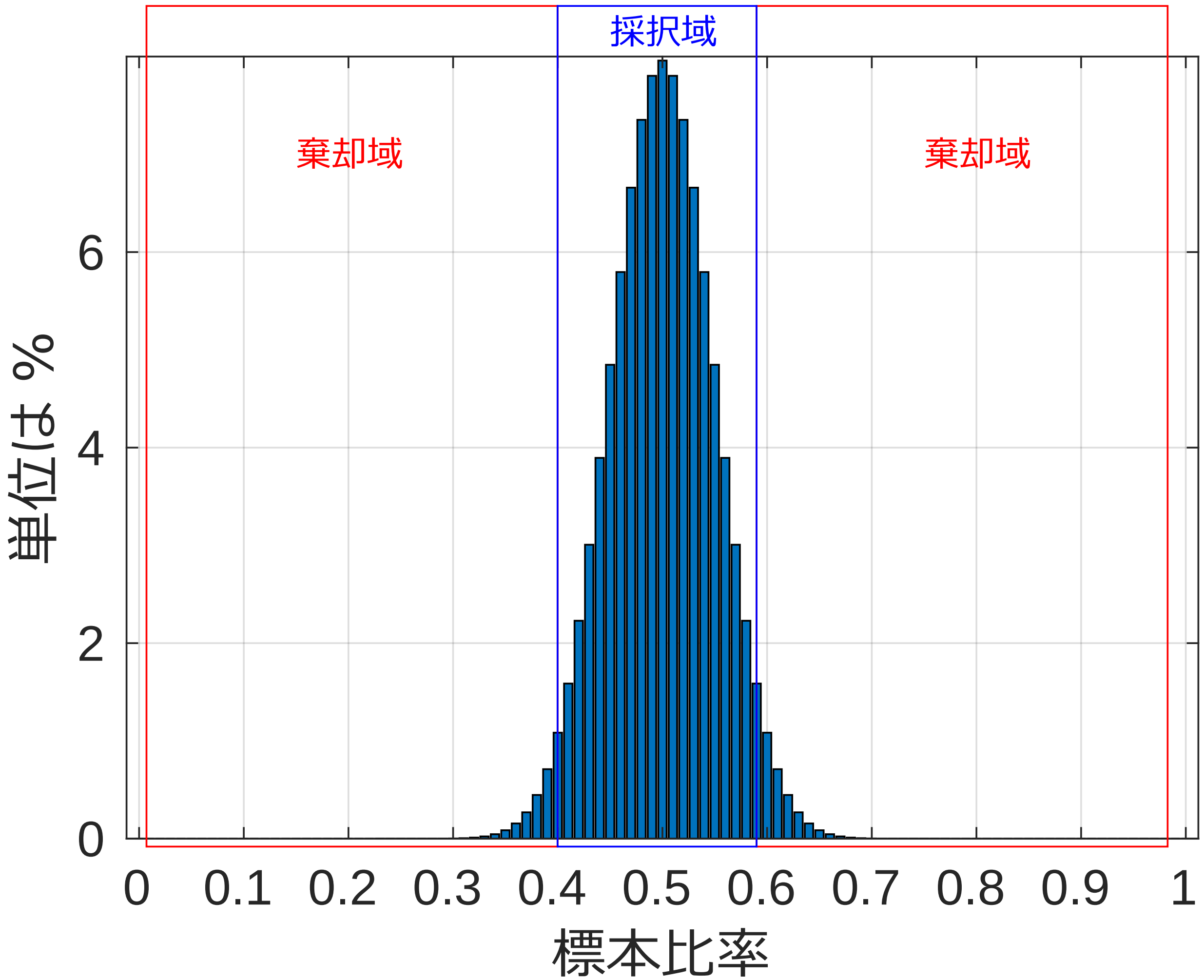


- 試行数 $n = 10$ のときは 11 本の柱がありましたが，試行数 $n = 100$ のときは 101 本の柱になりました．
- 左右の端のほうは，柱がないように見えますが，これは低すぎて印刷が見えないだけで，ちゃんと 0.00 から 1.00 まで柱は存在しています．

- 試行数 $n = 10$ と比較したときの，試行数 $n = 100$ の標本分布の一番の特徴は，分布の形状が中央部分にギュッと圧縮されていることです．

棄却域と採択域（5）

- 棄却域は $\{0.00 \leq$ 標本比率 ≤ 0.39 および $0.60 \leq$ 標本比率 $\leq 1.00\}$
- 採択域は $\{0.40 \leq$ 標本比率 $\leq 0.59\}$



- $n = 10$ の場合は，採択域の幅が 0.6 $(= 0.8 - 0.2)$ です．
- $n = 100$ の場合は，採択域の幅が 0.19 $(= 0.59 - 0.40)$ です．
- 一般的に有意水準と標本比率が同じなら，n の増加に伴って採択域は狭まり，逆に棄却域は広がります．

信頼区間による区間推定（1）

- 標本比率と母比率は，互いに異なった概念です．
- 標本比率は，母比率の推定値として利用することができます．
- 標本比率のような 1 点の値で母数を推定する方法を点推定といいます．
- 母数（母比率 π は母数の一種）は確率分布の特徴を決める数的指標です．
- それに対して母数を区間で推定する方法を区間推定といいます．
- 以下に，母比率 π を区間で推定する方法を説明します．
- 有意性検定の Step1 において，帰無仮説を

$$H_0 : \pi = 0.0 \sim 1.0 \tag{9}$$

のように，母比率が定義されている範囲で連続的に動かします．

信頼区間による区間推定（2）

- それぞれで有意水準 α の検定を実施し，帰無仮説が棄却されない下限 π_L と，上限 π_H とを調べます．
- このとき区間 (π_L, π_H) を $((1-\alpha)\times 100)\%$ 信頼区間といいます．
- $\alpha = 0.05$ の場合 95% 信頼区間，$\alpha = 0.01$ の場合 99% 信頼区間です．
- 原理的には有意水準 α の値は，どの値でもよいのですが，伝統的には 95% 信頼区間が最も頻繁に利用されます．
- 試行数 $n = 10$，成功数 $x = 7$ のとき 95% 信頼区間は $[0.348, 0.933]$ となり，幅は 0.585 です．
- 試行数 $n = 100$，成功数 $x = 70$ のとき 95% 信頼区間は $[0.600, 0.788]$ となり，幅は 0.188 です．

★ 標本比率と有意水準が同じならば，試行数が大きい方が信頼区間は狭い．

「未来の予感」の検定結果（1）

- 二項検定の仕組みを利用して，いよいよベムの「未来の予感」の有意性検定を再分析してみましょう．
- ヌード写真の位置は 1560 試行中 829 回的中しました．
- Step1 では帰無仮説 H_0 と対立仮説 H_1 を設定します．

$$H_0 : \pi = 0.5, \quad H_1 : \pi \neq 0.5$$

- Step2 では H_0 を真として検定統計量を計算します．
- 二項検定では成功数，標本比率はどちらでも同じ結果を与える検定統計量として利用できました．
- 成功数は $x = 829$，標本比率は 0.531（$= 829/1560$）です．
- 半分の試行数は 780（$= 1560/2$）ですから，49 回（$= 829 - 780$）余分に当てています．

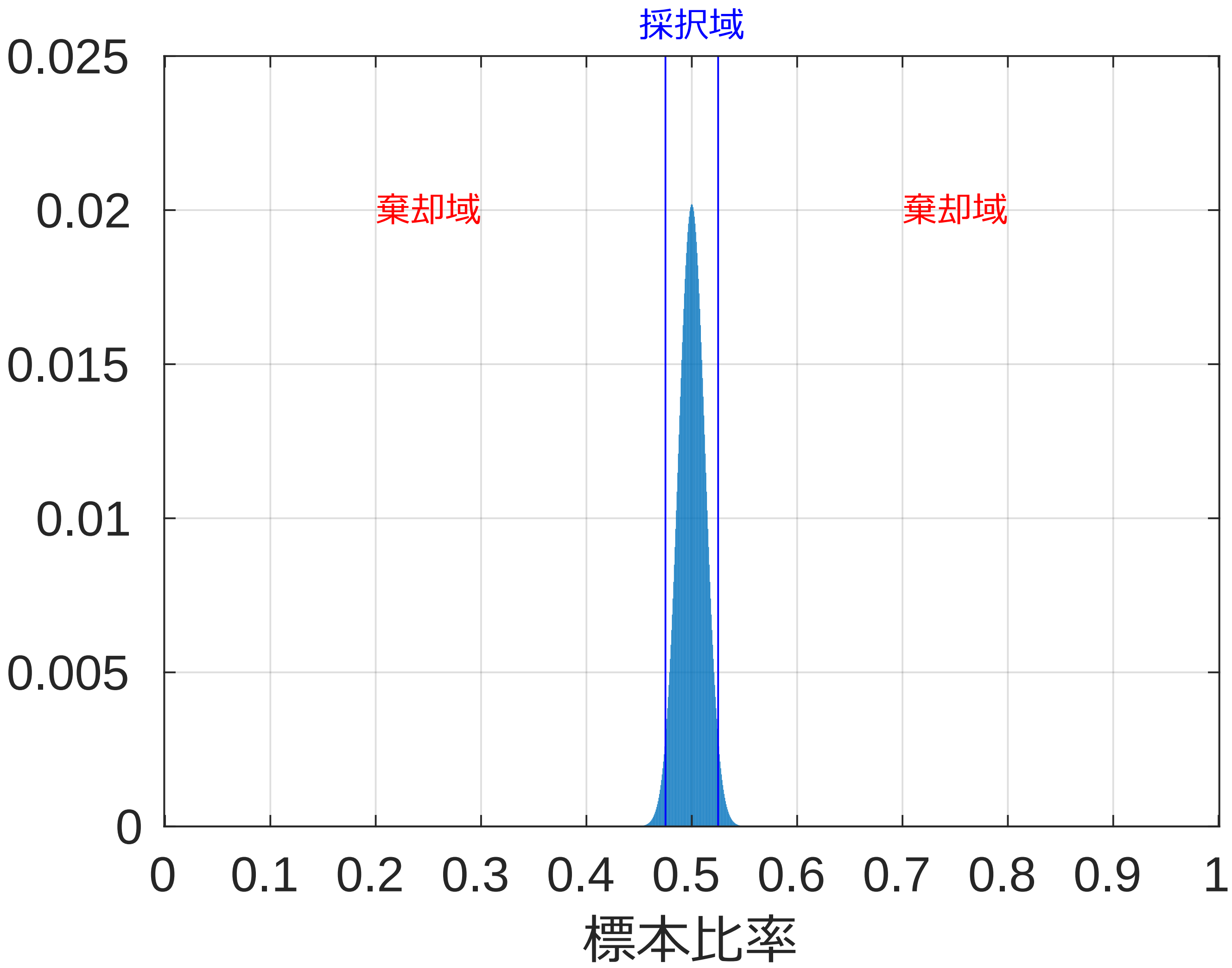
「未来の予感」の検定結果 (2)

- Step3* では標本分布から P 値を計算します.
- P 値は, 「成功数が 829 回以上になるか, または 731 $(= 780 - 49)$ 回以下になる」事象が生じる確率です.
- 帰無仮説が真 $(\pi = 0.5)$ であると仮定し, 試行数 $n = 1560$ の 2 項分布で評価します.
- 該当する成功数の柱の確率を全部足し上げると, P 値は 1.4% になります.
- Step4 では帰無仮説 H_0 を棄却または採択します.
- P 値が有意水準より小さい場合「帰無仮説が正しくかつ確率的に起きにくいことが起きたと考えるのではなく, 帰無仮説は間違っていた」と判定します.
- 二項検定の結果は $\mathbf{P = 0.014 < 0.05}$ で有意です.
- したがって予知能力は存在するとベムは主張しました.

* binocdf(731,1560,0.5) = 0.0070

「未来の予感」の検定結果（3）

- 学術雑誌 JPSP の査読者もその主張を認め，論文は公刊されました．
- 試行数 $n = 1560$ の 2 項分布による標本分布を下図に示します．



- 真ん中のわずかな部分を除いて，左右の端のほうは柱がないように見えますが，これは低すぎて印刷が見えないだけで，ちゃんと 0 から 1 まで 1561 本の柱が存在しています．

「未来の予感」の検定結果 (4)

- 試行数 $n = 1560$ のときの成功数の棄却域は

$$\{0 \leq x \leq 741 \quad \text{および} \quad 819 \leq x \leq 1560\}$$

- 標本比率の棄却域は

$$\{0.000 \leq \text{標本比率} \leq 0.475 \quad \text{および} \quad 0.525 \leq \text{標本比率} \leq 1.000\}$$

- 中央のわずかな範囲が採択域です．
- 52.5% 以上当てられれば，帰無仮説を 5% 水準で棄却して「有意である」とベムは論文に書くことができたのです．
- 「超能力の存在を示す」ためには，少なくとも $\pi \neq 0.5$ を示す必要がありますが，それは必要条件にしか過ぎません．
- 53.1% では，そもそも誰も予知能力があるとは思わないからです．

「念力実験」をしてみよう (1)

- もし試行数が，たとえば 10 万回になったら二項検定はどうなるでしょう．
- 写真の位置を 10 万回予言するのは大変ですから

「表が出ろ」と念じて 1 円玉を 10 万回投げ，表が出た枚数を数える

という「念力実験」はどうでしょうか．

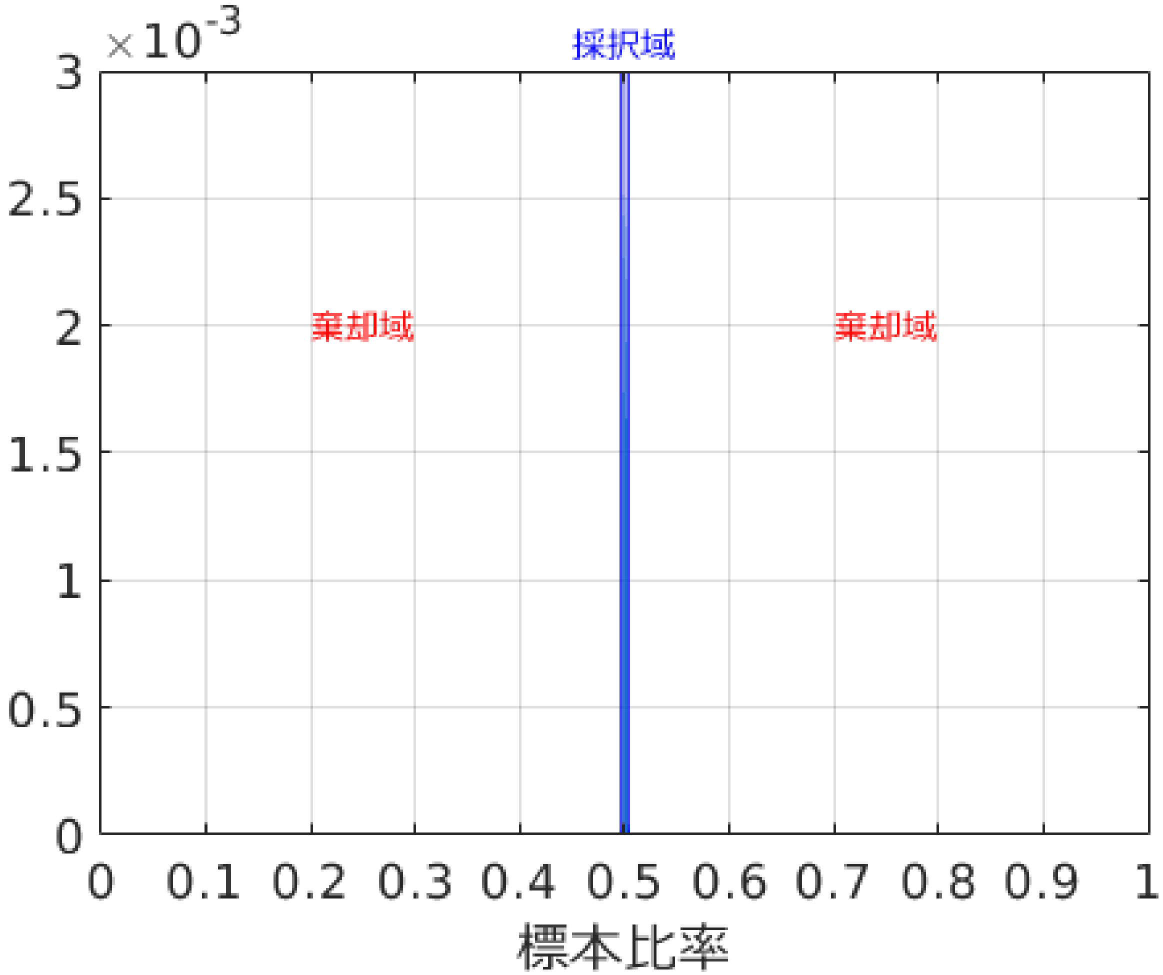
- この実験における二項検定の標本比率の棄却域は

$$\{0.0000 \leq \text{標本比率} \leq 0.4969 \quad \text{および} \quad 0.5031 \leq \text{標本比率} \leq 1.0000\}$$

- 50.3% 以上表が出れば「$P < 0.05$ で有意」と論文に書くことができます．
- 学術雑誌 JPSP のように，P 値が 5% を切ることを掲載の条件とする雑誌に対して，ベムのように「念力は存在する」と主張できます．

「念力実験」をしてみよう (2)

- 標本分布は棄却域で覆われた針状の下図で，採択域はまるで針の穴です．
- 試行数が $n = 10$ 万になると，採択域は 0.5 ± 0.0031 です．
- n の増加に伴って採択域は 0.5 を中心に，いくらでも狭まっていきます．



米国統計学会の声明

- 米国統計学会 (American Statistical Association, ASA) が 2016 年に「統計的有意性と P 値に関する声明」を発表しました.

統計的有意性と P 値に関する ASA 声明

(1) P 値は，データと特定の統計モデル（仮説も統計モデルの要素の一つ）が矛盾する程度を示す指標の一つである.
(2) P 値は，調べている仮説が正しい確率や，データが偶然のみで得られた確率を測るものではない.
(3) 科学的な結論や，ビジネス，政策における決定は，P 値がある値 (有意水準) を超えたかどうかにのみ基づくべきではない.
(4) 適正な推測のためには，すべてを報告する透明性が必要である.
(5) P 値や統計的有意性は，効果の大きさや結果の重要性を意味しない.
(6) P 値は，それだけでは統計モデルや仮説に関するエビデンスの，良い指標とはならない.

ダイエット法に効果はあるか（1）

ダイエット問題

あるダイエット法の効果を調べるために，20 名の女性に参加してもらいました． プログラム参加前体重と，参加後体重と，前後の体重差を下表に示します． さて，このダイエット法は有効でしょうか．

ダイエット参加前と参加後の体重の差（kg）

被験者番号	1	2	3	4	5	6	7	8	9	10
before 体重	53.1	51.5	45.5	55.5	49.6	50.1	59.2	54.7	53.0	48.6
after 体重	48.3	45.2	46.6	56.6	41.2	44.6	51.9	55.5	45.4	47.6
体重差	4.8	6.3	-1.1	-1.1	8.4	5.5	7.3	-0.8	7.6	1.0
被験者番号	11	12	13	14	15	16	17	18	19	20
before 体重	55.3	52.6	51.7	48.6	56.4	42.9	50.3	42.4	51.2	39.1
after 体重	50.6	54.5	49.0	43.9	53.8	40.1	52.8	35.3	55.6	38.0
体重差	4.7	-1.9	2.7	4.7	2.6	2.8	-2.5	7.1	-4.4	1.1

ダイエット法に効果はあるか (2)

- 測定とは，回答者や事物などの観測対象に対し，定められた操作に基づき，数値を割り当てることです．
- 測定によって割り当てられた数値を，測定値といいます．
- 測定値の集まりを，データといいます．
- 観測対象の性質を表現する 1 つの側面に関する測定値の集合が変数です．
- 「before 体重」「after 体重」「体重差」は 3 つの変数です．
- 第 i 番目の観測対象の「before 体重」と「after 体重」と「体重差」を，それぞれ x_{bi}，x_{ai}，$x_{差 i}$ と表記すると，3 つの変数の間には

$$x_{差 i} = x_{bi} - x_{ai} \quad (i = 1, \ldots, n) \tag{10}$$

  という関係があります．
- たとえば $i = 1$ の場合には，4.8 kg = 53.1 kg − 48.3 kg です．

ダイエット法に効果はあるか (3)

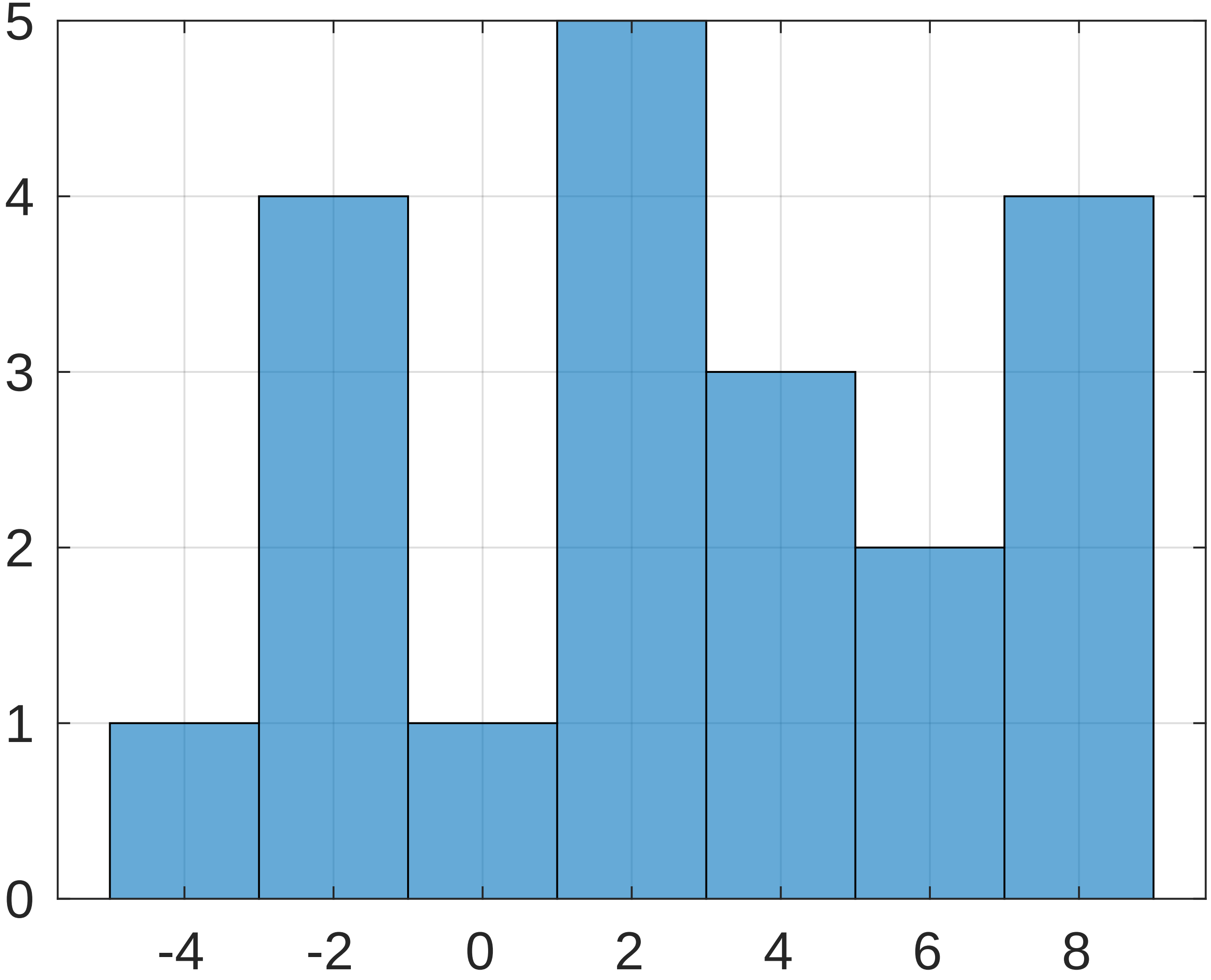
- 観測対象の数は $\boldsymbol{n}$ と表記し，この場合は $\boldsymbol{n = 20}$ です．
- $(\boldsymbol{i = 1, \ldots, n})$ は，添え字 $\boldsymbol{i}$ が 1 から 20 まで動くことを示しています．
- 小さな文字の「差，$\boldsymbol{b}$，$\boldsymbol{a}$」は変数を区別するための添え字です．
- 3 つの変数を同時に扱うと大変なので，ここでは当分の間，「体重差」$\boldsymbol{x}_{差\,\boldsymbol{i}}$ を中心にして，初等的な統計分析の方法を説明します．
- データの分析は 1 つ 1 つの測定値を丁寧に観察することから始めます．
- 「体重差」を目視で観察すると，負の値が 6 つ観察されます．
- つまり，20 人中 6 人は，ダイエットプログラム後に，逆に太っています．
- しかし，目視には限界がありますから，続いて変数の分布を調べます．
- 分布とは，どこにどれくらいのデータが観察されているかの様子です．

ヒストグラム (1)

- データには 2 種類あります.
- 「成功数」のような離散的に測定したデータを計数データといいます.
- 「体重差」のような連続量を測定したデータを計量データといいます.
- 計量データの分布を調べるにはヒストグラムを作成することが効果的です.
- ヒストグラムは, 横軸に階級, 縦軸に度数を配した統計グラフです.
- ここで階級とは測定値の区間, 度数は階級に観測された測定値の数です.
- 区間の長さを階級幅といいます.
- 階級の真ん中の値を階級値といいます.
- 統計グラフにより, データの性質を表現することを, 図的要約といいます.

ヒストグラム（2）

- たとえば一番左の柱は，-5 kg 以上 -3 kg 未満の区間に観測された測定値が 1 つある，ことを示しています．



- 左図の階級幅は 2 kg です．
- 一番左の階級の階級値は -4 kg です．

数値要約

- 図的要約は，ひと目でデータの状態が分かり，便利です．
- しかし手軽さに欠けます．
- そこで，データの特徴を要約的に記述する数的な指標を利用します．
- データを独立変数とみたときの，関数（3）式

$$\text{統計量} = f(\text{データ})$$

を統計量といいました．

- 統計量の中で有意性検定に利用するものを，特に検定統計量といいます．
- データの性質を縮約するための統計量を，要約統計量といいます．
- 要約統計量でデータの特徴を要約することを，数値要約といいます．
- 数値要約は，図的要約と対になる用語です．

代表値（1）

- 初等的な要約統計量には，代表値と散布度とがあります．
- 分布の位置を記述する要約統計量を，代表値といいます．
- データ全体の特徴を，1 つの数値で表す場合には，代表値を利用します．
- 具体的な代表値には，平均値・中央値・最頻値があります．
- 平均値は，すべての測定値の合計を，n で割って

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_{n-1} + x_n\right) \tag{11}$$

と求めます．

- ここで x_i は，i 番目の測定値です．
- データから得られたことを強調したい場合には，$\bar{x}$ を標本平均と呼びます．

代表値 (2)

- 「体重差」の標本平均は，$\bar{x}_{差} = 2.74$ kg でした．
- また $\bar{x}_b = 50.56$ kg，$\bar{x}_a = 47.83$ kg でした．
- データを小さい順に並べることをソートといいます．
- たとえば，「体重差」のソートされたデータは以下です．

-4.4	-2.5	-1.9	-1.1	-1.1	-0.8	1.0	1.1	2.6	2.7
2.8	4.7	4.7	4.8	5.5	6.3	7.1	7.3	7.6	8.4

- この分布から測定値を 1 つ取り出すという試行を考えます．
- この試行で x 以下の値が観察される確率が q であるとき

x の累積確率は q である

あるいは x は $q \times 100\%$ 点 といいます．

代表値（3）

- 中央値は，ソートしたデータの真ん中の測定値です．
- n が奇数の場合，$(n+1)/2$ 番目の測定値が中央値です．
- たとえば $n = 19$ の場合は，データを小さい順に並べたときの 10 番目の測定値が中央値です．
- n が偶数の場合，$n/2$ 番目と $(n/2)+1$ 番目の測定値の平均が中央値です．
- 「体重差」のデータは $n = 20$ なので，「体重差」の中央値は，10 番目と 11 番目の測定値の平均値，2.75kg（$= (2.7+2.8)/2$）となります．
- 最頻値は，最大度数を有する階級値です．
- 「体重差」のヒストグラムを観察すると，階級幅 2 kg の場合は，1.0 kg 以上 3.0 kg 未満の区間の度数が 5 で最大です．
- したがって最頻値は 2.0 kg です．

散布度（1）

- 分布の中心的な位置から，平均的に，どれほど測定値が散らばっているか，に関する要約統計量を，散布度といいます．
- 散布度の要約統計量としては，分散と標準偏差がよく使われます．
- 測定値 x_i から平均値 $\bar{x}$ を引いて，それを 2 乗した値の和を $n-1$ で割った統計量

$$s^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2+\cdots+(x_i-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right) \tag{12}$$

を分散といいます．

- データから計算したことを強調したい場合には不偏分散といいます．
- 「体重差」の不偏分散は，$s^2_{差} = 14.50$ です．
- ただし分散の単位は，測定値の 2 乗です．

散布度（2）

- このため，平均的な散布度として，分散の値は，具体的に解釈できません．
- この欠点を補うために

$$s = \sqrt{s^2}$$

  のように分散の平方根をとり，元の測定単位に戻します．
- これを**標準偏差**（standard deviation）といいます．
- 標準偏差は sd と略記することもあります．
- 標準偏差が大きい（小さい）なら，平均値から離れた（近づいた）測定値が平均的に観察されます．
- 「体重差」の標準偏差は，$s_{差} = 3.81$ kg でした．
- これは，平均的に平均から約 3.8 kg の距離で，測定値が散らばっている，ということを示しています．

対応ある 2 群の平均値差の検定 (1)

- 有意性検定は母比率の検定ばかりではありません.
- ここでは検定の考え方に慣れるために，母比率の検定と並んで，最も頻繁に利用される，母平均の差の検定を紹介します.
- 20 人の体重減少の平均値が，2.74 kg であるデータは，果たしてこのダイエット法の有効性を示しているのでしょうか.
- 同じ状況下で無数の追試実験を行ったと想定したとき，体重減少の平均値が，平均的に十分に大きいことが大切です.
- 平均的な平均値は，データから計算された標本平均 $\bar{x}$ とは異なっています.
- 区別するために，平均的な平均値を母平均 μ (ミュー) と別称します.

学術的には標本平均 $\bar{x}$ ではなく，母平均 μ の値に興味があるのです.

対応ある 2 群の平均値差の検定 (2)

- ここで登場するのが，対応ある 2 群の平均値差の検定です．
- 検定名の中の 2 群の平均値差とは，どういう意味でしょうか．
- 2 つの群「before 体重」と「after 体重」の母平均の差という意味です．
- 「対応ある」とは「before 体重」と「after 体重」は同一人物から測定されている，という意味です．
- Step1 では，帰無仮説 H_0 と対立仮説 H_1 を設定します．
- 帰無仮説を，2 つの群の母平均は等しい

$$H_0 : \mu_b = \mu_a \quad \text{または} \quad \mu_b - \mu_a = 0 \tag{13}$$

と設定します．

- μ_b は「before 体重」の母平均，μ_a は「after 体重」の母平均です．

対応ある 2 群の平均値差の検定 (3)

- 2 つの母平均には $\mu_{\text{差}} = \mu_b - \mu_a$ という関係があります．
- $\mu_{\text{差}}$ は「体重差」の母平均であり，上の関係を利用すると，帰無仮説は

$$H_0 : \mu_{\text{差}} = 0 \tag{14}$$

と書き換えることもできます．

- (13) 式，(14) 式に登場した 3 つの表現は，どれを用いても同じです．

体重減少の母平均は 0 である，つまり

「ダイエット法には効果がない」

という趣旨の仮説です．

- 対立仮説は，その否定ですから，$H_1 : \mu_{\text{差}} \neq 0$ です．

対応ある 2 群の平均値差の検定 (4)

- Step2 では，帰無仮説 H_0 を真として検定統計量を計算します．
- 「未来の予感」では 2 項分布を用いた検定を行い，検定統計量には標本比率を使いました．
- ここでは正規分布を利用した検定を行います．
- 正規分布は，私たちの身の回りで生じる様々な連続的変数の分布を近似するのに，最も頻繁に利用される確率分布です．
- 平均値の付近で度数が大きく，両側に離れるに従って，度数が小さくなるデータを記述するのに適しています．
- 変数 x が正規分布に従っている場合には

$$\text{正規分布}(x|\mu, \sigma)$$

と表記します．

対応ある 2 群の平均値差の検定 (5)

- 正規分布は，母平均 μ と母標準偏差 σ で形状が決まります．
- 一般的に，確率分布の形を決める数的な指標を母数といいます．
- 正規分布の形は μ と σ で決まりますから，あらゆる正規分布を $\mu = 0$, $\sigma = 1$ になるように変換しておけば，常に同じ形で比較ができます．
- この目的のために x に次の一次変換を施します．

$$\frac{x - \mu}{\sigma} \tag{15}$$

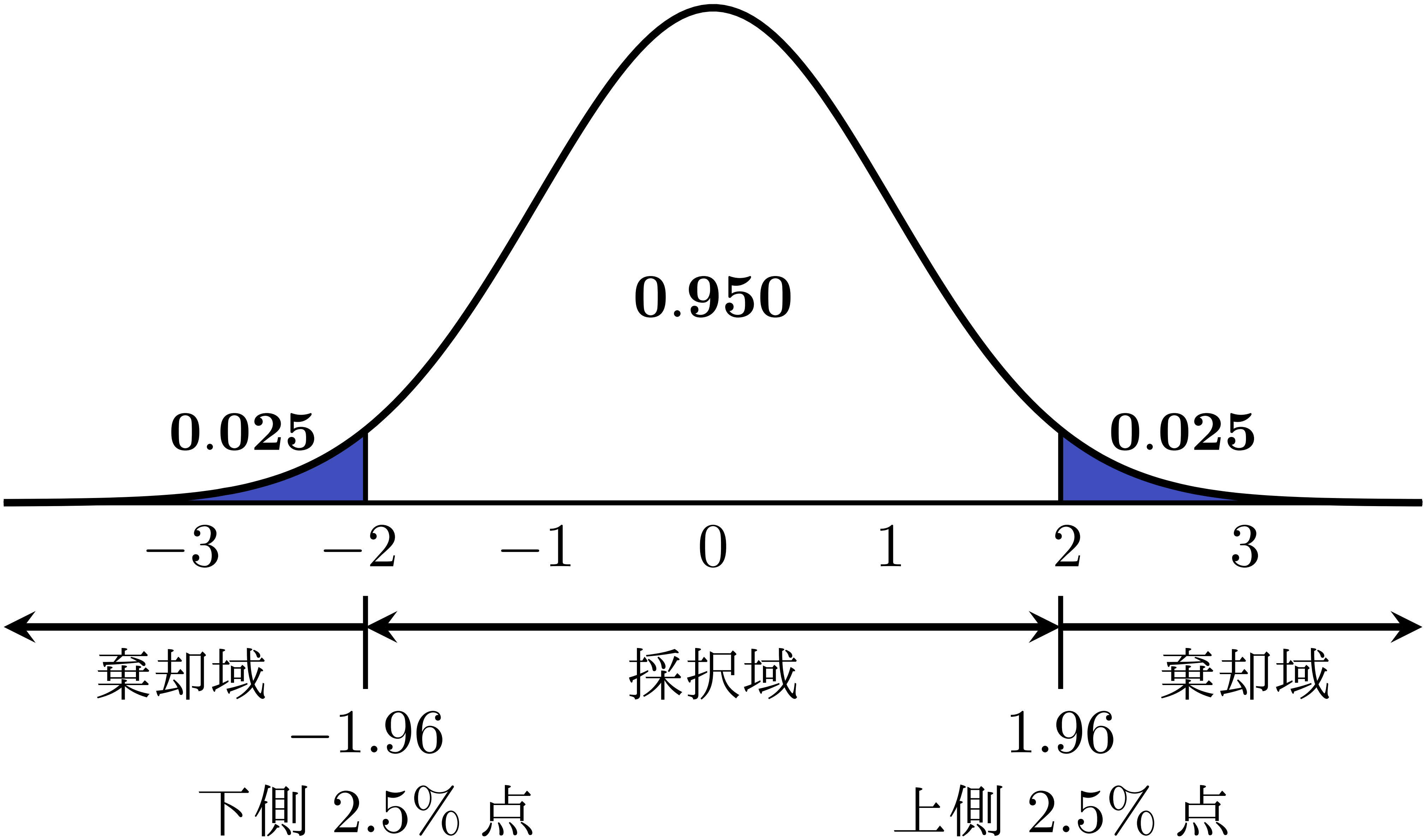
- このような変換を標準化といいます．
- 標準化された値は

$$\text{正規分布}(z|\mu = 0, \sigma = 1) \tag{16}$$

に従うことが知られています．

対応ある 2 群の平均値差の検定 (6)

- 母平均 $\boldsymbol{\mu = 0}$, 母標準偏差 $\boldsymbol{\sigma = 1}$ の正規分布を標準正規分布といいます.



対応ある 2 群の平均値差の検定 (7)

- (10) 式 $x_{差\,i} = x_{bi} - x_{ai} \quad (i = 1, \ldots, n)$ の「体重差」が正規分布

$$\text{正規分布}(x_{差\,i} | \mu_{差}, \sigma_{差}) \tag{17}$$

  に従い，$\mu_{差}$ と $\sigma_{差}$ は，それぞれ $x_{差\,i}$ の母平均と母標準偏差とします．

- いま，帰無仮説 (14) 式 $H_0 : \mu_{差} = 0$ が真であると仮定して，試行数 n の追試実験を無数に繰り返すことを想像します．

- さらにその実験ごとに，次の統計量を毎回計算し記録します．

$$z = \frac{\bar{x}_b - \bar{x}_a}{s_{差}} \times \sqrt{n} = \frac{\bar{x}_{差}}{s_{差}} \times \sqrt{n} \tag{18}$$

- ここで $s_{差}$ は，不偏分散の平方根による $x_{差\,i}$ の標準偏差です．

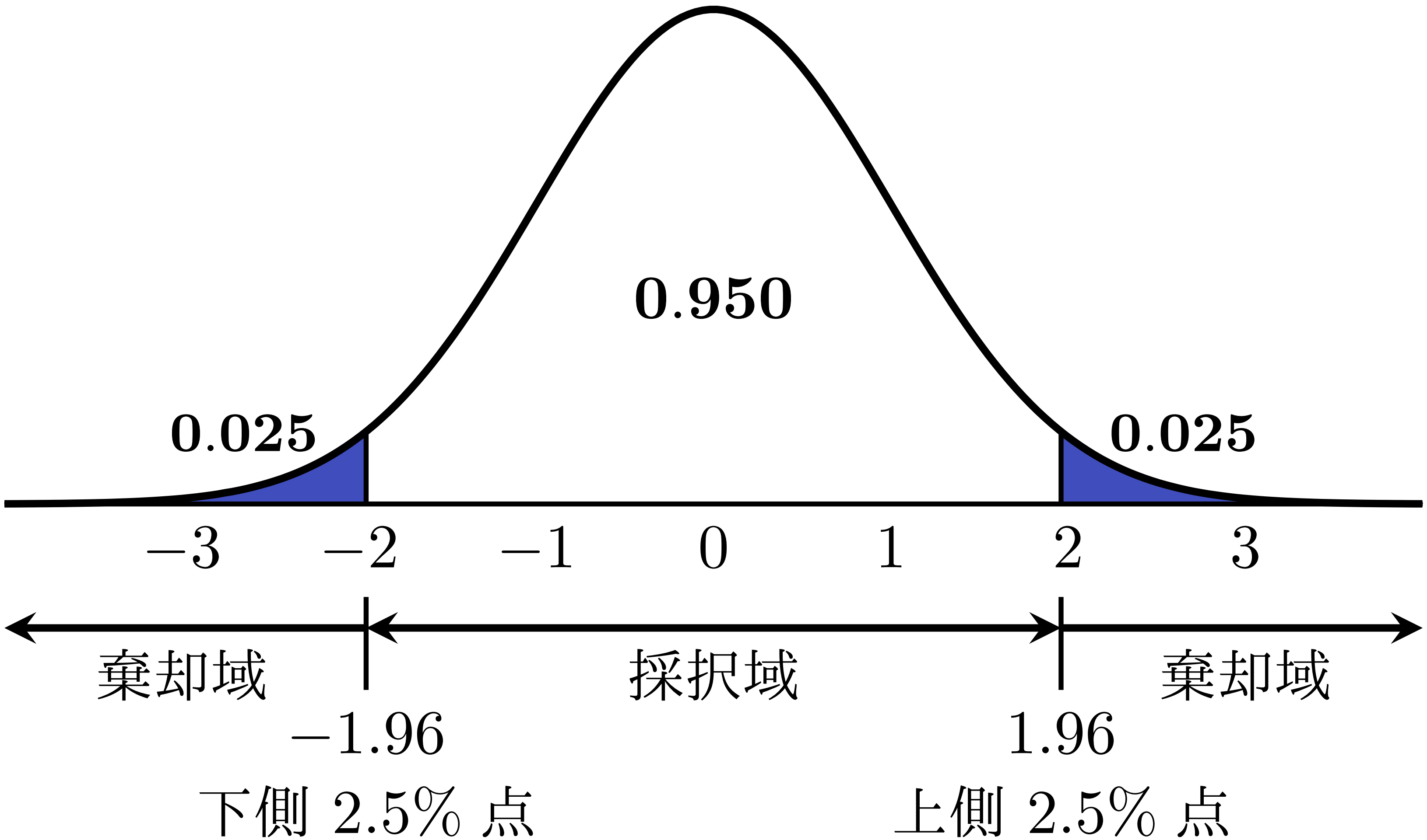
- $s_{差}$ を差得点の標準偏差と呼びます．

対応ある 2 群の平均値差の検定 (8)

- 本当は，実験は 1 回しか実施していないのですから，あくまでも想像です．
- 統計量の，この想像上の分布を標本分布といいます．
- このとき z の標本分布は，標準正規分布で近似できます．
- z は，帰無仮説（14）式 $\boldsymbol{H_0 : \mu_{差} = 0}$ に関する検定統計量となります．
- Step3 では標本分布から P 値を計算します．
- 検定統計量を定めたら，P 値を用いて確率的な評価をします．
- P 値とは「帰無仮説が真であるという仮定の下で，検定統計量がデータから計算された値以上に甚だしい値となる確率」でした．
- 計量データの確率分布は，正規分布に限らず，グラフの曲線と横軸で囲まれた面積が，その区間で値が観察される確率に一致します．

対応ある 2 群の平均値差の検定 (9)

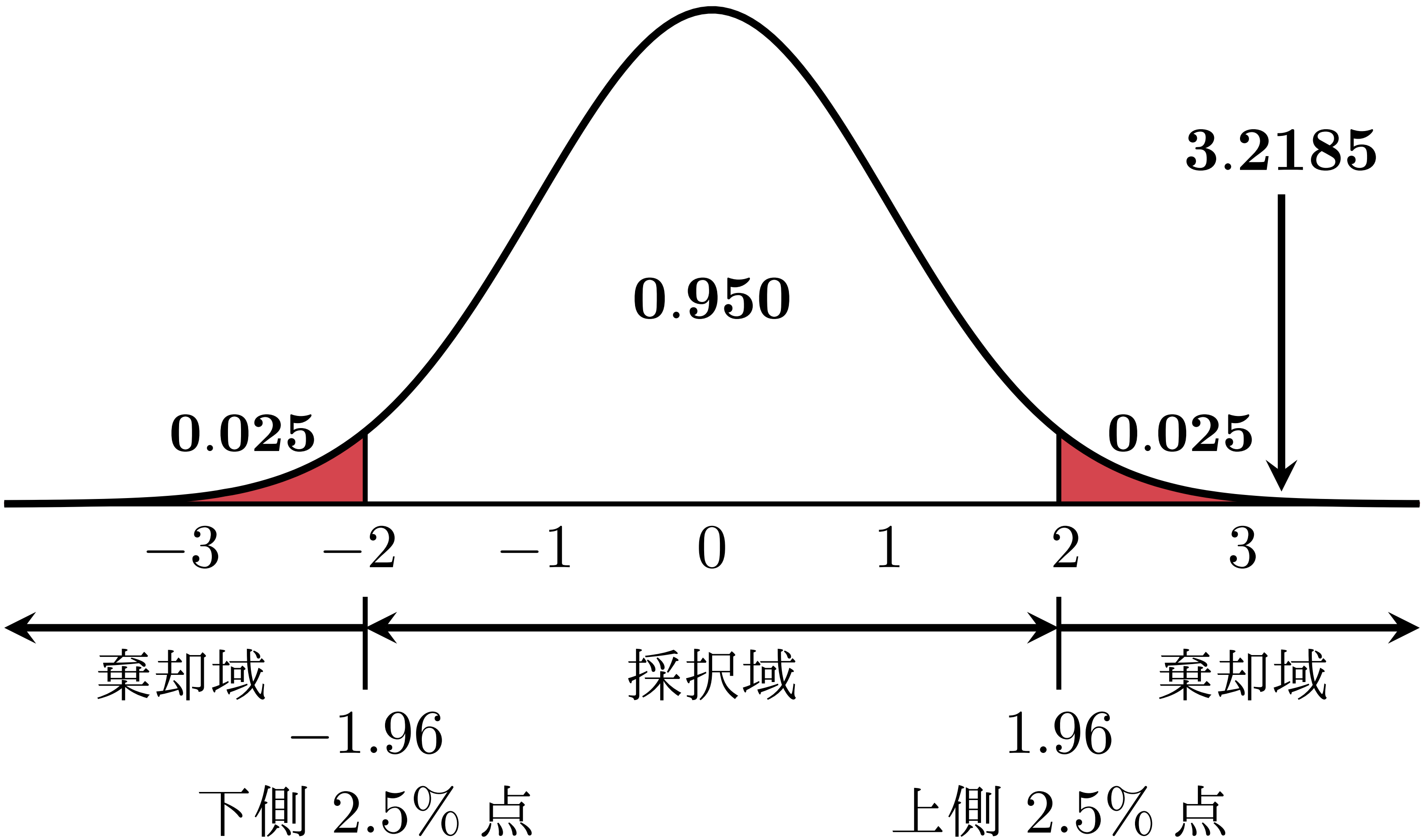
- +1.96 を上側 2.5% 点といい,+1.96 以上の値が観察される確率は 0.025
- -1.96 を下側 2.5% 点といい,-1.96 以下の値が観察される確率は 0.025
- 合わせると 5% で，絶対値を使って表現すると $\mathbf{P(1.96 \le |z|) = 0.05}$



対応ある 2 群の平均値差の検定（10）

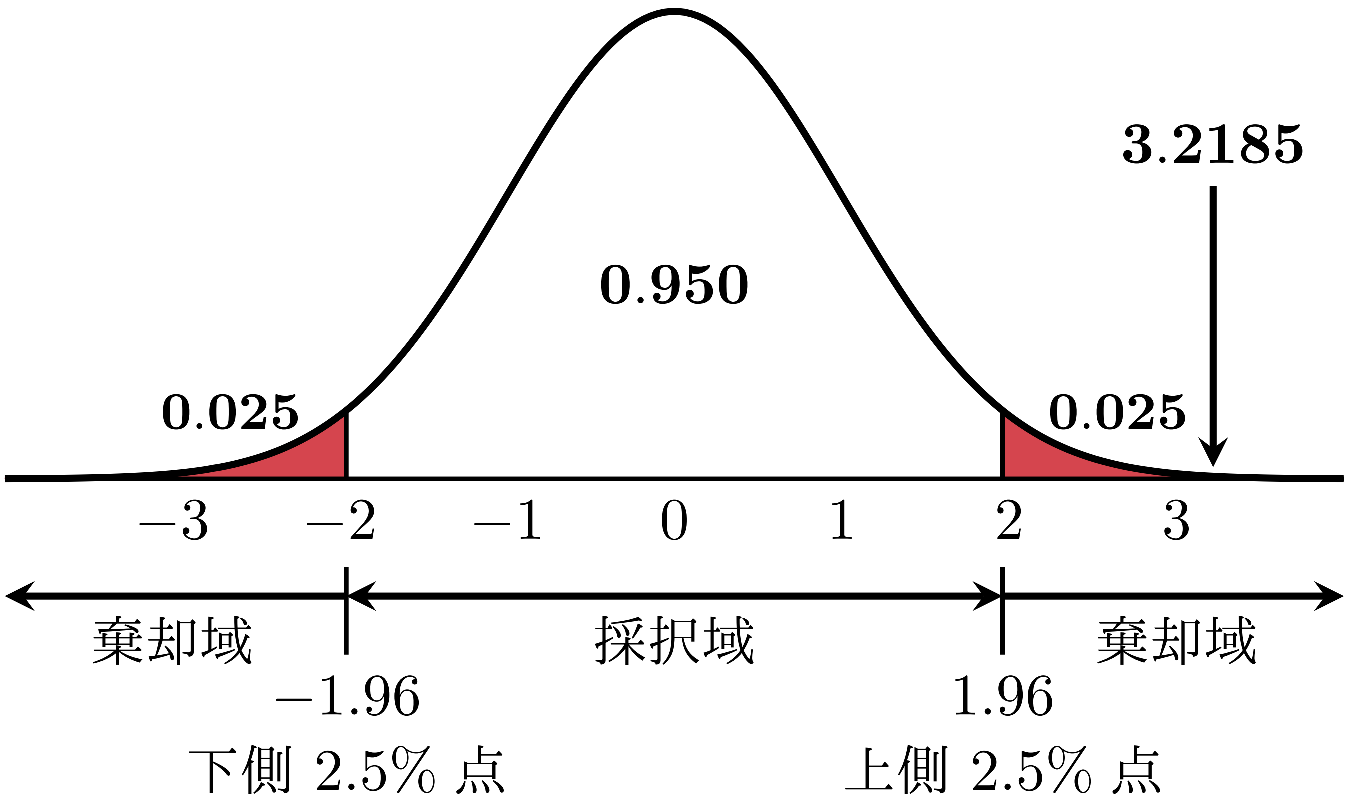
- ダイエットの実験の検定統計量は

$$z = \frac{50.57 - 47.83}{3.81} \times \sqrt{20} = \frac{2.74}{3.81} \times \sqrt{20} = 3.2185 \tag{19}$$



対応ある 2 群の平均値差の検定 (11)

- 検定統計量 z の絶対値が **3.2185** より大きくなる確率が P 値でした．
- ゆえに **P = 0.0013 < 0.05** です．[†]



- Step4 では帰無仮説 H_0 を棄却または採択します．
- **P < 0.05** ですから，帰無仮説 (2.5) 式を棄却し「体重差」の標本平均 $\bar{x}_{差} = 2.74$ kg は，統計的に有意であると結論します．

[†] `(1-normcdf(3.2185))*2 = 0.0013`

統計的に有意でも科学的に無意味（1）

- 検定統計量 z は（18）式

$$z = \frac{\bar{x}_b - \bar{x}_a}{s_{差}} \times \sqrt{n} = \frac{\bar{x}_{差}}{s_{差}} \times \sqrt{n}$$

から分かるように，標準偏差，標本平均，n の 3 つから計算されます．

- 減量問題なので上側だけを考えると，$1.96 < z$ なら統計的に有意です．
- ここに（18）式に代入して，式変形すると（$n > 0$）

$$1.96 < z \Rightarrow 1.96 < \frac{\bar{x}_{差}}{s_{差}} \times \sqrt{n} \Rightarrow \bar{x}_{差} > \frac{1.96 \times s_{差}}{\sqrt{n}} \tag{20}$$

- この式が成り立てば「統計的に有意」です．
- 標本平均 $\bar{x}_{差}$ は，平均的な体重の減少量であり，ダイエット法の有効性を本質的に反映しています．

統計的に有意でも科学的に無意味（2）

- それに対して，観測対象の数 n は実験者が自由に決められます．
- n は，ダイエット法の科学的性質や有用性とは，まったく無関係です．
- 標本標準偏差 $s_{差}$ は，ダイエット法の有効性を直接的には反映しないので，標本標準偏差 $s_{差}$ を，データから計算した $s_{差} = 3.81$ に固定します．

$$\bar{x}_{差} > \frac{1.96 \times 3.81}{\sqrt{n}} = f(n) \tag{21}$$

この式は，平均的な体重の減少量が何 kg 以上なら，言い換えるなら標本平均 $\bar{x}_{差}$ が何 kg 以上なら,「有意であった」のかを示しています．

- この式は観測対象の数 n の関数です．

統計的に有意でも科学的に無意味（3）

- $\bar{x}_{\text{差}} > \dfrac{1.96 \times 3.81}{\sqrt{n}} = f(n)$ で n を動かしてみると

$$\bar{x}_{\text{差}} > 1.670 = f(n = 20)$$
$$\bar{x}_{\text{差}} > 1.056 = f(n = 50)$$
$$\bar{x}_{\text{差}} > 0.747 = f(n = 100)$$
$$\bar{x}_{\text{差}} > 0.236 = f(n = 1000)$$
$$\bar{x}_{\text{差}} > 0.106 = f(n = 5000)$$
$$\bar{x}_{\text{差}} > 0.024 = f(n = 10000)$$

- 第 1 式は $n = 20$ の場合ですから，実験状況そのものです．
- 体重の平均減少量が 1670 g 以上であれば，この実験は「統計的に有意」であったことを示しています．
- 被験者を 100 人に増やすと，どうなるでしょう．

統計的に有意でも科学的に無意味（4）

- 被験者を 100 人に増やすと，体重の平均減少量が 747 g以上であれば，この実験は「統計的に有意」であったことを示しています．

$$\bar{x}_{差} > 0.747 = f(n = 100)$$

- $n = 10$万 ならば，$\bar{x}_{差}$ は 24 g 以上で，立派に「統計的に有意」です．

$$\bar{x}_{差} > 0.024 = f(n = 10000)$$

- しかし 24 g のダイエット法など，科学的にはまったく無意味です．
- 観測対象の数 n の増加に伴い，平均的体重減少は 0 g に近づきます．
- 逆に n が小さい場合，たとえば $n = 5$ とすると

$$\bar{x}_{差} > 3.34 = f(n = 5)$$

となり，仮に平均的に 3.3 kg 痩せても有意にはなりません．

帰無仮説は科学的には偽

- 帰無仮説 H_0 と対立仮説 H_1 は，それぞれ

$$H_0 : \mu_{差} = 0, \quad H_1 : \mu_{差} \neq 0$$

- 帰無仮説は点，その数学的な 1 点を除いて対立仮説は数直線です．
- たとえ $\mu_{差} = 1$ pg（ピコグラム，1 兆分の 1 g）でも，対立仮説が真，帰無仮説は偽です．
- データをとる前から，帰無仮説は厳密には偽です．
- 「学術的にまったく無意味でも，対立仮説が真であれば，観測対象の数 n が大きくなると必ず棄却される」という性質は，査読の判定に利用されているほとんどの有意性検定に，共通した性質なのです．
- この弊害を克服するために，学術的価値に統計分析の指標を連動させることが必要です．

参考文献

- 柳川 堯 (2018) P値：その正しい理解と適用，近代科学社
- 豊田 秀樹 (2020) 瀕死の統計学を救え!，朝倉書店
- Daryl J. Bem (2011) Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, **100**, 407-425.
- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, **70**, 129-133.

(順不同)