第 5 章：推測の基礎

OpenIntro Statistics 第 4 版（日本語版）

原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

点推定値と誤差

- 私たちはしばしば母集団のパラメータ（母数）に関心を持つ。
- 母集団全体からデータを収集することは難しいため、未知の母集団パラメータの推定に標本統計量を点推定値として使用する。
- 推定の誤差 ＝ 母集団パラメータと標本統計量の差
- 偏り（バイアス）とは、真の母集団パラメータを系統的に過大または過小に推定する傾向のことである。
- 標本誤差とは、ある標本から次の標本へと推定値がどれだけ変動しやすいかを表す。
- 統計学の多くは標本誤差の理解と定量化に焦点を当てており、標本サイズはこの誤差を定量化するのに役立つ。

アメリカの各州からそれぞれ 1,000 人の成人を無作為に抽出したとする。各州の身長の標本平均は同じになると思うか、やや異なると思うか、それとも大きく異なると思うか？

同じではないが、それほど大きな差はなく、やや異なる程度だろう。

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

誤差の範囲（許容誤差）

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%：38.1%から 43.9%の国民が、中年層や高齢者よりも若い成人が今日の経済で最も苦しんでいると考えていることを 95%の信頼度で確信している。
- 49% ± 4.4%：18〜34 歳の 44.6%から 53.4%が、請求書を支払うためだけに望まない仕事に就いたことがあると 95%の信頼度で確信している。

アメリカの成人で太陽光エネルギーの拡大を支持する割合が $p = 0.88$ であるとする（これが関心のあるパラメータである）。無作為に選ばれたアメリカの成人は、太陽光エネルギーの拡大を支持する可能性が高いか、低いか？

より支持する可能性が高い。

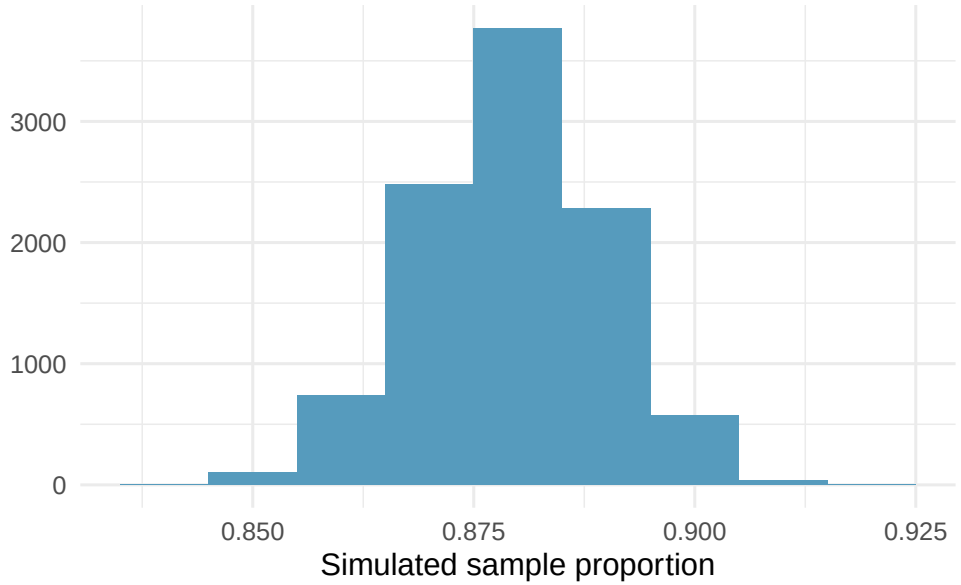
アメリカの成人全体の母集団にアクセスできない場合（これはよくある状況である）、太陽光エネルギーの拡大を支持するアメリカの成人の割合を推定するために、母集団から標本を抽出し、その標本割合を未知の母集団割合の最良の推定値として使用することが考えられる。

- 母集団からアメリカの成人 1,000 人を復元抽出し、太陽光エネルギーの拡大を支持するかどうかを記録する。
- 標本割合を求める。
- クラスの各メンバーが得た標本割合の分布をプロットする。

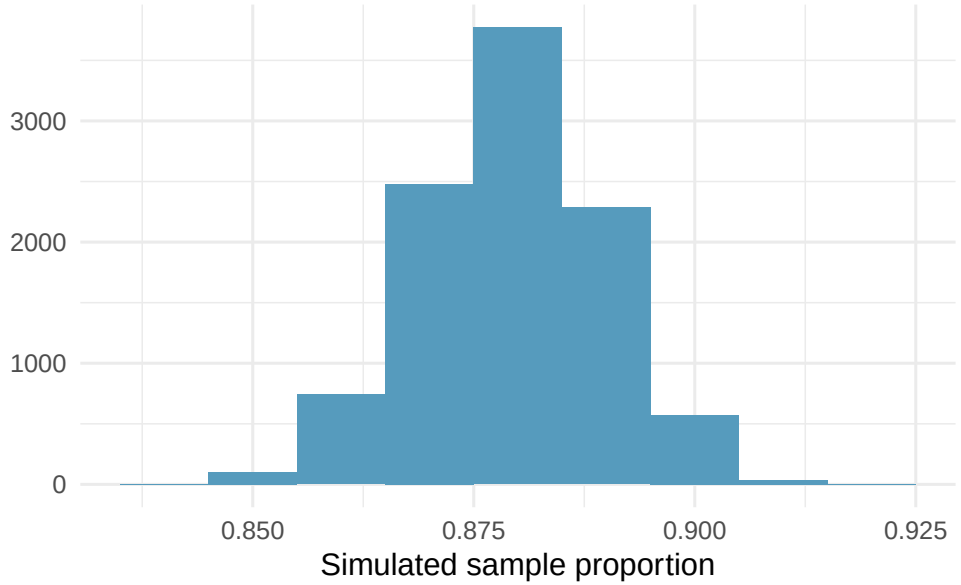
```
# 1. Create a set of 250 million entries, where 88% of
# them are "support" and 12% are "not".

pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size),
                      rep("not", 0.12 * pop_size))


# 2. Sample 1000 entries without replacement.

sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support",
# then divide by # the sample size.

sum(sampled sampled_entries == "support") / 1000
```

標本分布

このプロセスを何度も繰り返して多くの $\hat{p}$ を得たとする。この分布を標本分布と呼ぶ。



この分布の形と中心はどのようなものか？ この分布に基づいて、真の母集団割合はどれくらいだと思うか？



この分布は単峰性でほぼ対称である。真の母集団割合の合理的な推測は、この分布の中心で、約 *0.88* である。

標本分布は実際には観察されない

- 実際の応用では、標本分布を直接観察することは決してないが、点推定値はそのような仮想的な分布から得られるものとして常に考えることが有用である。
- 標本分布を理解することで、実際に観察する点推定値の特性を明らかにし、理解することができる。

中心極限定理

中心極限定理

標本割合は、母集団の割合 p を平均とし、標準誤差を $\sqrt{\frac{p\ (1-p)}{n}}$ とする正規分布にほぼ従う。

$$\hat{p} \sim N\left(\text{平均} = p, SE = \sqrt{\frac{p\ (1-p)}{n}}\right)$$

- 先ほど見た標本分布が対称で真の母集団割合を中心としていたことは偶然ではなかった。
- $SE = \sqrt{\frac{p\ (1-p)}{n}}$ の詳細な証明は省くが、n が増加すると SE が減少することに注目されたい。
  - n が増えるにつれて、標本はより一貫した $\hat{p}$ をもたらし、すなわち $\hat{p}$ 間の変動性が低くなる。

中心極限定理の適用条件

中心極限定理を適用するためには、一定の条件が満たされなければならない：

1. 独立性：抽出された観測値は互いに独立でなければならない。これを確認することは難しいが、以下の場合により可能性が高い：
   - 無作為抽出・割り当てが行われている場合、および
   - 非復元抽出の場合、n が母集団の 10%未満である場合。
2. 標本サイズ：観察された標本において、期待される成功数と失敗数がそれぞれ少なくとも 10 以上あること。
   母集団割合が不明な場合（または仮定できない場合）、この条件を確認することは難しい。そのような場合は、観察された成功数と失敗数がそれぞれ少なくとも 10 以上あることを確認する。

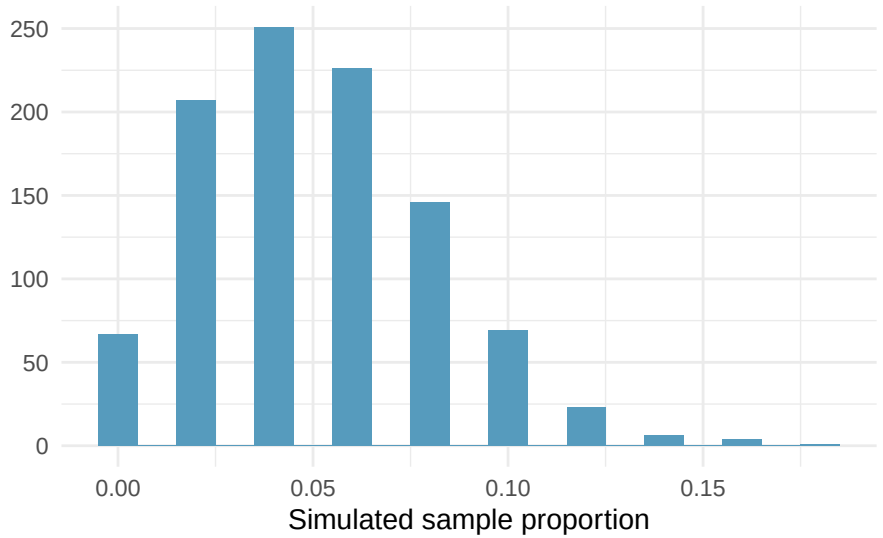
p が未知の場合

- 中心極限定理では $SE = \sqrt{\frac{p\ (1-p)}{n}}$ とされており、np と $n(1-p)$ がそれぞれ少なくとも 10 であることが条件だが、母集団割合 p の値が不明なことが多い。
- そのような場合は p の代わりに $\hat{p}$ を代入する。

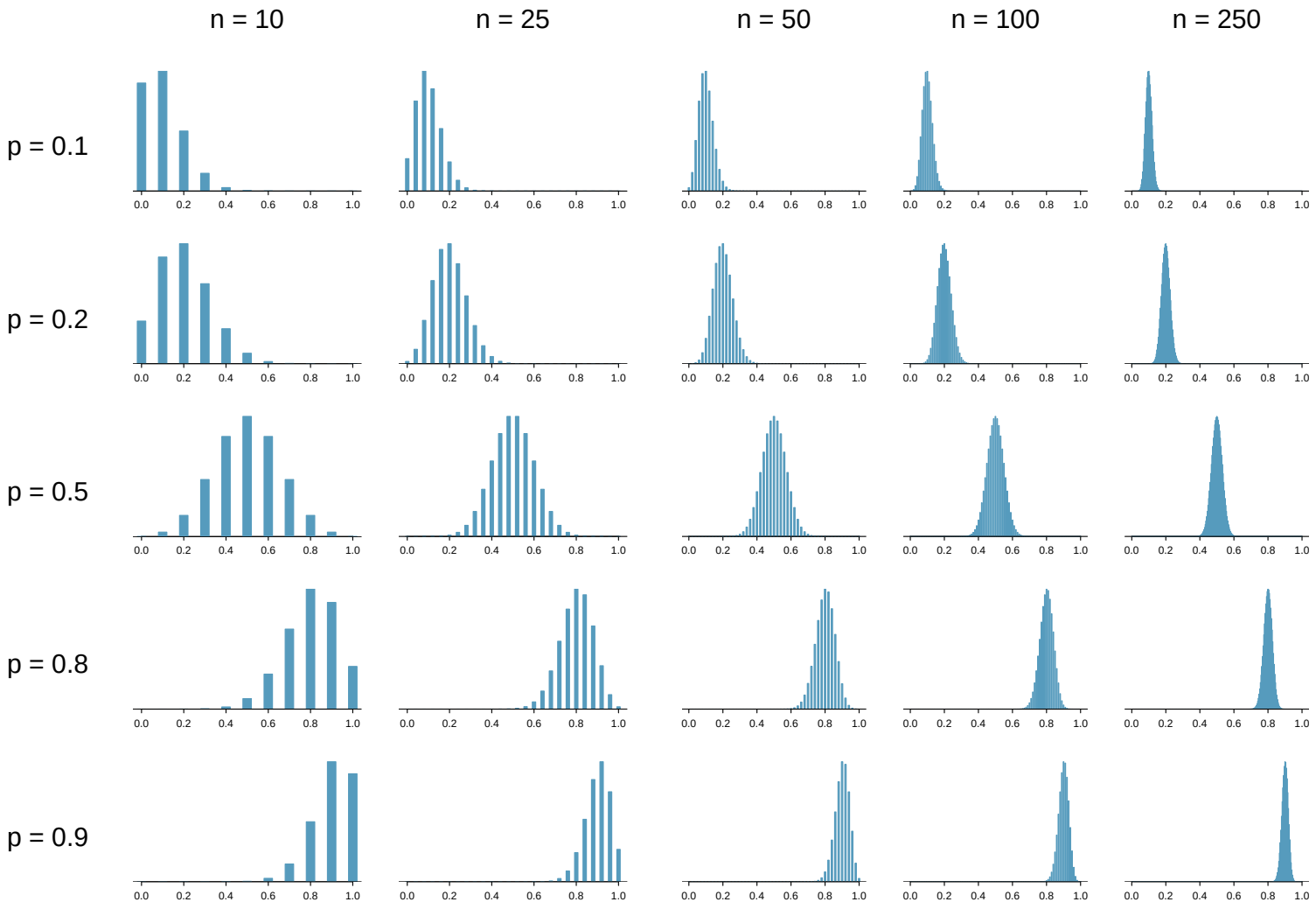
p が小さい場合

真の母集団割合が $p = 0.05$ である母集団があり、この母集団からサイズ $n = 50$ の無作為標本を抽出するとする。各標本の標本割合を計算してこれらの割合をプロットした場合、この分布はほぼ正規分布になると予想されるか？ なぜそうなるか、あるいはならないか？

いいえ、成功・失敗条件が満たされていないため
$(50 \times 0.05 = 2.5)$、標本分布がほぼ正規分布になるとは期待できない。



np および/または $n(1-p)$ が 10 未満の場合はどうなるか？



条件が満たされない場合…

- np または $n(1-p)$ のいずれかが小さい場合、分布はより離散的になる。
- np または $n(1-p)$ が 10 未満の場合、分布はよりゆがんだ形になる。
- np と $n(1-p)$ の両方が大きくなるほど、分布はより正規分布に近くなる。
- np と $n(1-p)$ の両方が非常に大きい場合、分布の離散性はほとんど見られなくなり、分布は正規分布にかなり近い形になる。

他の統計量への枠組みの拡張

- 標本統計量を使ってパラメータを推定するという戦略は非常に一般的であり、割合以外の統計量にも適用できる戦略である。
  - 大学の学生の無作為標本を抽出し、課外活動にいくつ参加しているかを聞いて、その大学の全学生が参加している課外活動の平均数を推定する。
- この章の原則と一般的な考え方は、詳細が若干変わるとしても、他のパラメータにも適用される。

信頼区間

- 母集団パラメータの妥当な値の範囲を信頼区間と呼ぶ。
- 標本統計量だけを使ってパラメータを推定することは、濁った湖で銛を使って魚を突こうとするようなものであり、信頼区間を使うことは網を使って魚を捕まえようとするようなものである。

魚を見た場所に銛を投げても、たいていは外れてしまう。その辺りに網を投げれば、魚を捕まえる可能性が高い。

- 点推定値を報告しても、正確な母集団パラメータには当たらないかもしれない。妥当な値の範囲を報告すれば、パラメータを捉える可能性が高くなる。

写真提供：Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) および Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Facebookによるユーザー興味分類

多くの商業ウェブサイト（ソーシャルメディアプラットフォーム、ニュースサイト、オンライン小売業者など）はユーザーの行動データを収集し、ターゲットを絞ったコンテンツ、おすすめ、広告の配信に活用している。アルゴリズムによる分類システムがアメリカ人の実際の生活をどの程度反映しているかを理解するため、ピュー・リサーチはアメリカのFacebookユーザー850人の代表的な標本に対し、Facebookの「興味のあるもの」ページに表示されたカテゴリーが自分や自分の興味を正確に表しているかどうかを尋ねた。回答者の67%が、表示されたカテゴリーは正確だと答えた。Facebookが自分の興味を正確に分類していると思うアメリカのFacebookユーザーの真の割合を推定せよ。

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook によるユーザー興味分類

$$\hat{p} = 0.67 \qquad n = 850$$

おおよその 95%信頼区間は次のように定義される：

$$\text{点推定値} \pm 1.96 \times SE$$

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

$$\begin{aligned} \hat{p} \pm 1.96 \times SE &= 0.67 \pm 1.96 \times 0.016 \\ &= (0.67 - 0.03, 0.67 + 0.03) \\ &= (0.64, 0.70) \end{aligned}$$

この信頼区間の正しい解釈はどれか？

95%の確信を持って言えることは：

(a) この標本のアメリカの Facebook ユーザーの 64%から 70%が、Facebook は自分の興味を正確に分類していると思っている。

(b) アメリカの Facebook ユーザー全体の 64%から 70%が、Facebook は自分の興味を正確に分類していると思っている

(c) 無作為に選ばれたアメリカの Facebook ユーザーの興味が正確に分類されている確率が 64%から 70%である。

(d) アメリカの Facebook ユーザーの 95%の興味が正確に分類されている確率が 64%から 70%である。

この信頼区間の正しい解釈はどれか？

95%の確信を持って言えることは：

(a) この標本のアメリカの Facebook ユーザーの 64%から 70%が、Facebook は自分の興味を正確に分類していると思っている。

(b) *アメリカの Facebook ユーザー全体の 64%から 70%が、Facebook は自分の興味を正確に分類していると思っている*

(c) 無作為に選ばれたアメリカの Facebook ユーザーの興味が正確に分類されている確率が 64%から 70%である。

(d) アメリカの Facebook ユーザーの 95%の興味が正確に分類されている確率が 64%から 70%である。

「95%の確信」とはどういう意味か？

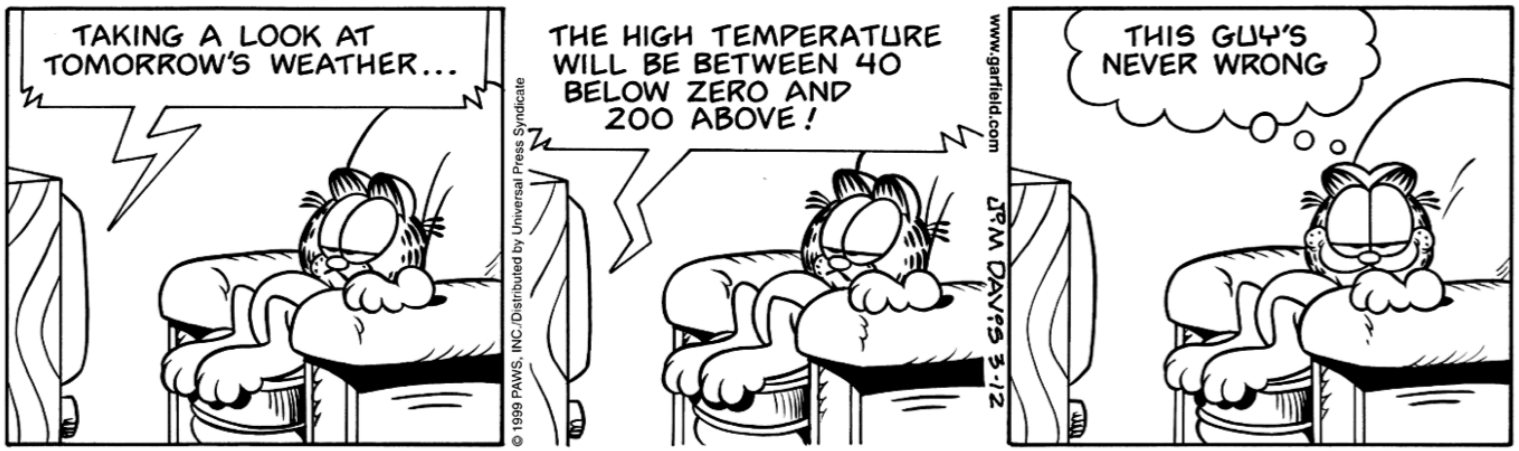
- 多くの標本を抽出し、点推定値 $\pm 1.96 \times SE$ という式を使って各標本から信頼区間を構築したとする。
- すると、その区間の約 95%が真の母集団割合（p）を含むことになる。

区間の幅

母集団パラメータをより確実に捉えたい場合、すなわち信頼水準を高めたい場合、より広い区間を使うべきか、それとも狭い区間を使うべきか？

より広い区間。

より広い区間を使うことの欠点はあるか？



区間が広すぎると、あまり有用な情報が得られないかもしれない。

画像出典：http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

信頼水準の変更

$$\text{点推定値} \pm z^\star \times SE$$

- 信頼区間において、$z^\star \times SE$ を誤差の範囲（許容誤差）と呼び、ある標本に対して、信頼水準が変わると許容誤差も変わる。
- 信頼水準を変えるには、上の式の $z^\star$ を調整する必要がある。
- 実際によく使われる信頼水準は 90%、95%、98%、99%である。
- 95%信頼区間では、$z^\star = 1.96$ である。
- ただし、標準正規（z）分布を使えば、任意の信頼水準に対応する適切な $z^\star$ を求めることができる。

98%信頼区間を計算する際の適切な $z^\star$ はどの Z 値か？

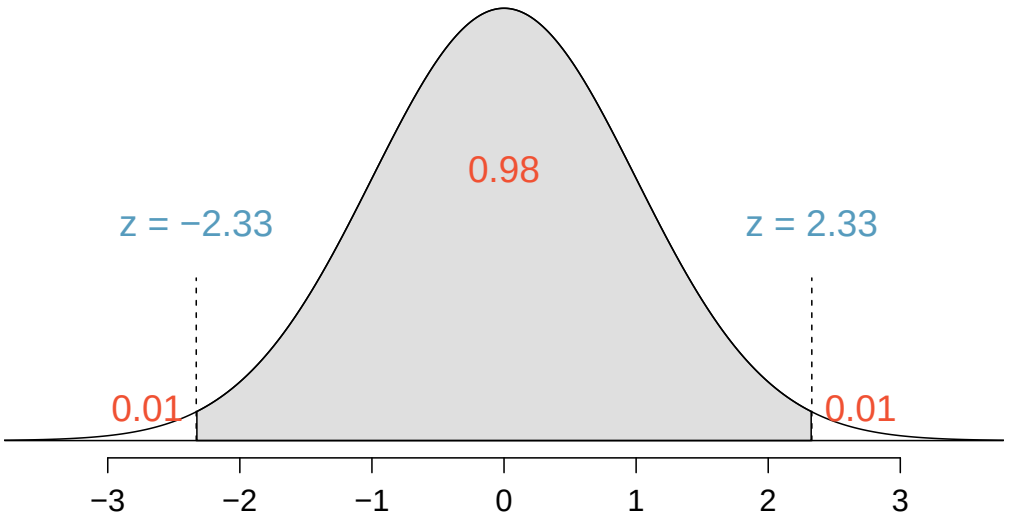
(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$



98%信頼区間を計算する際の適切な $z^\star$ はどの Z 値か？

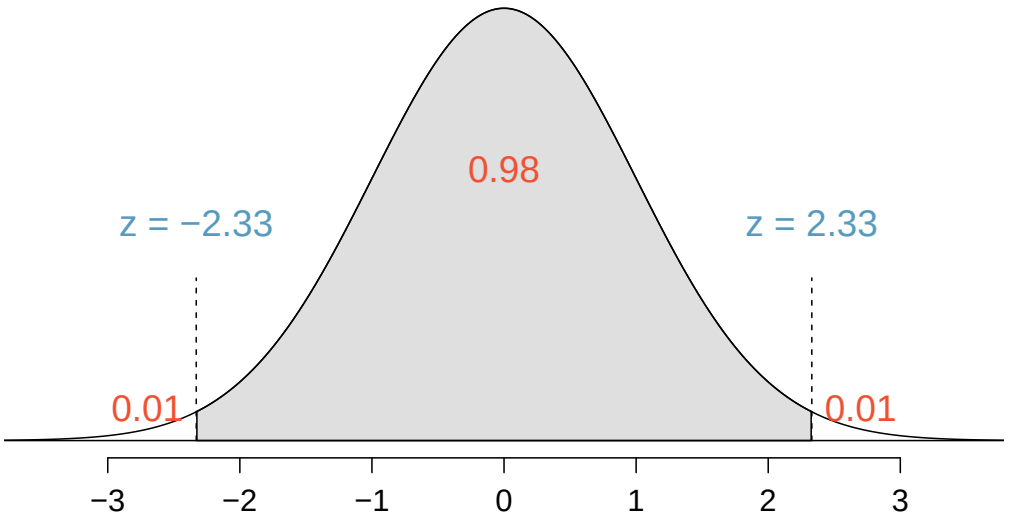
(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$



信頼区間の解釈

信頼区間は…

- 常に母集団についてのものである
- 確率の陳述ではない
- 個々の観測値ではなく、母集団パラメータのみについてのものである
- 基となる標本統計量が母集団パラメータの不偏推定量である場合にのみ信頼できる

思い出してみよう…

性差別実験：

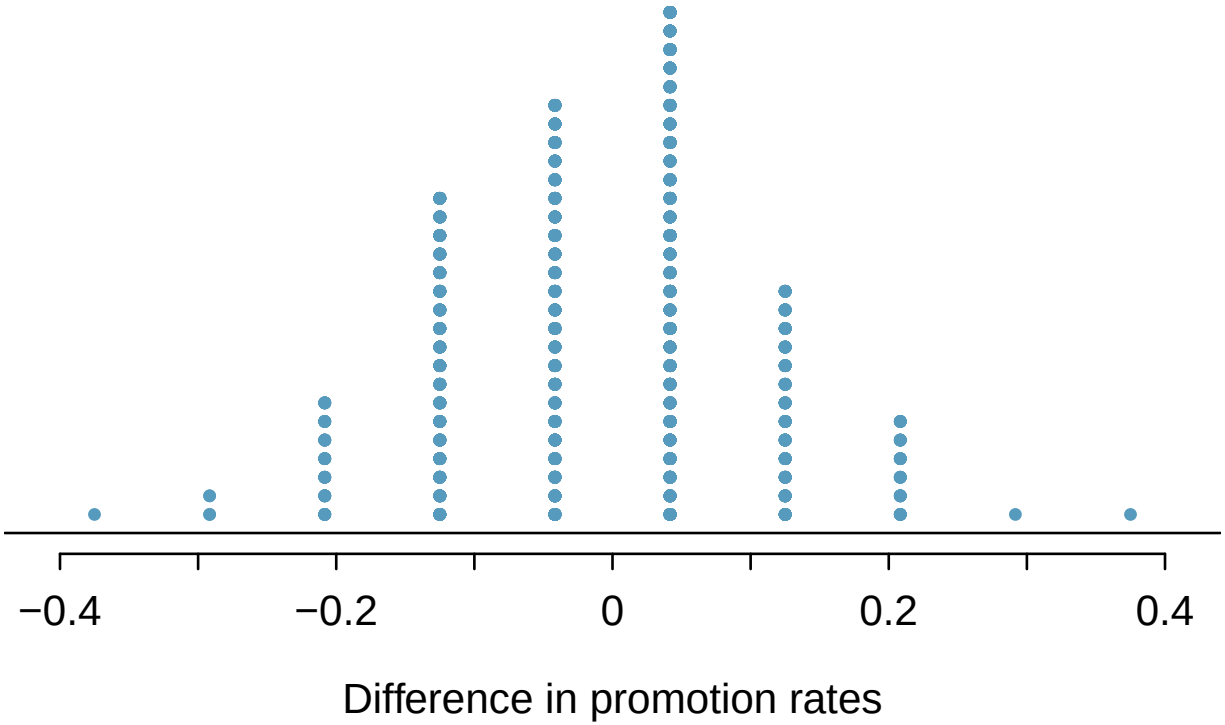
		昇進		
		昇進あり	昇進なし	合計
性別	男性	21	3	24
	女性	14	10	24
	合計	35	13	48

$$\hat{p}_{\text{男性}} = 21/24 \approx 0.88 \quad \text{および} \quad \hat{p}_{\text{女性}} = 14/24 \approx 0.58$$

考えられる説明：

- 昇進と性別は独立であり、性差別はなく、割合の差は単なる偶然による。$\rightarrow$ 帰無仮説 -（何も起きていない）
- 昇進と性別は従属であり、性差別があり、割合の差は偶然ではない。$\rightarrow$ 対立仮説 -（何かが起きている）

結果



シミュレーションにおいて実際のデータと同等またはそれ以上の極端な結果（男性の昇進率が女性より 30%以上高い）が得られる可能性は非常に低かったため、帰無仮説を棄却して対立仮説を採択することにした。

まとめ：仮説検定の枠組み

- 現状を表す帰無仮説（H_0）から始める。
- また、研究上の問い（検定したいこと）を表す対立仮説（H_A）も設定する。
- 帰無仮説が真であるという前提のもとで、シミュレーションまたは中心極限定理に基づく伝統的な手法によって仮説検定を行う（次に説明する）。
- 検定結果が、対立仮説を支持する説得力のある証拠がデータに含まれないことを示す場合は、帰無仮説を維持する。そうでない場合は、帰無仮説を棄却して対立仮説を採択する。

割合に関する主張を検定する例を使って、仮説検定の枠組みを正式に紹介する。

信頼区間を用いた仮説検定

先ほど、Facebook が自分の興味を正確に分類していると思うアメリカの Facebook ユーザーの割合に対する 95%信頼区間を 64%から 67%と計算した。この信頼区間に基づいて、アメリカの Facebook ユーザーの過半数が Facebook による興味の分類は正確だと思っているという仮説をデータは支持するか？

- 対応する仮説は次の通りである：
  - H_0: $p = 0.50$：アメリカの Facebook ユーザーの 50%が Facebook による興味の分類は正確だと思っている
  - H_A: $p > 0.50$：アメリカの Facebook ユーザーの 50%以上が Facebook による興味の分類は正確だと思っている
- 帰無値が区間に含まれていない $\rightarrow$ 帰無仮説を棄却する。
- これは仮説検定の簡便な方法だが、帰無仮説のもとでの特定の結果の可能性（p 値）は示さない。

決定誤差

- 仮説検定は完璧ではない。
- 司法制度では、無実の人が誤って有罪判決を受けたり、有罪の人が無罪放免になったりすることがある。
- 同様に、統計的仮説検定でも誤った決定をすることがある。
- 違いは、統計学では誤りを犯す頻度を定量化するための道具があることである。

決定誤差（続き）

帰無仮説と対立仮説という２つの競合する仮説がある。仮説検定では、どちらが真である可能性が高いかを決定するが、その選択が誤っている場合がある。

		決定	
		H_0 を棄却しない	H_0 を棄却する
真実	H_0 が真	✓	第一種の過誤
	H_A が真	第二種の過誤	✓

- 第一種の過誤とは、H_0 が真であるときに帰無仮説を棄却することである。
- 第二種の過誤とは、H_A が真であるときに帰無仮説を棄却しないことである。
- H_0 または H_A のどちらが真であるかは（ほぼ）わからないが、すべての可能性を考慮する必要がある。

仮説検定を裁判として考える

仮説検定を刑事裁判として考えると、帰無仮説と対立仮説の観点から評決を組み立てることが理にかなっている：

$$H_0: \text{被告は無実である}$$
$$H_A: \text{被告は有罪である}$$

次の状況ではどの種類の誤りが犯されているか？

- 実際には有罪の被告を無実と宣告する
  第二種の過誤
- 実際には無実の被告を有罪と宣告する
  第一種の過誤

どちらの誤りがより深刻だと思うか？

「10 人の有罪者を逃がすことは、1 人の無実の人を苦しめることより好ましい」
——ウィリアム・ブラックストーン

第一種の過誤の確率

- 一般的なルールとして、p 値が 0.05 未満のとき H_0 を棄却する。すなわち、有意水準 0.05、$\alpha = 0.05$ を使用する。
- これは、H_0 が実際に真である場合に、5%を超えて誤って棄却したくないということを意味する。
- 言い換えると、5%の有意水準を使用する場合、帰無仮説が真であれば第一種の過誤を犯す確率は約 5%である。

$$P(\text{第一種の過誤} \mid H_0 \text{ が真}) = \alpha$$

- これが α の小さい値を好む理由である。α を大きくすると第一種の過誤の確率が高まる。

Facebook の興味カテゴリー

同じ調査で 850 人の回答者に対し、Facebook が自分のカテゴリーリストを作成することにどの程度快適に感じるかを尋ねた。回答者の 41%が快適だと答えた。Facebook が自分の興味カテゴリーリストを作成することに快適に感じるアメリカの Facebook ユーザーの割合が 50%と異なるという説得力のある証拠をこのデータは提供しているか？

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

仮説の設定

- 関心のあるパラメータは、Facebook が自分の興味カテゴリーを作成することに快適に感じるすべてのアメリカの Facebook ユーザーの割合である。
- 標本割合が 0.50（少数派）より低い理由として、2 つの説明が考えられる：
  - 真の母集団割合が 0.50 と異なる。
  - 真の母集団割合は 0.50 であり、真の母集団割合と標本割合の差は単なる自然な標本変動性によるものである。

仮説の設定

- アメリカの Facebook ユーザーの 50%が Facebook による興味カテゴリーの作成に快適に感じているという仮定から始める：

$$H_0 : p = 0.50$$

- Facebook による興味カテゴリーの作成に快適に感じるアメリカの Facebook ユーザーの割合が 50%と異なるという主張を検定する：

$$H_A : p \neq 0.50$$

Facebook の興味カテゴリー — 条件

この仮説検定を進めるために満たす必要のある条件として、次のうち正しくないものはどれか？

(a) 標本の回答者は、Facebook による興味分類に快適かどうかという点で互いに独立していること。

(b) 抽出は無作為に行われていること。

(c) 標本サイズはアメリカの Facebook ユーザー全体の母集団の 10%未満であること。

(d) 標本には少なくとも 30 人の回答者がいること。

(e) 期待される成功数と失敗数がそれぞれ少なくとも 10 以上あること。

Facebook の興味カテゴリー — 条件

この仮説検定を進めるために満たす必要のある条件として、次のうち正しくないものはどれか？

(a) 標本の回答者は、Facebook による興味分類に快適かどうかという点で互いに独立していること。

(b) 抽出は無作為に行われていること。

(c) 標本サイズはアメリカの Facebook ユーザー全体の母集団の 10%未満であること。

(d) 標本には少なくとも *30* 人の回答者がいること。

(e) 期待される成功数と失敗数がそれぞれ少なくとも 10 以上あること。

検定統計量

観察された標本割合が仮説上の標本分布に対してどれほど異常かを評価するために、帰無仮説から何標準誤差離れているかを求める。これを検定統計量と呼ぶ。

$$\hat{p} \sim N\left(\mu = 0.50, SE = \sqrt{\frac{0.50 \times 0.50}{850}}\right)$$

$$Z = \frac{0.41 - 0.50}{0.0171} = -5.26$$

標本割合は仮説値から 5.26 標準誤差離れている。これは異常に低いと考えられるか？ つまり、この結果は統計的に有意か？

はい。そして、p 値を使ってどれほど異常かを定量化できる。

p 値

- この検定統計量を使って p 値を計算する。p 値とは、帰無仮説が真であった場合に、現在のデータセットと同等かそれ以上に対立仮説に有利なデータを観察する確率である。
- p 値が低い場合（有意水準 α より低い場合、通常は 5%）、帰無仮説が真であればこのデータを観察することは非常にありそうにないと判断し、H_0 を棄却する。
- p 値が高い場合（α より高い場合）、帰無仮説が真であってもこのデータを観察することは十分ありそうだと判断し、H_0 を棄却しない。

Facebook の興味カテゴリー — p 値

p 値：帰無仮説が真であった場合（真の母集団割合が 0.50 であった場合）に、現在のデータセット（標本割合が 0.41 未満）と同等かそれ以上に H_A に有利なデータを観察する確率。

$$P(\hat{p} < 0.41 \quad \text{または} \quad \hat{p} > 0.59 \mid p = 0.50)$$
$$= P(|Z| > 5.26) < 0.0001$$

Facebook の興味カテゴリー — 決定

- p 値 < 0.0001
  - アメリカの Facebook ユーザーの 50%がこれらの興味カテゴリーの作成に快適であるならば、850 人のアメリカの Facebook ユーザーの無作為標本において 41%以下または 59%以上が快適と感じるという観察結果が得られる確率は 0.01%未満である。
  - 観察された標本割合またはそれ以上に極端な結果が偶然に起こる可能性は非常に低い。
- p 値が低い（5%未満）ため、H_0 を棄却する。
- Facebook による興味カテゴリーリストの作成に快適に感じるアメリカの Facebook ユーザーの割合が 50%と異なるという説得力のある証拠がデータに含まれている。
- 帰無値 0.50 と観察された標本割合 0.41 の差は偶然または標本変動性によるものではない。

有意水準の選択

- 伝統的な水準は 0.05 であるが、応用に応じて有意水準を調整することが有用である。
- 検定から得られる結論の結果に応じて、0.05 より小さいまたは大きい水準を選択する。
- 第一種の過誤が危険または特にコストが高い場合は、小さい有意水準（例：0.01）を選択すべきである。このような場合、帰無仮説を棄却することに非常に慎重になりたいため、H_0 を棄却する前に H_A を支持する非常に強力な証拠を求める。
- 第二種の過誤が第一種の過誤よりも比較的危険またはコストが高い場合は、より高い有意水準（例：0.10）を選択すべきである。この場合、帰無仮説が実際に偽であるときに H_0 を棄却しないことに注意が必要である。

片側対両側仮説検定

- 両側仮説検定では、p が帰無値 p_0 より上か下かのいずれかに関心がある：$H_A : p \neq p_0$。
- 片側仮説検定では、p が帰無値 p_0 から一方向にずれることに関心がある：
  - 母集団パラメータが p_0 より小さいかどうかを検出することにのみ価値がある場合：$H_A : p < p_0$。
  - 母集団パラメータが p_0 より大きいかどうかを検出することにのみ価値がある場合：$H_A : p > p_0$。
- 両側検定はしばしばより適切である。なぜなら、データが対立仮説と逆方向に明確に向いている場合も検出したいことが多いからである。