

MAT220 第13講：推定についてもう少し

目次

1	推定の復習	2
1.1	復習： σ が既知の場合の μ の推定	2
2	σ が未知の場合の推定	3
2.1	既知の σ を使う方法が破綻する場所	3
2.2	σ を s で置き換える	4
2.3	Student の t 分布	4
2.4	t の臨界値	6
2.5	σ が未知の場合の信頼区間	7
2.6	例題： σ が未知の場合のネジ	8
3	二項比率の推定	8
3.1	標本比率の定義	8
3.2	標本分布の定義	9
3.3	p の信頼区間	10
3.4	方法	11
4	例題：バスケットボールのスカウト	11
4.1	例1：95回中65本成功	12
4.2	例2：1000回中773本成功	12

第12講では、推定という概念を導入し、ある標本平均 $\bar{x}$ が与えられたときに、母平均 μ の値を推測するためにそれを使用しました。その際、真の母集団標準偏差である σ が何らかの理由ですでに分かっている、という仮定を意図的に置きました。この仮定のおかげで、関係する概念が単純化され、その時点では非常に役に立ちました。しかし、現実世界では、 μ が分からないなら、おそらく σ も分からないはずだと考えるのが自然です。この講義では、 σ が既知であると仮定せずに推定を行う考え方を調べます。これを二つの設定で行います。第12講と同じように μ を推定する場合と、二項実験における成功確率 p を推定する場合です。

1 推定の復習

前回の講義では、真の母平均 μ を推定するために信頼区間という概念を導入しました。主な考え方は、標本平均 $\bar{x}$ を取り、中心極限定理を使って $\bar{X}$ の標本分布

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

を構成することでした。この分布は真の母平均 μ をちょうど中心に持つので、 $\bar{x}$ と μ を関係づけるために使うことができます。そこから、 $\bar{x}$ の周りに小さな区間を作りました。この区間は、構成上、ある割合で真の平均を含むことになります。

このような推定がよくある現実的な使い道は、品質管理に現れます。説明のため、たくさんのネジを製造する工場があるとしましょう。これらのネジは、設計された部品にきちんと収まるために、およそ 20 mm の長さである必要があり、どちらの方向にも 0.1 mm の許容誤差があります。つまり、例えば 19.9 mm のネジは許容されますが、長さ 19.89 mm のネジは許容されません。

原理的には、工場が製造するすべてのネジを一つずつ検査し、その長さを測定すれば、ネジの高い品質を保証できます。そうすれば、これらすべての数値を足し合わせ、ネジの総数で割ることで、真の平均長さを得ることができます。しかし実際には、これはあまりにも時間がかかり、現実的には不可能です。そこで、このような状況では、代わりにいくつかの標本データを取り、ネジの真の平均長さを単に推定します。この数値が 19.9 から 20.1 の範囲内にあれば、そのネジは許容できると判断します。これが品質管理の考え方です。

品質管理

品質管理では、標本データを使って、生産工程が許容できる状態で動いているかどうかを判断します。すべての製品を一つずつ検査するのは非現実的なことが多いので、代わりに統計的な推定を使う方が簡単で安価です。

1.1 復習： σ が既知の場合の μ の推定

工場の例を続けましょう。長期的な研究により、母集団標準偏差が既知であり、その正確な値が

$$\sigma = 1.2 \text{ mm}$$

であることが分かっているとします。また、品質管理担当者が $n = 50$ 本のネジからなる単純ランダム標本を取り、標本平均が $\bar{x} = 19.98 \text{ mm}$ であったとしましょう。

ここで、真の平均長さ μ に対する 95% 信頼区間を構成します。第12講より、 σ が既知の場合の μ の信頼区間は、誤差の限界 E から計算できます。観測された $\bar{x}$ の値を中心に区間を置くと、信頼区間は次のようになります。

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \right].$$

95% 信頼区間では、第12講の表より $z_c = 1.96$ です。したがって、誤差の限界は次のように計算できます。

$$E = z_c \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{1.2}{\sqrt{50}} \approx 0.333.$$

したがって、観測された標本平均 $\bar{x} = 19.98$ を中心とする 95% 信頼区間は

$$CI = [19.98 - 0.333, 19.98 + 0.333] = [19.647, 20.313].$$

解釈

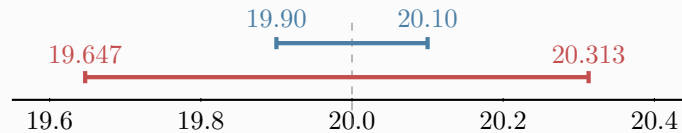
設定によると、平均長さとして許容される範囲は

$$[19.90, 20.10].$$

です。しかし、私たちが求めた 95% 信頼区間は

$$[19.647, 20.313].$$

です。この区間は、要求されている工学的許容範囲よりもかなり広がっています。

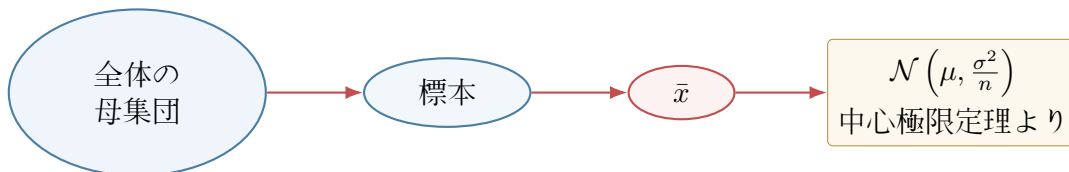


したがって、この標本は、真の平均が必要な許容範囲内にあるという強い証拠を与えていません。しかし、これは直ちに工場を停止すべきだという意味ではありません。むしろ、標本データはどちらの結論にも十分な説得力を持たない、ということを受け入れる必要があります。真の平均が望ましい $20 \text{ mm} \pm 0.1 \text{ mm}$ の外にあると結論するだけの十分な証拠もなく、またその範囲の中にあると結論するだけの十分な証拠もないのです。

2 σ が未知の場合の推定

2.1 既知の σ を使う方法が破綻する場所

前の例では、次の手順を使いました。



そこから、

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

を使って誤差の限界を計算し、 μ を推定しました。

しかし、多くの現実世界の状況では、 σ を単純に知らないのです。つまり、上の公式にその数値を代入することができず、したがってこの方法で μ を推定することはできません。そこで、 σ を私たちが知っている何かで置き換え、その置き換えがもたらす影響を評価する必要があります。

2.2 σ を s で置き換える

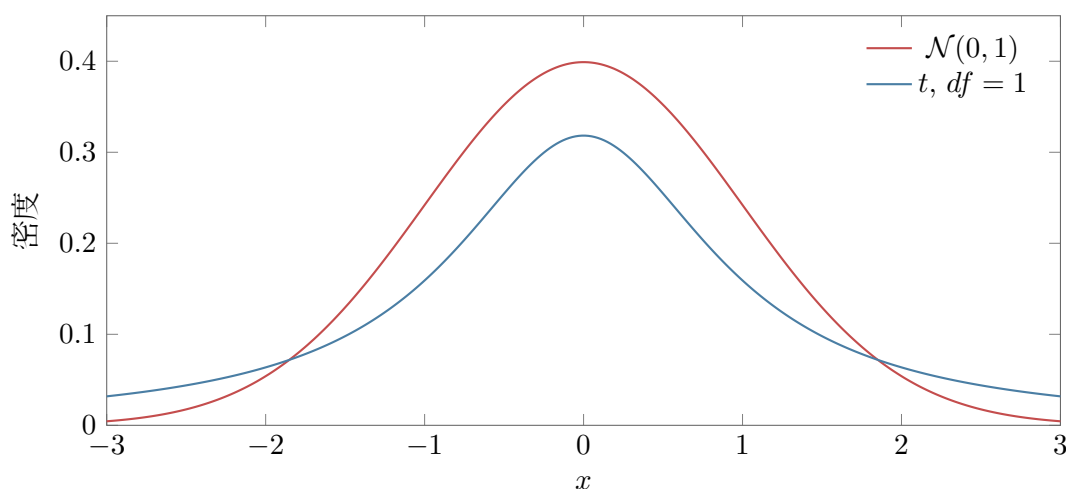
真の母集団標準偏差 σ が分からないなら、次善の策は、それを標本標準偏差で置き換えることのように見えます。

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}.$$

しかし、この置き換えを行うと、以前よりも多くの不確実性を導入することになります。したがって、推定を導くためにどの確率分布を使うかについて注意する必要があります。実際、この場合、標準正規分布を使うことはもはや適切ではありません。代わりに、Student の t 分布と呼ばれる修正版を使う必要があります。これから、この t 分布がどのようなものかを詳しく見ていきます。

2.3 Student の t 分布

Student の t 分布は、標準正規分布と関係があります。主な違いは、より多くの面積が尾部に分布していることです。簡単な比較を下に示します。



ここでは、標準正規分布が赤、 t 分布が青で描かれています。青い曲線もやはり対称であり、中心は 0 です。実際、明らかな違いは尾部だけです。つまり、尾部が太いのです。¹ これには十分な理由があります。 σ が未知であるとき、私たちはより多くの不確実性を持っているため、この分布に基づいて作られる信頼区間は、正規分布に基づく対応物よりも広くなるべきだからです。

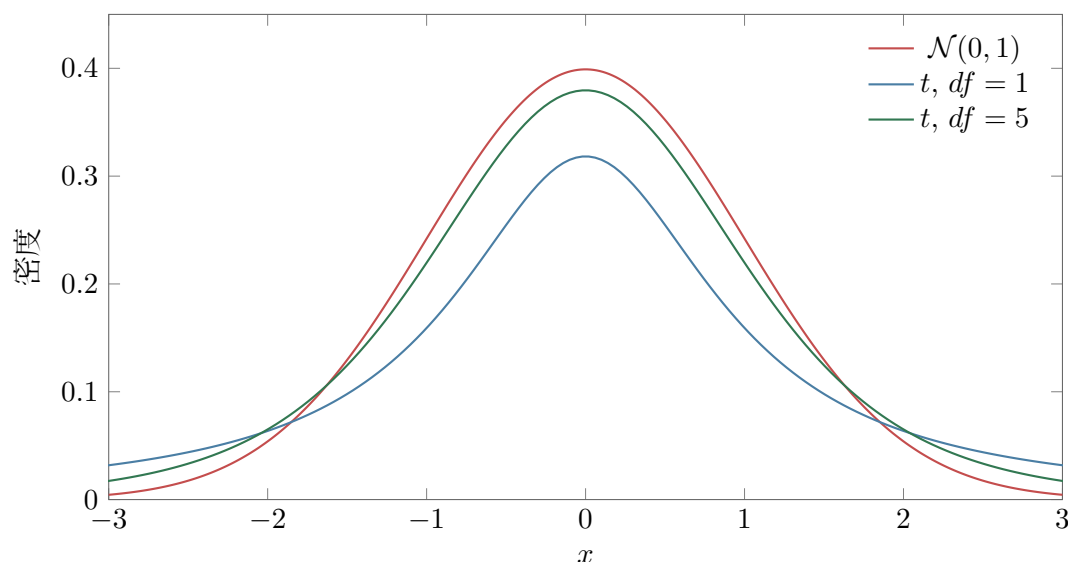
上のグラフで、 t 分布のラベルには“ df ”という追加の項があることに注意してください。実際、これは t 分布を完全に記述するために必要な追加の入力です。この追加情報は自由度と呼ばれます。だから df という文字を使います。これは単純に次で定義されます。

$$df = n - 1.$$

ここでの考え方は、この数が標本サイズとともに増加するということです。標本サイズが大きくなるにつれて、一般には標本標準偏差 s はランダムなノイズの影響を受けにくくなり、した

¹これに対する専門用語は尖度です。これは分布の尾部がどれほど「重い」かを表す尺度です。

がって真の母集団パラメータ σ のより良い推定値になると期待されます。 n が増えるにつれて s が σ に近づくなれば、 t 分布の尾部に組み込まれた不確実性は減少するはずだと考えられます。 n が非常に大きくなると、標本標準偏差 s は σ に非常に近くなり、したがって標準正規分布 $\mathcal{N}(0,1)$ と非常によく似た t 分布を期待できます。実際、その通りのことが起こります。 df が増えるにつれて、 t 分布は標準正規分布により近くなります。



df がどんどん大きくなるにつれて、得られる t 分布は正規分布にますます似ていきます。これは、標準正規分布を記述する別の方法を与えます。すなわち、標準正規分布は、 $df \rightarrow \infty$ としたときの標準 t 分布の極限です。

Student の t 分布の性質

Student の t 分布には、次の性質があります。

1. 0 について対称である。
2. 曲線の下での総面積は 1 である。
3. $\mathcal{N}(0,1)$ よりも大きな尾部を持つ。
4. 追加の入力が一つ必要である。

$$df = n - 1.$$

5. df が大きくなるにつれて、 t 分布は $\mathcal{N}(0,1)$ に近づく。

2.4 t の臨界値

Student の t 統計量

もとの母集団が平均 μ の正規分布であると仮定します。サイズ n の標本について、標本平均を $\bar{X}$ 、標本標準偏差を S とすると、^a 統計量

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

は、自由度

$$df = n - 1$$

の Student の t 分布に従います。

^aここでは、標本標準偏差をそれ自体一つの確率変数へと格上げしています。

この式は、 σ を s で置き換えたことを除けば、 z スコアの公式とほとんど同じ形をしています。観測された標本が得られると、公式は次のように変わります。

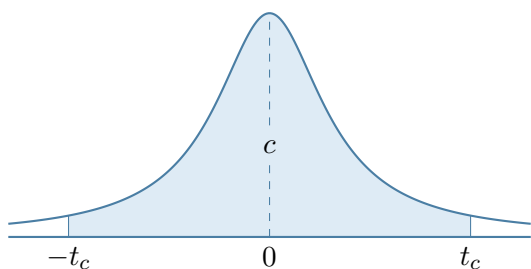
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{未知の } \sigma} t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

σ が既知のときは、標準正規分布から臨界値 z_c を使いました。幾何学的には、これは標準正規分布の中で面積がちょうど c になる領域に対応していました。そして、 z スコアの公式を使ってこの領域を標本分布 $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ に戻すことで、 μ を推定する信頼区間を構成しました。

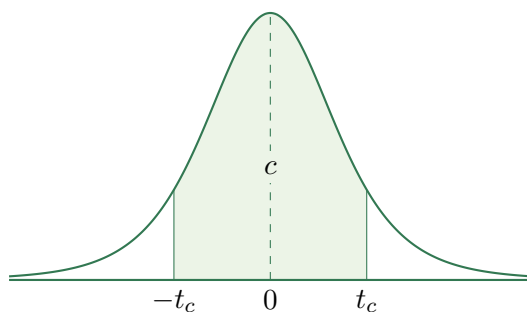
これに対して、 σ が未知のときは、Student の t 分布から臨界値を使います。臨界値 z_c との類推で、 t 分布における臨界値を t_c と書くことにします。幾何学的には、これらの臨界値はまったく同じ役割を果たします。つまり、原点を中心とし、面積がちょうど c になる t 分布内の領域の端点を与えます。 t 分布は自由度 df に依存するので、 t_c の正確な値は次の二つの情報に依存します。

1. 信頼水準 c 、
2. 自由度 $df = n - 1$ 。

例えば、先ほど描いた $df = 1$ と $df = 5$ の t 分布を考えましょう。この二つの曲線は形が異なるので、臨界値 t_c も変わります。



$df = 1$ のときの面積 c



$df = 5$ のときの面積 c

見て分かるように、 df が小さいとき、 t_c の臨界値は大きくなります。よく使われる t_c の値の表を下に示します。対応する z_c の表よりもかなり複雑であることに注意してください。

df	80%	90%	95%	98%	99%	99.5%
1	3.078	6.314	12.706	31.821	63.657	127.321
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
12	1.356	1.782	2.179	2.681	3.055	3.428
15	1.341	1.753	2.131	2.602	2.947	3.286
20	1.325	1.725	2.086	2.528	2.845	3.153
25	1.316	1.708	2.060	2.485	2.787	3.078
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
120	1.289	1.658	1.980	2.358	2.617	2.860
∞	1.282	1.645	1.960	2.326	2.576	2.807

ここで、各列は信頼水準 c のよく使われる値に対応しています。各行は、よく使われる自由度に合った t_c の値を与えています。最下行は、 df が任意に大きくなる場合を表しています。この場合、 t_c の値は z_c の値にどんどん近づき、最終的には極限でそれに近づきます。

2.5 σ が未知の場合の信頼区間

臨界値 t_c が使えるようになったので、通常の誤差の限界の導出に従って、観測値 $\bar{x}$ の周りの信頼区間を記述します。この場合、通常の z スコアの公式ではなく、 t 統計量

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

を使う必要があります。臨界値 $\pm t_c$ を使って解くと、誤差の限界は

$$E = t_c \frac{s}{\sqrt{n}}.$$

となります。 σ が既知の場合との明確な類似性に注意してください。ここでは単に σ を s で置き換え、通常の z_c の代わりに t 分布の臨界値を使っています。観測された標本平均 $\bar{x}$ の周りの信頼区間は、次のように記述できます。

σ が未知の場合の μ の信頼区間

母集団標準偏差 σ が未知の場合、標本標準偏差 s を使います。 μ の信頼区間は次のようになります。

$$[\bar{x} - E, \bar{x} + E] = \left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}} \right],$$

ただし $df = n - 1$ です。

既知の場合と未知の場合の主な違いを下にまとめます。

	σ 既知	σ 未知
スコア	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
臨界値	z_c	t_c
誤差の限界	$E = z_c \frac{\sigma}{\sqrt{n}}$	$E = t_c \frac{s}{\sqrt{n}}$
信頼区間	$[\bar{x} - E, \bar{x} + E]$	$[\bar{x} - E, \bar{x} + E]$

2.6 例題： σ が未知の場合のネジ

第1節の工場の例に戻りましょう。今度は、工場が真の母集団標準偏差 σ を知らないとします。代わりに、50 本のネジからなる標本が次を与えたとします。

$$n = 50, \quad \bar{x} = 19.98, \quad s = 1.1 \text{ mm.}$$

真の平均長さ μ に対する 95% 信頼区間を構成したいなら、 σ の代わりに s を使い、 z スコアの代わりに t 統計量を使う必要があります。 $n = 50$ なので、自由度は $df = n - 1 = 49$ です。 t 表より、 $df = 49$ の 95% 信頼区間に対する臨界値は、およそ

$$t_c \approx 2.010.$$

です。 $s = 1.1$ を使うと、誤差の限界は次のように計算できます。

$$E = t_c \frac{s}{\sqrt{n}} = 2.010 \cdot \frac{1.1}{\sqrt{50}} \approx 0.313.$$

したがって、観測された $\bar{x}$ の値を中心とする 95% 信頼区間は

$$\text{CI} = [19.98 - 0.313, 19.98 + 0.313] = [19.667, 20.293].$$

ここでも、前の例と同じ解釈になります。

3 二項比率の推定

3.1 標本比率の定義

二項実験の一部を観察するとします。たまたま n 回の試行を観察し、そのうち r 回の成功を数えたとします。標本内のランダムな変動のため、これは真の成功確率 p の値を推論するには十分な情報ではありません。しかし、この情報を使って p の値を推定することはできます。前節および第12講と同様に、標本データの周りに信頼区間を作り、推定値の周りに適切な誤差範囲を与えます。まず必要なのは、意味のある標本統計量を得ることです。

標本比率

n 回の二項試行で r 回の成功を観測した場合、標本比率を次のように定義します。

$$\hat{p} = \frac{r}{n}.$$

これにより、 $\hat{q} = 1 - \hat{p}$ も定義します。

量 $\hat{p}$ は、真の成功確率 p に対する点推定値になります。

3.2 標本分布の定義

第11講で二項分布の正規近似を議論したとき、二項実験内の成功回数を数える通常の二項確率変数 X は、 Y_i の和に分解できると述べました。ここで Y_i は、実験の各個別試行に対して定義される確率変数です。

$$Y_i = \begin{cases} 1 & \text{試行 } i \text{ が成功した場合,} \\ 0 & \text{試行 } i \text{ が失敗した場合.} \end{cases}$$

記号で書くと、第11講では次を見ました。

$$X = Y_1 + Y_2 + \cdots + Y_n$$

したがって、 X を n で割ると、標本平均のように振る舞う量が得られました。これを標本比率と結びつけます。次のように定義すれば、 $\hat{P}$ と書かれる確率変数を導入できます。

$$\hat{P} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

実際、これは第11講で名前を付けずに行っていたことそのものです。このように書くと、 $\hat{P}$ は確率変数であり、その観測値 $\hat{p}$ はサイズ n の標本から得られる標本平均のように振る舞うことが分かります。上の観測値 r は、 n 個の個別の確率変数 Y_i の結果の和であり、それを n で割って平均を得ています。

第11講で、これらの個別の確率変数 Y_i はそれぞれ

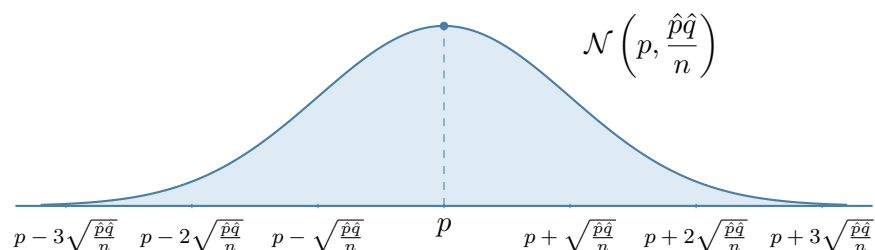
$$E(Y_i) = p \quad \text{かつ} \quad \text{Var}(Y_i) = pq$$

を持つことを見ました。中心極限定理は、すべての標本平均の集合が近似的な正規分布に従うことを教えます。その正規分布の中心は、標本を作るもとの変数の平均であり、新しい分散は、もとの変数の分散を標本サイズ n で割ったものです。この場合、中心極限定理は、 $\hat{P}$ が次のパラメータを持つ近似正規分布に従うことを教えます。

$$\mathcal{N}\left(p, \frac{pq}{n}\right).$$

これは、まさに必要なものを与えているように見えます。つまり、標本比率が近似正規分布を持ち、その真の平均が中心にあることを確認できたからです。ここから信頼区間を作り始められそうです。しかし、根本的な問題が一つあります。私たちは真の成功確率 p を推定しようとしています。 p の正確な値が分からないなら、 $\frac{pq}{n}$ の値もまた分かりません。言い換えると、このまま進めて何らかの誤差の限界を作ろうとした場合、 p の値をすでに知らずに、どうやってそれを実際に計算すればよいのでしょうか。

ここで、 $\hat{P}$ の真の標本分布は私たちの目的には使えないことを受け入れなければなりません。代わりに次善の策を取ります。つまり、分散の項を点推定値で置き換えます。実際には、 p と q を観測値 $\hat{p}$ と $\hat{q}$ で置き換え、次の近似を得ます。



これが私たちの推定の設定になります。

$\hat{P}$ の標本分布

二項実験に対して、標本比率は近似的に正規分布に従います。

$$\hat{P} \approx \mathcal{N}\left(p, \frac{\hat{p}\hat{q}}{n}\right).$$

この近似は、次が成り立つとき妥当です。

$$n\hat{p} > 5 \quad \text{かつ} \quad n\hat{q} > 5.$$

3.3 p の信頼区間

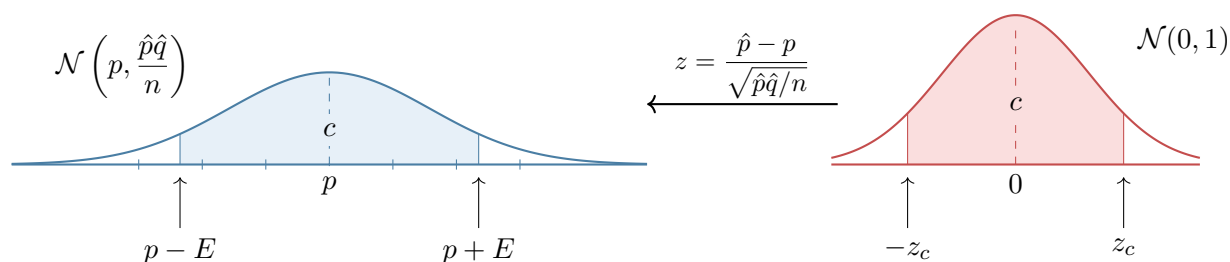
ここからは、以前とまったく同じように信頼区間を構成します。推定された標準誤差は

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

したがって、 z スコアの公式は次のようになります。

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}.$$

信頼水準 c の信頼区間を導くには、 $\mathcal{N}(0, 1)$ の中で $P(-z_c < Z < z_c) = c$ を与える臨界値 z_c を求めます。 z スコアの公式を逆向きに使うことで、この領域を標本分布へ戻します。



誤差の限界 E は、 z_c に標準誤差を掛けたものです。この場合、

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

したがって、観測値 $\hat{p}$ の周りの信頼区間を次のように定義できます。

二項比率の信頼区間

n 回の試行で r 回の成功が与えられたとき、

$$\hat{p} = \frac{r}{n}, \quad \hat{q} = 1 - \hat{p}.$$

真の成功確率 p の信頼区間は

$$\left[\hat{p} - z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} \right].$$

3.4 方法

二項実験の成功確率 p を推定するには、次の手順に従います。

Step 1: 次を計算する。

$$\hat{p} = \frac{r}{n}, \quad \hat{q} = 1 - \hat{p}.$$

Step 2: 正規近似が妥当であることを確認する。

$$n\hat{p} > 5 \quad \text{かつ} \quad n\hat{q} > 5.$$

Step 3: 選んだ信頼水準に対する臨界値 z_c を求める。

Step 4: 誤差の限界を計算する。

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Step 5: 信頼区間を書く。

$$[\hat{p} - E, \hat{p} + E].$$

4 例題：バスケットボールのスカウト

あるバスケットボールチームは、真のフリースロー成功率が少なくとも 70% である選手だけを採用したいとします。記号で書けば、チームは $p \geq 0.70$ である証拠を求めています。

あなたはある選手をスカウトし、彼がフリースローを打つのを観察します。一日中観察することはできないので、彼のパフォーマンスから標本を取り、真のシュート成功率 p を推定するために 95% 信頼区間を計算することにします。

4.1 例1：95回中65本成功

選手が $n = 95$ 本シュートを打ち、そのうち $r = 65$ 本を決めたとします。このとき、成功確率に対する点推定値は

$$\hat{p} = \frac{65}{95} \approx 0.684,$$

したがって $\hat{q} = 1 - \hat{p} \approx 0.316$ です。先に進む前に、近似条件を確認すべきです。次が成り立ちます。

$$n\hat{p} = 95(0.684) \approx 65, \quad \text{かつ} \quad n\hat{q} = 95(0.316) \approx 30.$$

どちらも 5 より大きいので、正規近似は妥当です。95% 信頼区間では、 $z_c = 1.96$ です。したがって、誤差の限界は

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.684)(0.316)}{95}} \approx 0.093.$$

観測値 $\hat{p} = 0.684$ の周りの 95% 信頼区間は、したがって

$$CI \approx [0.684 - 0.093, 0.684 + 0.093] = [0.591, 0.778].$$

解釈

設定によると、私たちは望ましいしきい値 $p = 0.70$ を探しています。しかし、私たちの信頼区間は

$$[0.591, 0.778].$$

です。この区間はしきい値をまたいでいます。



したがって、第1.1節と同じような状況になります。つまり、選手の真のシュート成功率が 70% より上か下かを、決定的には言えません。簡単に言えば、この標本はどちらの方向にも十分に説得的な証拠を与えていません。

4.2 例2：1000回中773本成功

最初の観察が結論を出せないものだと分かった後、あなたは再びその選手を観察しに行き、今度はずっと長い時間観察します。今度は、選手が $n = 1000$ 本のシュートを打ち、そのうち $r = 773$ 本を決めたとします。このとき、成功確率に対する点推定値は

$$\hat{p} = \frac{773}{1000} = 0.773,$$

したがって $\hat{q} = 1 - \hat{p} = 0.227$ です。ここでも、まず近似条件が成り立つことを確認します。

$$n\hat{p} = 1000(0.773) = 773, \quad \text{かつ} \quad n\hat{q} = 1000(0.227) = 227.$$

どちらも 5 より大きいので、正規近似は妥当です。再び臨界値 $z_c = 1.96$ を使い、誤差の限界を計算します。

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.773)(0.227)}{1000}} \approx 0.026.$$

観測値 $\hat{p} = 0.773$ の周りの 95% 信頼区間は

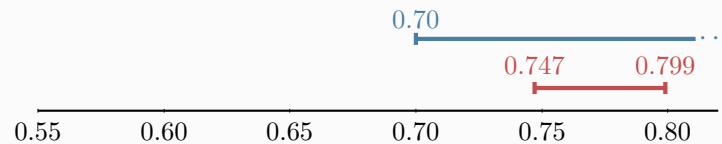
$$CI = [0.773 - 0.026, 0.773 + 0.026] = [0.747, 0.799].$$

解釈

この場合、信頼区間は

$$[0.747, 0.799].$$

です。この区間全体は、望ましいしきい値 0.70 を十分に上回っています。



この方法は真の p の値をおよそ 95% の割合で捉えるので、真の p が 0.70 を上回っているという説得力のある証拠があります。したがって、この標本は、その選手がチームの基準を満たしているという説得力のある証拠を与えています。