

MAT220 第12講：推定入門

目次

1	中心極限定理の復習	2
1.1	標本分布	2
1.2	中心極限定理	2
2	推定の考え方	3
2.1	σ が既知の場合の推定	3
2.2	標本平均の周りの区間	4
2.3	信頼水準の解釈	6
3	信頼区間の構成	6
3.1	区間についての確認	6
3.2	標本分布に対する z スコア	7
3.3	50% 信頼区間を求める	8
3.4	95% 信頼区間を求める	9
4	信頼区間の一般公式	10
4.1	臨界値	10
4.2	一般の信頼区間	11
4.3	誤差の限界	12

この講義では、正規分布の理論を推定に使います。基本的な問いは単純です。母集団全体を測定できない場合でも、 μ のような未知の母集団パラメータについて、妥当な推測をすることはできるでしょうか。これから見るように、答えは「できる」です。ここでは当面、母集団標準偏差 σ がすでに分かっていると仮定します。そうすることで、中心極限定理を使って、観測された標本データを真の母平均に結びつけることができます。そして、正規分布の仕組みを使って、標本平均 $\bar{x}$ の周りに許容できる誤差範囲を構成することで、 μ の値に対する推定値を作ります。この構成を信頼区間と呼びます。

1 中心極限定理の復習

1.1 標本分布

平均 μ 、標準偏差 σ を持つ何らかの母集団分布があるとします。サイズ n の単純ランダム標本を取ると、その標本平均を計算できます。

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

この過程を何度も繰り返すと、たくさんの可能な標本平均が得られます。これらすべての標本平均を集めたものは、新しい分布を形成します。これを標本分布と呼びます。

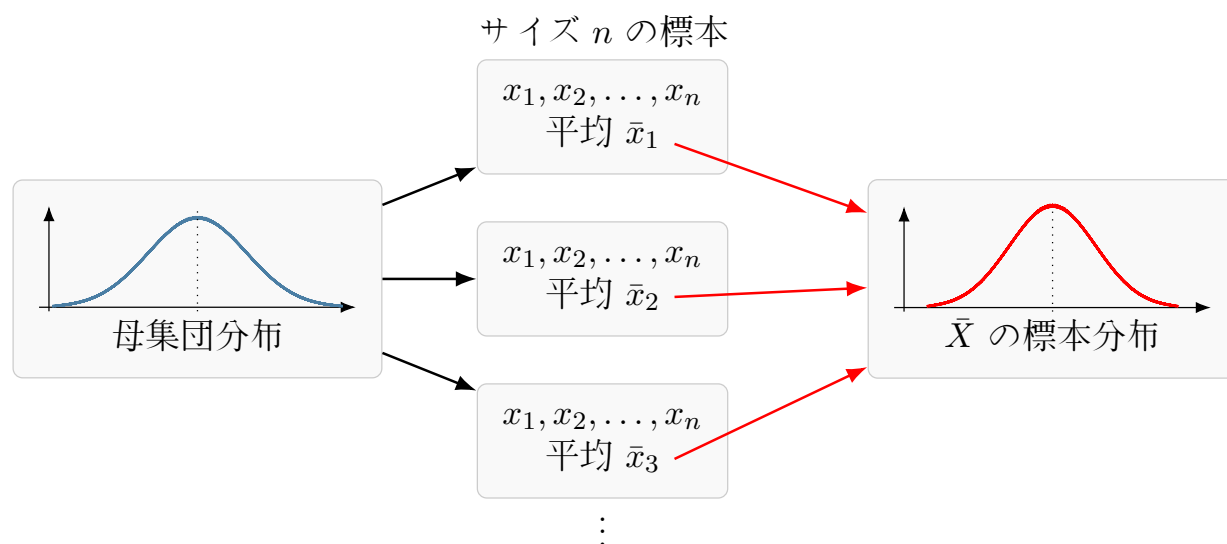
標本平均の標本分布

ある確率分布が与えられたとき、サイズ n の標本に対する $\bar{X}$ の標本分布とは、サイズ n のすべての単純ランダム標本から得られるすべての標本平均によって作られる確率分布のことです。

ここで重要なのは、標本がランダムに取られるという点です。これは、 $\bar{x}$ の正確な値もランダムになることを意味します。特定の標本を取る前には、実際にどの値が得られるかは分かりません。したがって、 $\bar{x}$ の正確な値も分かりません。そこで、標本平均を一つの確率変数へと格上げし、 $\bar{X}$ と書きます。これは、母集団から取り得るすべてのランダム標本に応じて、標本平均が取り得るすべての値を表すための記号です。

1.2 中心極限定理

一般に、中心極限定理は、 $\bar{X}$ の標本分布がどのような形になるかを教えてくれる数学的な結果です。もとの母集団が正規分布であれば、 $\bar{X}$ の標本分布は正確に正規分布になります。



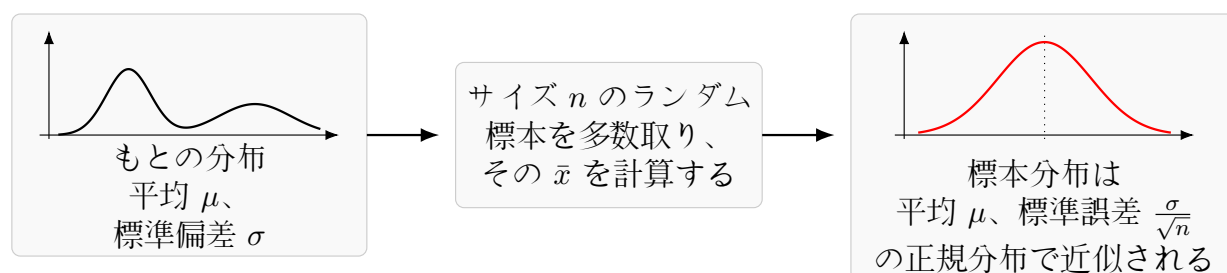
さらに、もとの母集団が正規分布でない場合でも、 n が十分大きければ、中心極限定理は、得られる $\bar{X}$ の標本分布が近似的に正規分布になると述べています。形式的には、中心極限定理の一つの述べ方は次の通りです。

中心極限定理

母集団が平均 μ 、標準偏差 σ を持つとき、十分大きな n に対して、 $\bar{X}$ の標本分布は次を満たす。

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

これは、標本分布が真の母平均 μ を中心としているが、その広がり（分散）はもとの母集団の広がりより小さい、という意味です。標本サイズ n が大きいほど、標本平均の広がり（分散）は小さくなります。



中心極限定理の記述では、標本分布の分散は特定の量 $\frac{\sigma^2}{n}$ に等しくなります。この平方根である $\frac{\sigma}{\sqrt{n}}$ は、標本分布の標準偏差です。これをもとの母集団の標準偏差 σ と混同しないように、別の用語を使います。

標準誤差

$\bar{X}$ の標本分布の標準偏差は

$$\frac{\sigma}{\sqrt{n}}.$$

これを**標準誤差**と呼びます。

2 推定の考え方

2.1 σ が既知の場合の推定

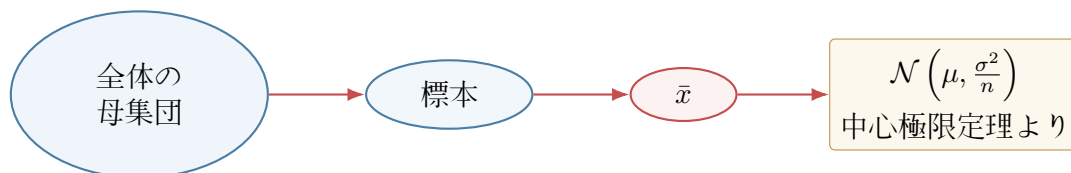
実際の統計問題では、通常、真の母平均 μ は分かりません。母集団の真の平均のようなものを知ることは非常に有用であり、確かに知りたいと思う代表的なパラメータです。しかし実際には、真の平均を計算するために、母集団のすべての対象を一つずつ測定しに行くことはできません。

この講義では、次を仮定します。

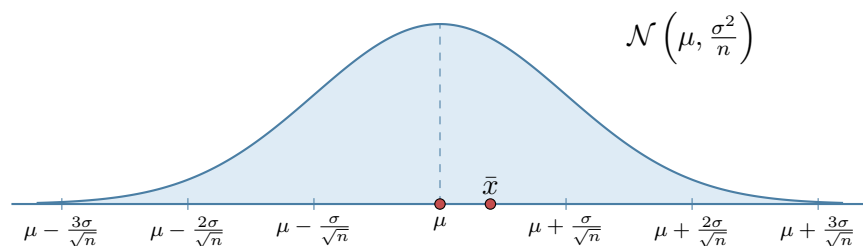
- 母平均 μ は未知であるが、
- 母集団標準偏差 σ はたまたま既知である。

これは明らかに非現実的な状況です。しかし理論的には、 μ を推定するための最も単純な設定です。次回の講義ではこの仮定を緩め、 σ も未知であるという現実的な状況で μ を推定します。

σ が既知であると仮定する利点は、中心極限定理を使って標本平均を真の母平均に結びつけられることです。



図で言えば、単純ランダム標本を取ると、中心極限定理により、観測された $\bar{x}$ の値はこの正規分布のどこかに位置することになります。

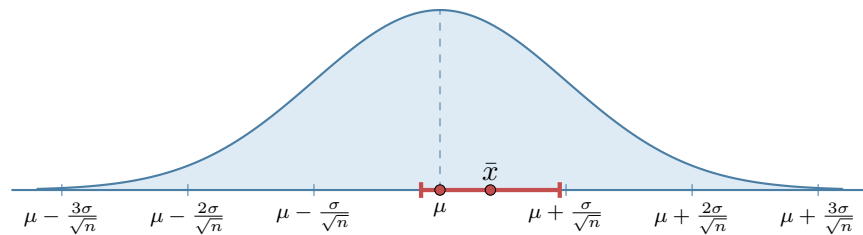


σ が既知であると仮定すると、標本分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ の構造を利用できます。例えば、 μ がどこにあるかは分かりませんが、 $\bar{X}$ は μ から一つの標準誤差以内に入る確率がおおよそ 68% であることは分かります。なぜでしょうか。 $\bar{X}$ のランダム性は分布の形に符号化されているからです。第10講の z スコア計算により、分布の面積の 68% が中心から一つの標準誤差以内にあることが分かります。

重要なのは、このような考え方は、実際に μ がどこにあるか分からなくてもなお正しいということです。中心極限定理によれば、 $\bar{X}$ のランダム性は、真の母平均を中心とする正規分布に従います。すぐ後で見えるように、 μ が何であるか分からなくても、少なくとも観測された値 $\bar{x}$ とそれを関係づけることはできます。

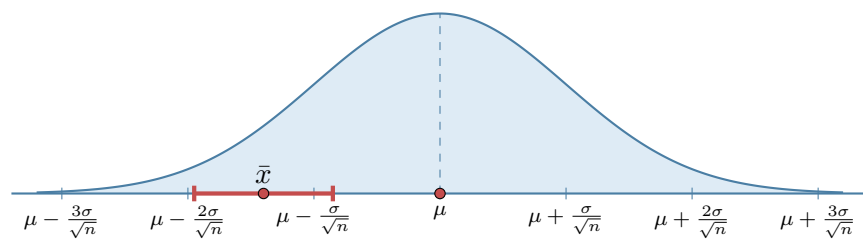
2.2 標本平均の周りの区間

ランダムに標本を選んだ場合、その結果として得られる平均 $\bar{x}$ が μ と完全に同じになる保証はありません。実際、一般には、ランダムなノイズのために、それぞれの標本平均は真の平均から少しずつずれると考えられます。しかし少なくとも、 $\bar{x}$ が μ に十分近いという意味を定義することは常にできます。 μ が何であるか分からない以上、「 μ は正確にここにある」と言う代わりに、私たちが言える最善のことは「 μ はおそらくここにあり、ある許容できる誤差範囲内にある」ということです。数学的には、この「許容できる誤差範囲」は区間として表されます。



上の図では、ランダムに選ばれた $\bar{x}$ の値の周りに、固定された長さの区間があります。 μ は $\bar{x}$ とは異なりますが、それでも $\bar{x}$ の近傍の中にあります。

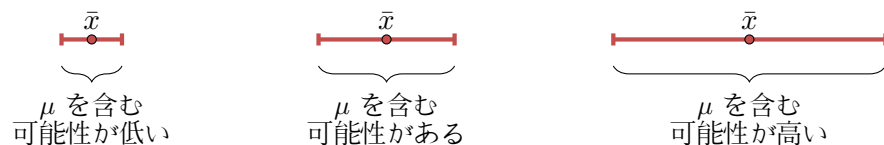
もちろん、標本が統計的に起こりにくいものである場合もあり、その場合、得られる $\bar{x}$ は中心から遠く離れることがあります。



これはランダム性の性質上、必ず起こり得ます。実際、 $\bar{X}$ のランダムな値が与えられたとき、私たちが言える最善のことは、次のようなものです。

この区間が μ を含むと、 $c\%$ の信頼度で考える。

これは結局、不確実性についての主張です。私たちは μ がどこにあるかを知っているわけではありません。その代わり、この特定の長さの区間が、ある割合で μ を捉えると言うことができます。もちろん、区間を長くすれば μ に当たる可能性は高くなり、区間を短くすれば μ を含む可能性は低くなります。



したがって、私たちの信頼の程度は、区間の長さで直接関係しているように見えます。ランダムに得られた $\bar{x}$ の周りの区間が μ を含むことに強い確信を持ちたいなら、その区間を非常に大きくすべきです。逆に、区間を非常に小さくすると、それが μ を含むという確信は弱くなります。これが信頼区間の考え方です。

信頼区間

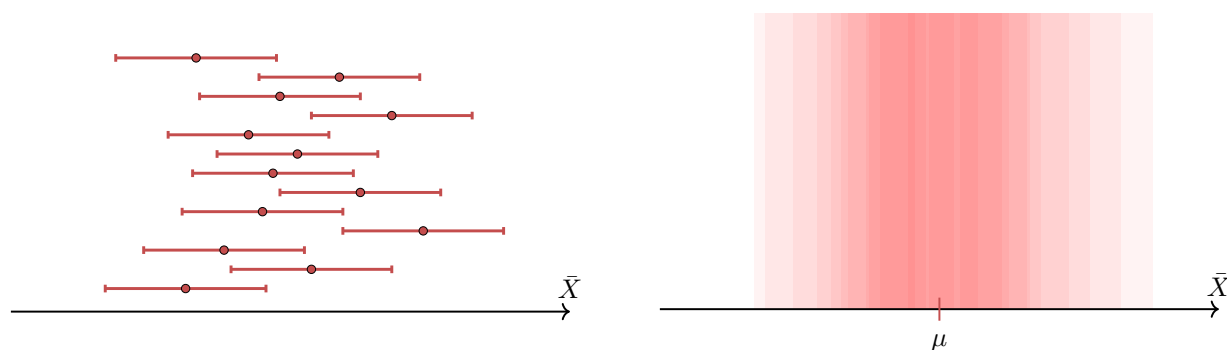
信頼区間とは、未知の母集団パラメータに対する区間推定です。信頼水準は、同じ方法を繰り返し用いたときに、その方法が真のパラメータをどれくらいの頻度で捉えるかを表します。

信頼水準はよく文字 c で書きます。ここでの考え方は、 c をパーセントとして書けるということです。そして、異なる信頼水準に対応して異なる信頼区間があります。より高い信頼度を求めるなら、 c を大きくします。すると、結果として得られる区間の長さも大きくなります。

2.3 信頼水準の解釈

説明のため、信頼水準 $c = 95\%$ の信頼区間を構成したと仮定しましょう。この構成により、この方法から作られるランダムな区間は、真の母平均 μ を 95% の割合で含むことになります。言い換えると、もし 100 個の単純ランダム標本を取り、その標本平均の周りに 95% 信頼区間を計算したなら、平均的にはそのうち 95 個の区間が μ を含み、5 個は含まないと期待されます。

下の図は、信頼水準の長期的な意味を説明しています。標本を何度も繰り返し取り続けると、得られる信頼区間のほとんどは μ を含むことになります。



右の図では、左側のさまざまな信頼区間を重ねて表示しています。見て分かるように、多くの区間は真の母平均 μ の近くで重なっています。

信頼区間の構成について議論する前に、よくある誤解を取り除いておく必要があります。

よくある誤解

観測された $\bar{x}$ の周りに 95% 信頼区間を構成したとする。このとき、「95%」は、その区間が μ を含む確率が 95% である、という意味ではありません。代わりに、それは方法全体の長期的な成功率を指します。

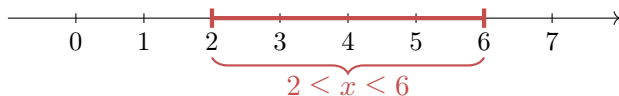
「この区間が μ を含む確率は 95% である」のように言うと、 μ が何らかの形でランダムに値を変えているように聞こえます。これは正しくありません。 μ は未知ですが、その値は完全に固定されています。選ばれた $\bar{x}$ の周りの任意の区間について、平均 μ はその区間の中にあるか、ないかのどちらかです。代わりに、分布の中でランダムに動き回っているのは $\bar{x}$ の値です。このことは $\bar{X}$ という記法で表されています。信頼区間は $\bar{x}$ の値とともに動き、信頼水準はこの動きに関するものです。

3 信頼区間の構成

3.1 区間についての確認

一般に、連結な区間はその端点によって完全に記述されます。区間がどこで始まり、どこで

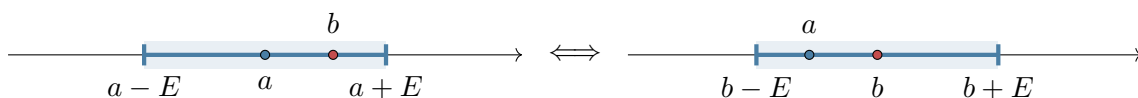
終わるかを書けば、その間のすべてが区間そのものだと理解されます。例えば、区間 $[2, 6]$ は、 $2 \leq x \leq 6$ を満たすすべての数 x を意味します。



今、 a と b が実数直線上の二つの点であるとしします。「 b が a から距離 E 以内にある」と言うことは、「 a が b から距離 E 以内にある」と言うことと同じです。記号で書けば、

$$a - E \leq b \leq a + E \quad \Longleftrightarrow \quad b - E \leq a \leq b + E.$$

幾何学的には、 b が a を中心とする長さ $2E$ の区間に入っているなら、 a も b を中心とする長さ $2E$ の区間に入っていることが分かります。



信頼区間を構成するとき、私たちの目標は、それぞれの区間の二つの端点を指定することです。観測された $\bar{x}$ の両側にある二つの端点 $\bar{x}_1$ と $\bar{x}_2$ を探します。ただし、 $\bar{x}$ はちょうど中心に来るようにします。すると、記号では次のようになります。

$$\mu \text{ が } \bar{x} \text{ を中心とする区間にある} \quad \Longleftrightarrow \quad \bar{x} - E \leq \mu \leq \bar{x} + E.$$

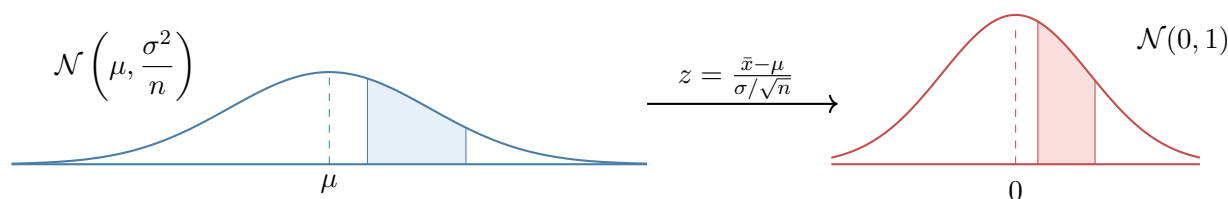
この観察が、信頼区間の背後にある重要な考え方です。すなわち、 $\bar{X}$ が μ に近ければ、 μ もまた $\bar{X}$ に近いのです。

3.2 標本分布に対する z スコア

中心極限定理によれば、十分大きな n に対して、 $\bar{X}$ の標本分布は $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ で近似できることが分かっています。これは正規分布なので、 z スコアを作ることによって標準正規分布 $\mathcal{N}(0, 1)$ と関係づけることができます。 z スコアの公式は、ある値と中心との距離を測り、それを分布の標準偏差で割ります。標本分布の場合、中心は μ であり、分布の広がり（標準偏差）は標準誤差 $\frac{\sigma}{\sqrt{n}}$ によって測られます。したがって、使うべき z スコアの公式は

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}.$$

です。これは $\bar{X}$ の標本分布を標準正規分布へ変換し、 z 表を使って標本分布内の面積を計算できるようにします。

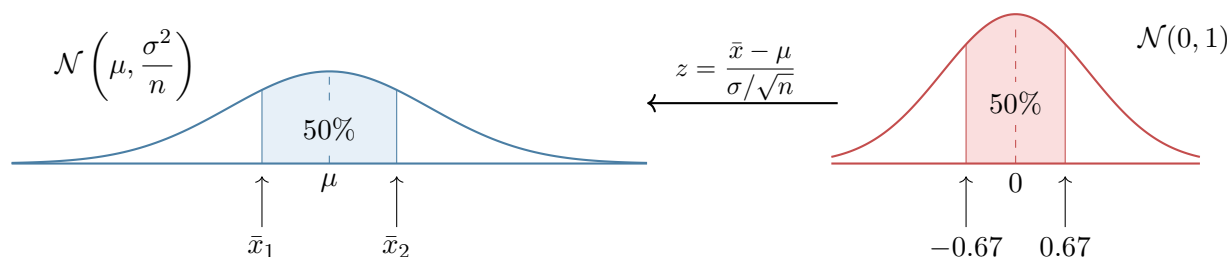


3.3 50% 信頼区間を求める

$\bar{X}$ が μ に近い確率を 50% にしたいとします。まず標準正規分布において、0 を中心とする面積 50% の領域を探すことで、 $\bar{X}$ の値の周りの信頼区間を記述できます。そこから、 z スコアの公式を使って逆向きに計算し、同じ面積に対応する境界を、今度は標本分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ の中で求めます。 z 表によると、両側の尾の面積が等しく、

$$P(z_1 < Z < z_2) = 0.5$$

を満たす z_1 と z_2 の値は、ちょうど $z_1 = -0.67$ と $z_2 = 0.67$ です。この二つの値を使って、 z スコアの公式を逆にたどり、標本分布における対応する値を求めます。



これらの値は、 μ の周りの面積の 50% を囲みます。つまり、 $\bar{X}$ は 50% の確率でこの領域に入ります。

$\bar{x}_1$ と $\bar{x}_2$ の値は、 z に ± 0.67 を代入し、代数的に解くことで求められます。 $z = 0.67$ を使うと、

$$0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu + \frac{0.67\sigma}{\sqrt{n}}.$$

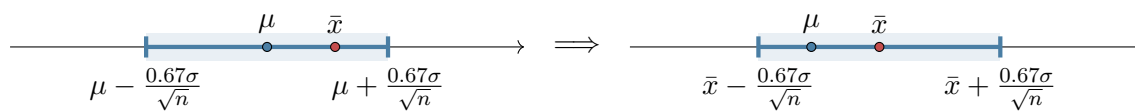
同様に、 $z = -0.67$ を使うと、

$$-0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu - \frac{0.67\sigma}{\sqrt{n}}.$$

これら二つの端点を使って、 $\bar{X}$ が 50% の確率で中に入るような、 μ の周りの区間を記述できます。記号で書くと、構成により、この区間は次を満たします。

$$P\left(\mu - \frac{0.67\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{0.67\sigma}{\sqrt{n}}\right) = 0.5.$$

ここで、3.1節の観察を使います。 $\bar{X}$ が 50% の割合で μ の近くにあるなら、 μ もまた確率 0.5 で $\bar{X}$ の近くにある、ということです。



したがって、区間の長さを同じに保ったまま、その中心を $\bar{X}$ に移すと、構成により、この区間は 50% の割合で μ を含みます。言い換えると、 $\bar{X}$ の周りに 50% 信頼区間を作ったことになります。これは次のように書きます。

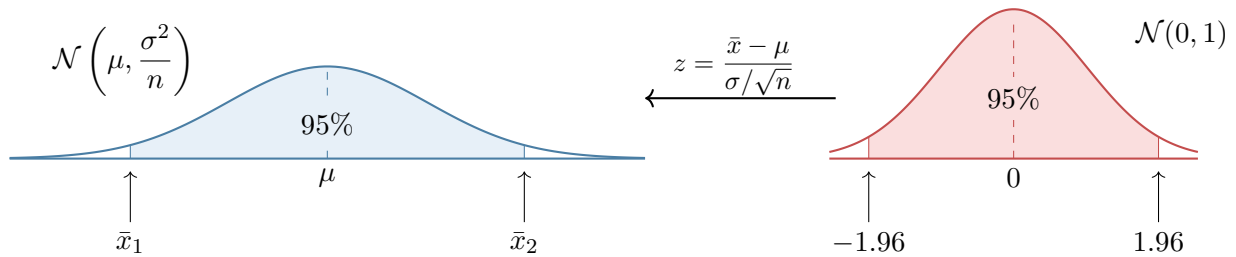
$$\left[\bar{X} - \frac{0.67\sigma}{\sqrt{n}}, \bar{X} + \frac{0.67\sigma}{\sqrt{n}} \right].$$

特定の標本平均 $\bar{x}$ を観測した後では、その周りの信頼区間を次のように書きます。

$$\left[\bar{x} - \frac{0.67\sigma}{\sqrt{n}}, \bar{x} + \frac{0.67\sigma}{\sqrt{n}} \right].$$

3.4 95% 信頼区間を求める

今度は 95% 信頼区間を構成したいとします。同じ論理が成り立ちますが、今回は $\mathcal{N}(0, 1)$ において、0 を中心とする面積が 95% になるような z の値を探すだけです。再び z 表を使うと、それは ± 1.96 であることが分かります。そこで、 z スコアの公式を使って、標本分布における対応する端点の値を求めます。



今度は ± 0.67 の代わりに $z = \pm 1.96$ を使うと、次が得られます。

$$1.96 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \implies \bar{x} = \mu + \frac{1.96\sigma}{\sqrt{n}}$$

$$-1.96 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \implies \bar{x} = \mu - \frac{1.96\sigma}{\sqrt{n}}$$

先ほどと同じ論理により、 $\bar{x}_1$ と $\bar{x}_2$ の間の領域は、95% の割合で $\bar{X}$ を含みます。

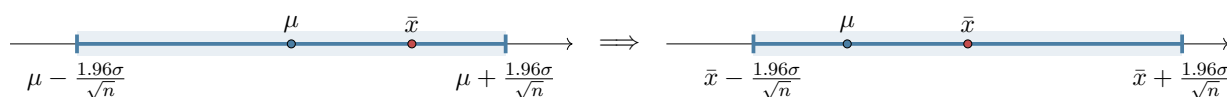
$$P\left(\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

したがって、同じ区間を取り、その中心を代わりに $\bar{X}$ に置けば、この新しい区間もまた 95% の割合で μ を含むことになります。

$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

今回は、幾何学的には $\bar{x}$ の周りにより広い区間を作ったことになります。¹

¹この図は縮尺通りではありません。実際には、本当の区間は前節で描いた区間のおよそ 3 倍の大きさになるはずです。



記号で書くと、 $\bar{X}$ の周りの 95% 信頼区間は次のようになります。

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}} \right],$$

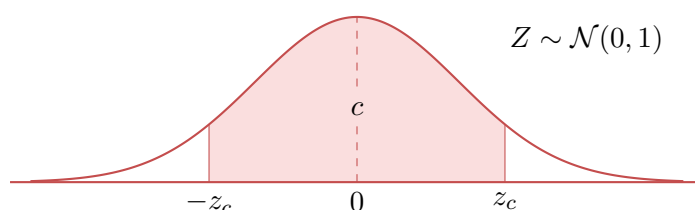
あるいは、 $\bar{x}$ の値がすでに観測されているなら、次のように書けます。

$$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right].$$

4 信頼区間の一般公式

4.1 臨界値

今、信頼水準 c の信頼区間を構成したいとします。前節と同じ考え方に従って、まず信頼水準 c に対応する標準正規分布の面積を探索する必要があります。



一般に、 c は異なる値を取ることができます。したがって、面積をとれくらい大きく、または小さくしたいかによって、 $\pm z_c$ の正確な値は変わります。この数 z_c には特別な名前があります。それは、信頼水準 c に対する臨界値と呼ばれます。

臨界値

信頼水準 c に対して、**臨界値** z_c とは、

$$P(-z_c < Z < z_c) = c,$$

を満たす正の数です。ただし、

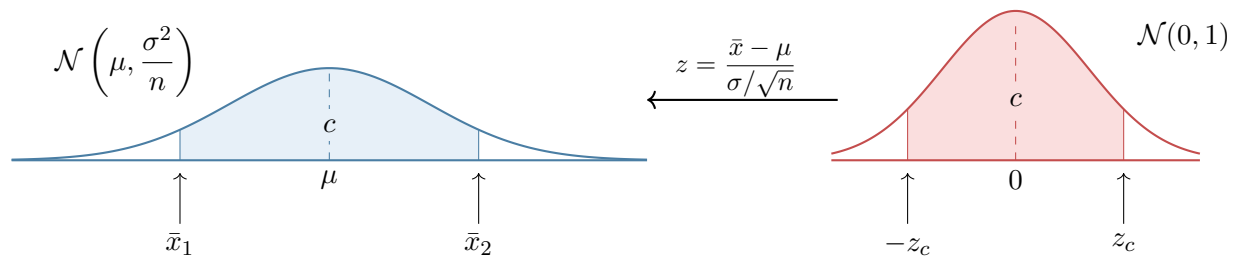
$$Z \sim \mathcal{N}(0, 1).$$

z_c の正確な値は、 z 表や電卓から求められます。ただし実際には、90% や 99% のようなよく使われる信頼水準で信頼区間を計算することが多いです。便利のため、よく使われる臨界値を下の表にまとめます。

信頼水準 c	臨界値 z_c
0.70, または 70%	1.04
0.75, または 75%	1.15
0.80, または 80%	1.28
0.85, または 85%	1.44
0.90, または 90%	1.645
0.95, または 95%	1.96
0.98, または 98%	2.33
0.99, または 99%	2.58

4.2 一般の信頼区間

今、信頼水準 c の信頼区間を構成したいとします。前節と同じ考え方に従って、まず標準正規分布の中央の面積 c を探す必要があります。 c に対応する臨界値が見つければ、 $\bar{X}$ に対応する端点を解くことができます。



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

を $\bar{x}_1$ と $\bar{x}_2$ について解くと、標本分布における端点は

$$\mu - \frac{z_c \sigma}{\sqrt{n}} \quad \text{と} \quad \mu + \frac{z_c \sigma}{\sqrt{n}}$$

になります。

したがって、構成により、 μ を中心とし、確率 c で $\bar{X}$ を含む区間が見つかりました。記号で書くと、

$$P\left(\mu - \frac{z_c \sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

「近い」という関係の対称性により、この区間を取り、ランダム変数 $\bar{X}$ の周りにランダムな区間として中心を置くことができます。すると、この区間は同じ確率で μ を含みます。

$$P\left(\bar{X} - \frac{z_c \sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

これにより、信頼区間の一般公式が得られます。

σ が既知の場合の μ の信頼区間

標本平均 $\bar{x}$ 、既知の母集団標準偏差 σ 、標本サイズ n 、信頼水準 c が与えられたとき、 μ の信頼区間は

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

4.3 誤差の限界

一般の信頼区間の公式には、

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

という量が現れます。この距離 E は、推定における許容できる誤差の範囲を表します。そのため、 E は**誤差の限界**と呼ばれます。この用語を使うと、信頼区間は単純に次のように書けます。

誤差の限界

σ が既知の場合の μ の信頼区間について、誤差の限界は

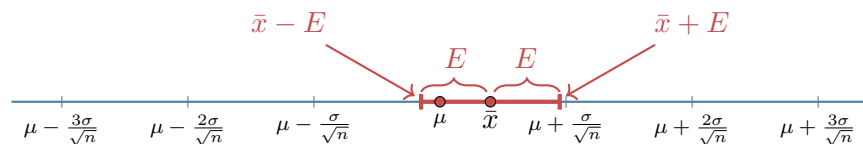
$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

信頼区間は

$$[\bar{x} - E, \bar{x} + E].$$

です。

幾何学的には、長さ E は信頼区間全体の長さの半分です。



したがって、 σ が既知の場合に μ の信頼区間を求めるには、次の手順に従います。

Step 1: 信頼水準 c を使って臨界値 z_c を求める。

Step 2: 誤差の限界を計算する。

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

Step 3: 区間を書く。

$$[\bar{x} - E, \bar{x} + E].$$