

MAT220 第11講：中心極限定理

目次

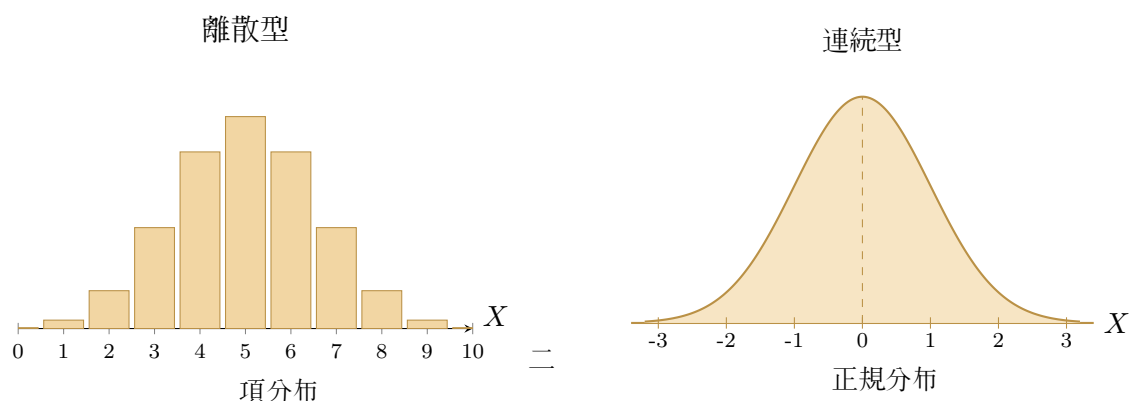
1	正規分布の簡単な復習	2
1.1	離散型確率変数と連続型確率変数	2
1.2	正規分布を記述する	2
1.3	z スコア	3
1.4	標準正規分布	4
1.5	その他の確率を計算する	4
2	標本分布	5
2.1	確率変数から標本を取る	5
2.2	標本平均の標本分布	6
3	中心極限定理	7
3.1	バージョン1：もとの母集団が正規分布の場合の標本分布	7
3.2	バージョン2：もとの母集団が必ずしも正規分布ではない場合	7
3.3	なぜ中心極限定理が重要なのか	8
3.4	例題：モルモットの体重	9
4	二項分布の正規近似	10
4.1	大きな二項実験における問題	10
4.2	中心極限定理を二項分布に適用する	12
4.3	連続性補正	14
5	例題：古着	15
5.1	方法1：正確な二項計算	15
5.2	方法2：正規近似	16

第10講では、正規分布と z スコアの使い方を学びました。この講義では、なぜ正規分布が統計学においてこれほど頻繁に現れるのかを説明します。中心となる考え方は、いわゆる中心極限定理です。これは、もとのシステムが最初から正規分布に従っていない場合であっても、多くのシステムから得られる標本平均に対して、正規分布を対応させる方法を与えるものです。実際、この結果は非常に重要なので、この講義全体をこの定理に使います。まず、定理をきちんと述べるために少し丁寧に準備します。その後、二項分布を正規曲線で近似する方法を説明します。

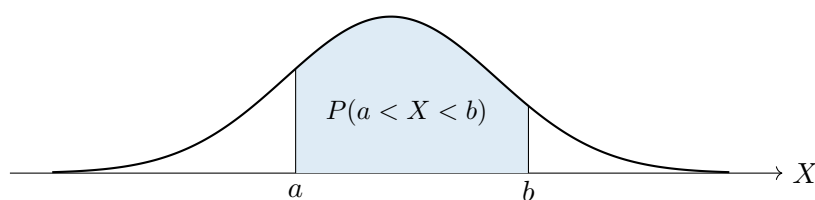
1 正規分布の簡単な復習

1.1 離散型確率変数と連続型確率変数

前回の講義では、正規分布を導入しました。これは、離散型確率変数ではなく、連続型確率変数から作られる確率分布です。これは、これまで見てきた二項分布とはかなり違って見えます。



ここでの根本的な違いは、 x 軸上で取り得る値にあります。二項分布の場合、確率変数は成功回数を数えます。したがって、確率変数が取り得る値は離散的になり、 X が取り得るそれぞれの値に確率が対応します。一方、正規分布の場合、確率変数 X は任意の実数値を取ることができます。そのため、棒の高さではなく、確率に関する問題は面積として表されます。つまり、 X がある範囲の値を取る確率は、下の図のように、曲線の下面積の大きさによって与えられます。



ここでは、確率変数 X が a と b の間の領域に入る確率は、青く塗られた領域の面積を計算することで与えられます。

1.2 正規分布を記述する

一般に、平均と標準偏差は、データ集合の特徴を記述する二つの異なる方法です。正規分布の場合、この二つの数値だけで分布全体を定義するために必要な情報がすべて揃います。つまり、正規分布は次のように数学的な記法で書くことができます。

正規分布

正規分布は、次の記法で表される連続型確率分布です。

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

ここで、 μ は平均、 σ は標準偏差、 σ^2 は分散です。

平均と標準偏差があれば、正規分布を完全に記述できます。データ集合の場合と同じように、平均 μ は分布の中心がどこにあるかを表し、標準偏差 σ は分布がどれくらい広がっているかを表します。正規分布は常に次の有用な性質を満たします。

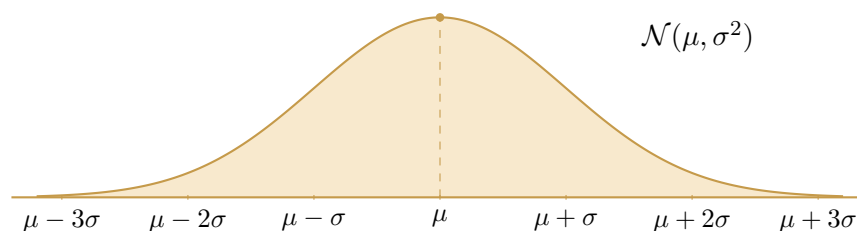
四つの重要な性質

正規分布には次の性質があります。

1. 曲線はベル型であり、最も高い点は平均 μ の上にある。
2. 曲線は μ を通る縦線に関して対称である。
3. 両端の尾は横軸に近づくが、横軸に触れたり交差したりしない。
4. 曲線全体の下側の総面積は 1 である。

1.3 z スコア

上に書いた四つの性質によって、正規分布を次のように扱うことができます。



ここでは、グラフは対称で、ちょうど μ を中心にしています。 σ は μ と同じ物理的な単位で定義されているので、グラフの中心から出発して、 σ の増加分だけ「外側へ歩く」ことができます。これにより、 X が取り得る値を、 μ からの距離として記述できます。例えば、 X が μ からかなり離れた値 x を取るなら、 $x - \mu$ は 3σ や -3σ かもしれません。また、 X が μ に比較的近い値を取るなら、それは 0.5σ だけ離れているかもしれません。

この「 μ からの距離」の測り方から σ を取り除きたいなら、単純に σ で割れば、単位を持たない純粋な数が残ります。この数を z スコアと呼びます。

$$z = \frac{x - \mu}{\sigma}.$$

要するに、値 x の z スコアは、 x が μ からどれだけ離れているかを、 σ を単位として数えるものです。 z が大きいほど、 x は μ から遠く離れています。全体の z スコアが負であれば、値 x は μ の左側にあることを意味し、 z スコアが正であれば、値 x は μ の右側にあることを意味します。

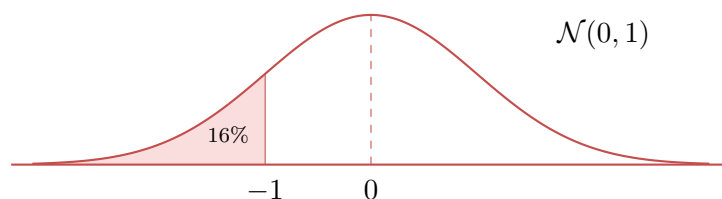
1.4 標準正規分布

第10講では、「標準」正規分布についても少し時間を使って議論しました。ここでの考え方は、これは正規分布の中で最も単純な形だということです。平均は0、標準偏差は1です。この特別な分布は非常に扱いやすいので、その確率変数を表すために特別な文字 Z を使いました。数学的な記法では、標準正規分布は $Z \sim \mathcal{N}(0, 1)$ と書けます。

便利なことに、 z 表を使うことで、標準正規分布の面積、したがって確率を近似できます。これらの表には、よく使われる Z の値の左側の面積が並んでいます。例えば、この表の小さい版は次のようになります。

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

より詳しい表は、第10講の付録にあります。視覚的には、この表の値は、ある値の左側の面積を表しています。例えば、表によると $P(Z < -1) \approx 0.16$ です。これは、 $\mathcal{N}(0, 1)$ において -1 の左側の面積がおよそ全体の 16% であることを意味します。



上のような表は、与えられた値の左側の確率しか与えません。他の確率は、正規分布の性質を使って計算できます。よく使う事実を下にまとめます。

確率に関する一般的な規則

$Z \sim \mathcal{N}(0, 1)$ に対して：

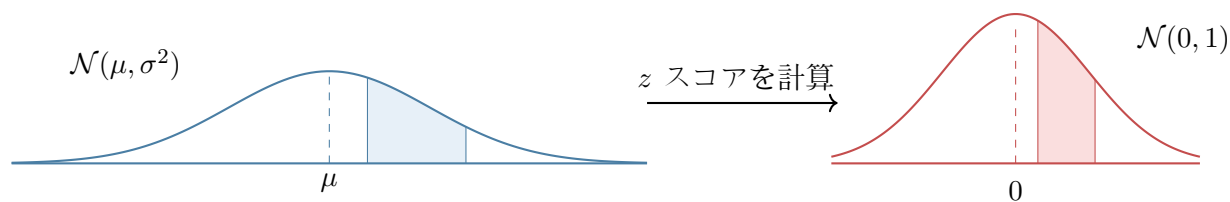
- $P(Z > z) = 1 - P(Z < z)$
- $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$
- $P(Z > z) = P(Z < -z)$
- $P(Z = z) = 0$

1.5 その他の確率を計算する

$X \sim \mathcal{N}(\mu, \sigma^2)$ が一般の正規分布であるとき、値 x は z スコアを計算することで標準正規分布へ変換できます。

$$z = \frac{x - \mu}{\sigma}.$$

この変換は面積を保存します。したがって、 X が特定の領域に入る確率は、まず標準正規分布に変換し、そこで計算することで求められます。



例：身長

この点を説明するために、日本の男性の身長が、平均 171 cm、標準偏差 5.6 cm の正規分布に従うとしましょう。例えば、身長が 180 cm より高い人に会う確率は、 z スコアを使って計算できます。ここで求めたいのは $P(X > 180)$ なので、対応する z スコアを計算します。

$$z = \frac{180 - 171}{5.6} \approx 1.61.$$

この変換は確率を保存するので、求める答えは次を計算することで得られます。

$$P(X > 180) \approx P(Z > 1.61),$$

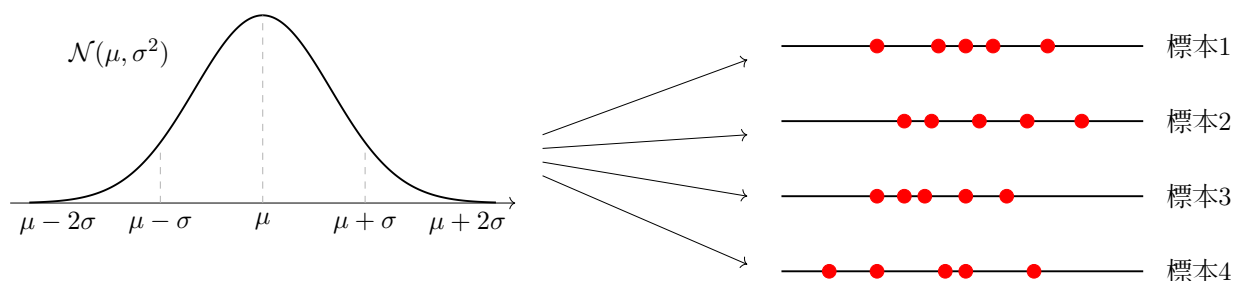
これはおよそ 5.4% です。

2 標本分布

2.1 確率変数から標本を取る

一般の正規分布 $X \sim \mathcal{N}(\mu, \sigma^2)$ を考えます。すでに述べたように、 X が任意の領域に入る確率は、その領域における曲線の下面積によって決まります。したがって、 X はより大きな面積に対応する領域に落ちる可能性が高いと考えられます。

ここで、この分布からランダムな値の標本を取るとしましょう。簡単のため、ランダムに 5 個の値を選びます。第1講の用語を使えば、これはサイズ 5 の単純ランダム標本と呼ばれます。一般に、この 5 個の値は x 軸上のどこにでも現れ得ます。したがって、サイズ 5 の単純ランダム標本を何度も繰り返し取ると、それらはかなり異なって見えるかもしれません。



それぞれの標本について、標本平均を計算できます。

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}.$$

ここで重要な観察があります。標本はランダムな値から成るので、標本平均 $\bar{x}$ の正確な値もランダムになります。したがって、標本平均そのものを一つの確率変数として解釈することができます、この確率変数を $\bar{X}$ と書きます。これから、確率変数 $\bar{X}$ の分布がどのような形になるのかを見ていきます。

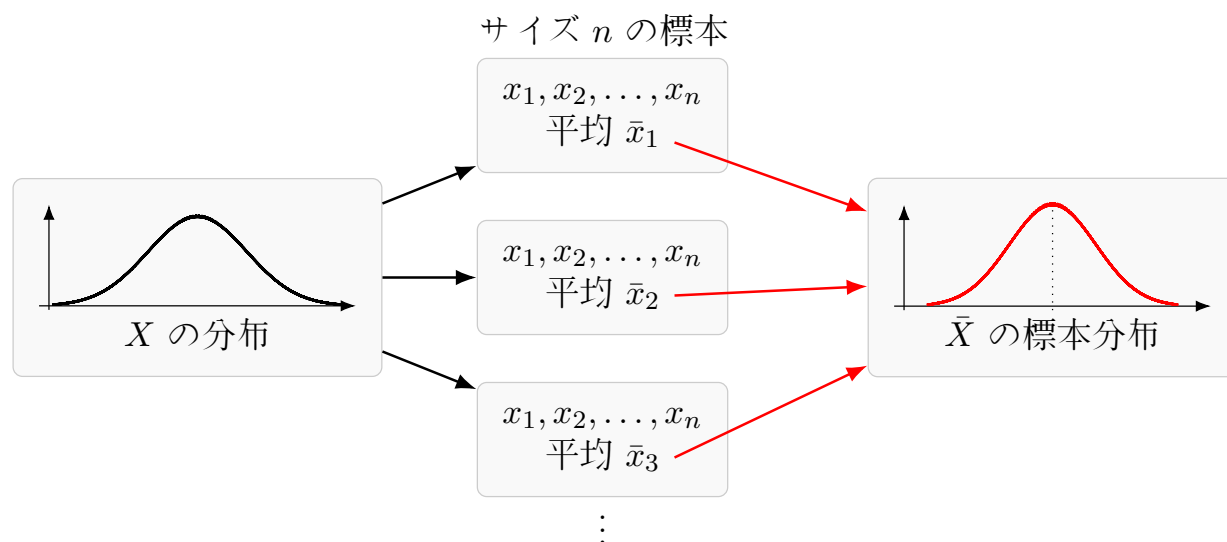
2.2 標本平均の標本分布

母集団から取り得るすべてのサイズ n の単純ランダム標本を考えます。それぞれの標本に対応する標本平均を計算し、それらをすべて集めると、新しい確率分布ができます。この新しい対象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。

標本分布の構成は、次のようなものとして想像できます。



言い換えると、まず一つの確率分布から始め、サイズ n のすべての単純ランダム標本を取り、それぞれに対応する標本平均 $\bar{x}$ をすべて計算し、それらを一つに集めます。標本はもともとランダムに抽出されたものなので、標本平均 $\bar{x}$ は確率変数 $\bar{X}$ の値として扱うことができます。したがって、取り得るすべての $\bar{x}$ の集まりには、確率分布の構造があります。

サイズ n の標本は、 X と同じ分布を持つ確率変数 $X_1, \dots, X_n$ と考えるべきであり、

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}.$$

3 中心極限定理

予想されるように、異なる確率分布は一般に異なる標本分布を持ちます。中心極限定理は、もとの分布の形に応じて、これらの標本分布がどのように見えるかを教えてくれる結果です。ここでは、二つの関連した結果を提示します。一つは正規分布に対する結果であり、もう一つはより一般の確率分布に対する結果です。

この授業では、これらの定理を証明しません。その代わり、主な焦点は、推定や仮説検定のような技法における中心極限定理の応用にあります。したがって、ここでの目的は、結果を述べて理解することです。

3.1 バージョン1：もとの母集団が正規分布の場合の標本分布

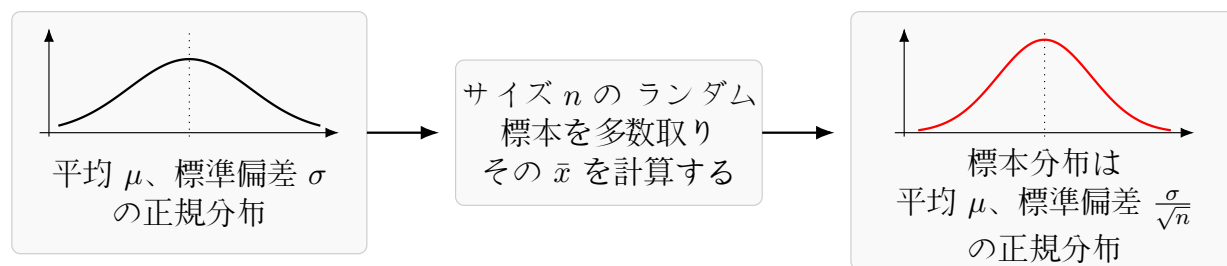
まず、もとの確率分布が正規分布であるとします。つまり、 $X \sim \mathcal{N}(\mu, \sigma^2)$ です。最初の結果は、 $\bar{X}$ の標本分布をパラメータ μ と σ に結びつけるものです。

中心極限定理：バージョン1

X を $X \sim \mathcal{N}(\mu, \sigma^2)$ を満たす確率変数とする。このとき：

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

この結果は、平均 μ 、分散 σ^2 の正規分布から始めると、すべての標本平均の集まりは、同じ中心 μ を持つが、分散が $\frac{\sigma^2}{n}$ になった別の正規分布として並ぶ、ということを述べています。



大まかに言えば、 $\mathcal{N}(\mu, \sigma^2)$ から値の標本を取ると、標本平均 $\bar{x}$ は個々の値に含まれるランダムなばらつきを平均化し、より中心に近い値を返します。したがって、平均的には、これらの標本平均は分布の真の中心に近くなると期待されます。この考え方は、標本分布の分散 $\frac{\sigma^2}{n}$ がもとの分布の分散より小さくなる理由を部分的に説明しています。公式から分かるように、 n の値が大きくなるほど標本分布はより密集します。これは、サイズの大きい標本は、平均を歪める可能性のある個々の外れ値の影響を受けにくいからです。

3.2 バージョン2：もとの母集団が必ずしも正規分布ではない場合

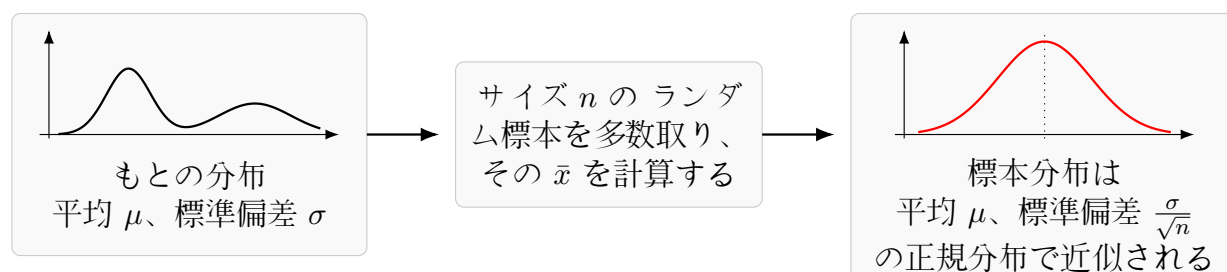
最初の結果は、正規分布から $\bar{X}$ の標本分布を作ると、その結果も正規分布になることを述べています。驚くべきことに、中心極限定理にはさらに強い形があります。

中心極限定理

X が有限の平均 μ と有限の分散 σ^2 を持つ確率分布に従うとする。十分大きな標本サイズ n に対して、しばしば $n \geq 30$ とするが、標本平均 $\bar{X}$ は、次のパラメータを持つ正規分布で近似される。

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

言い換えると、中心極限定理は、 $\bar{X}$ の標本分布が、平均 μ 、分散 $\frac{\sigma^2}{n}$ の正規分布に近い形になることを述べています。



どちらの結果でも、標本分布の標準偏差は同じ一般的な形をしていました。実は、この形には特別な名前があります。それを標準誤差と呼びます。

標準誤差

$\bar{X}$ の標本分布の標準偏差は

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

これを標準誤差と呼びます。

標準誤差は、標本平均が標本ごとにどれくらい変動するかを測るものです。標本サイズ n が大きくなると、標準誤差は小さくなります。つまり、得られる標本分布は中心 μ の周りにより密集します。先ほど述べたように、これは、大きい標本ほど外れ値の影響を受けにくく、より安定した平均を与えるという直感と一致します。

3.3 なぜ中心極限定理が重要なのか

現実の世界では、通常、母集団パラメータに直接アクセスすることはできません。例えば、東京に住む人々の平均身長を調査する場合、市内の何百万人もの人の身長をすべて測ることはできません。実際には、私たちは母集団から標本を取り、そこから標本統計量を計算することしかできません。

中心極限定理は、標本統計量と真の母平均を結びつける数学的な方法を与えます。母集団の分布の真の構造を完全には理解できないとしても、少なくとも標本を十分大きく取れば、通常は $n \geq 30$ と考えれば、その標本平均全体の分布はほぼ正規分布になると期待できます。特に、標本分布の中心は依然として真の母平均です。したがって、中心極限定理は、私たちに新しい表現力を与えます。標本平均 $\bar{x}$ と母平均 μ の間の理論的な関係を語れるようになるのです。標本統計量から母集団パラメータを推論するというこの考え方は、推測統計学の中心にあります。

次の数回の講義では、この考え方を詳しく調べます。具体的には、いくつかの標本データに基づいて μ を実際に推定する方法を見ていきます。

3.4 例題：モルモットの体重

ある大きなモルモットの母集団について、平均が 950g、標準偏差が 150g であるとします。この母集団は正規分布ではないと仮定します。ここで、30 匹のモルモットからランダム標本を取り、標本平均 $\bar{x}$ を計算するとします。次の問いに答えます。

30 匹のモルモットの平均体重が 900g から 1000g の間になる確率はいくらか。

$n = 30$ なので、中心極限定理を使って $\bar{X}$ の標本分布を近似できます。

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(950, \frac{150^2}{30}\right).$$

同じことを標準誤差で書くと、

$$\sigma_{\bar{X}} = \frac{150}{\sqrt{30}} \approx 27.39.$$

したがって、次のようにも書けます。

$$\bar{X} \approx \mathcal{N}(950, 27.39^2).$$

上の問いに答えるためには、確率 $P(900 \leq \bar{X} \leq 1000)$ を求めたいことになります。そのためには、まず両端の値を z スコアに変換する必要があります。

$$z_1 = \frac{900 - 950}{150/\sqrt{30}} \approx -1.83,$$

そして

$$z_2 = \frac{1000 - 950}{150/\sqrt{30}} \approx 1.83.$$

したがって、

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83).$$

標準正規表を使うと、

$$P(Z < 1.83) = 0.96638 \quad \text{かつ} \quad P(Z < -1.83) = 0.03362,$$

なので、次のように結論できます。

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83) = 0.96638 - 0.03362 = 0.93276.$$

したがって、30 匹のモルモットのランダム標本を取ったとき、その平均体重が 900g から 1000g の間になる確率は、およそ 93% です。

3.4.1 なぜこれが有用なのか

もしこの確率を直接計算しなければならないとしたら、非常に面倒な計算になってしまいます。まず、1000 匹のモルモットから 30 匹の標本を取るすべての場合を列挙しようとする、可能な標本の数は

$$\binom{1000}{30} \approx 2.43 \times 10^{57}.$$

古典的確率の黄金公式を使って正確な確率を直接計算するには、次のことをしなければなりません。

1. サイズ 30 のすべての標本を列挙する。
2. すべての標本について標本平均を計算する。
3. その標本平均のうち 900 から 1000 の間にあるものを数える。
4. そして、可能な標本の総数である $C_{1000,30}$ で割る。

もちろん、これは中心極限定理を使って解を近似するよりもはるかに難しい方法です。

4 二項分布の正規近似

4.1 大きな二項実験における問題

第8講を思い出すと、二項実験とは次の性質を持つ確率的なシステムでした。

1. 同じ試行が何度も繰り返される。
2. 各試行には常に二つの結果しかない。それらを成功と失敗と呼んだ。
3. 各試行の確率は同じである。
4. 試行は独立である。

二項分布は主に二つの数によって指定されました。 n は試行回数を表し、 p は成功確率を表します。二項実験における中心的な問題は、 n 回の試行で、特定の成功回数、または成功回数の範囲が起る正確な確率を求めることです。これを数学的にきちんと書くために、 n 回の試行における成功回数を数える離散型確率変数 X を定義しました。そして、 n 回の試行で r 回成功する確率は次の公式で与えられることを見ました。

$$P(X = r) = C_{n,r} p^r q^{n-r},$$

ここで $q = 1 - p$ です。

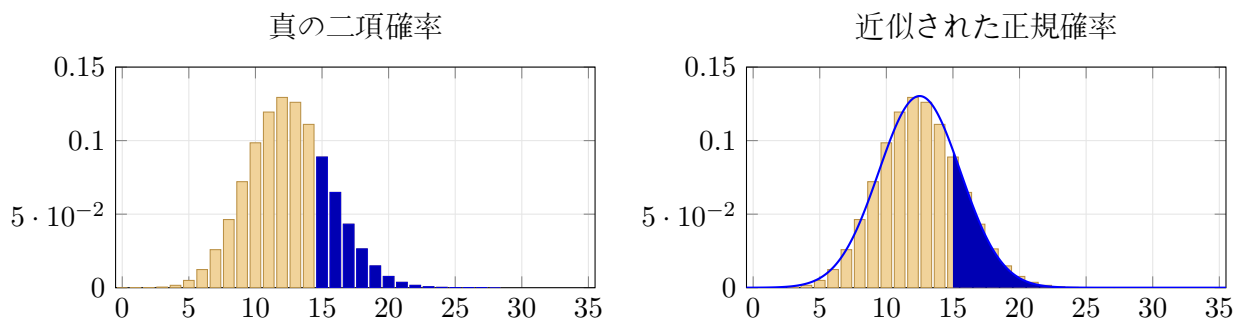
ここで、かなり大きな二項実験、例えば $n = 50$ 、 $p = 0.25$ の場合を考えましょう。もし $P(X \geq 15)$ のような複雑な確率を計算するよう求められたら、多くの個別の確率を計算して足し合わせる必要があります。

$$\begin{aligned} P(X \geq 15) &= P(X = 15) + P(X = 16) + \cdots + P(X = 49) + P(X = 50) \\ &= C_{50,15} p^{15} q^{35} + C_{50,16} p^{16} q^{34} + \cdots + C_{50,49} p^{49} q^1 + C_{50,50} p^{50} q^0, \end{aligned}$$

または、余事象を使うなら、

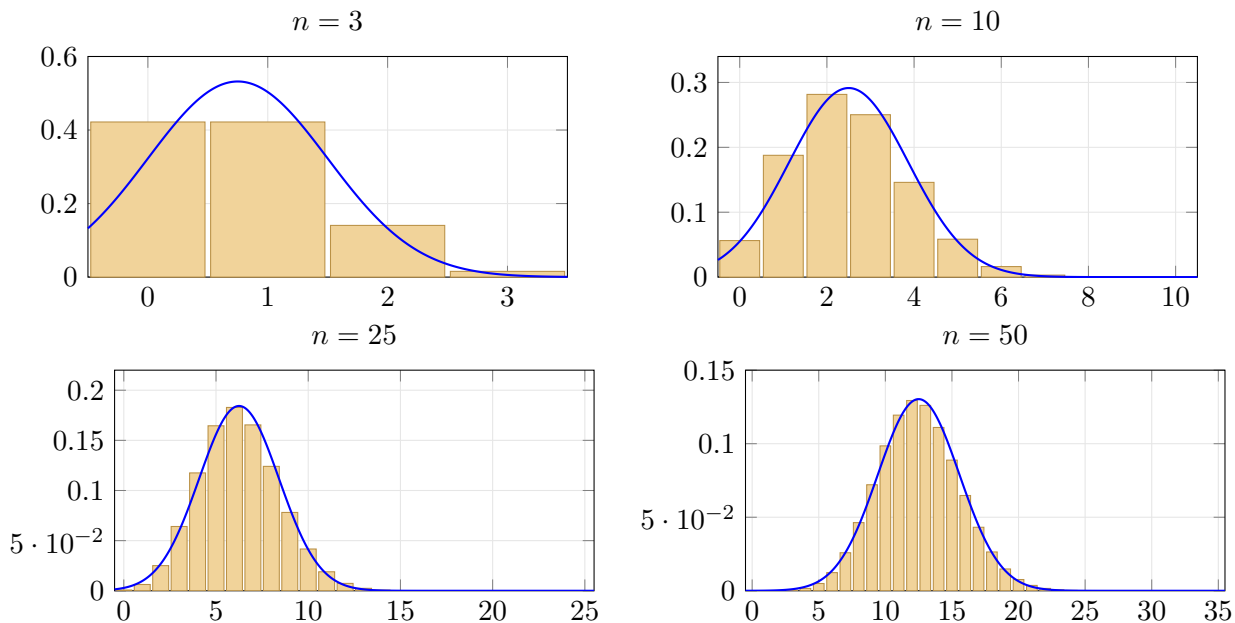
$$\begin{aligned}
 P(X \geq 15) &= 1 - P(X < 15) \\
 &= 1 - (P(X = 0) + P(X = 1) + \cdots + P(X = 13) + P(X = 14)) \\
 &= 1 - P(X = 0) - P(X = 1) - \cdots - P(X = 13) - P(X = 14) \\
 &= 1 - (C_{50,0}p^0q^{50}) - (C_{50,1}p^1q^{49}) - \cdots - (C_{50,13}p^{13}q^{37}) - (C_{50,14}p^{14}q^{36})
 \end{aligned}$$

どちらの方法で進めても、大きくて難しい計算に出会います。その代わり、このような状況では、分布を正規曲線で近似し、 z スコアを使って大まかな答えを出す方が簡単な場合があります。図で見ると、真の確率は 15 から 50 までのすべての棒の高さの合計に対応します。一方、正規近似では、分布にだいたい合う正規曲線の下で、15 の右側の面積を計算します。



もちろん、この点は、もっと大きな二項実験、例えば $n = 1000$ を考えるとより明確になります。突然、 $P(X \geq 400)$ のような確率を計算するには、公式 $C_{n,r}p^r q^{n-r}$ を 601 回計算し、その後にかなり大きな和を取る必要があります。

第10講では、二項分布が適切なスケーリングの後にどのように正規曲線に似てくるかを見ることで、正規分布を動機づけました。視覚的には、 n をとんとん大きくすると、棒の数が増え、最終的に図は連続的な曲線に似てきます。これは、 n が増えるにつれて、二項分布がますます正規曲線のように見えてくることを意味します。したがって、 z スコアを使った確率の近似は、ますます適切になります。下の図では、より小さい二項分布に正規曲線を当てはめています。 n が大きくなるにつれて、正規曲線が二項分布によりよく合っていくことに注目してください。



これから、これらの正規曲線が中心極限定理の応用によってどのように作られるかを見ます。

4.2 中心極限定理を二項分布に適用する

固定された n, p, q を持つ二項実験を考えます。 X を成功回数を数える確率変数とします。ある意味で、各試行は、それ自身の確率変数を持つ別々の確率的システムとして理解できます。次のように定義します。

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

すると、 X は次の和として書き直せます。

$$X = Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n.$$

この和の各項は、成功なら 1、失敗なら 0 を寄与します。したがって、和 X は n 回の試行後の成功総数を数えます。ここで次のことに気づきます。 X を n で割ると、都合よく標本平均のような形になります。

$$\frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n}{n}.$$

中心極限定理によると、確率変数 $Y_1, \dots, Y_n$ の標本平均は、近似的に正規分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ の形に並びます。ここで、 μ と σ^2 は、単一の試行変数 Y_i の平均と分散です。この場合、もとの変数は Y_i の一つ、つまり実験の一回の試行に対応する変数です。二項実験を扱うとき、平均に対応するものは期待値であることを知っています。単一の変数 Y_i に対して、

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

期待値¹は、

$$E(Y_i) = P(Y_i = 0) \cdot 0 + P(Y_i = 1) \cdot 1 = q \cdot 0 + p \cdot 1 = p.$$

さらに、分散は、

$$\begin{aligned} \text{Var}(Y_i) &= E((Y_i - E(Y_i))^2) \\ &= E((Y_i - p)^2) \\ &= P(Y_i = 0)(0 - p)^2 + P(Y_i = 1)(1 - p)^2 \\ &= q(-p)^2 + p(q)^2 \\ &= qp^2 + pq^2 \\ &= pq(p + q) \\ &= pq. \end{aligned}$$

まとめると、変数 $\frac{X}{n}$ は、二項実験における各試行の結果の一種の標本平均として働きます。これらの各試行は、同じ期待値 p と同じ分散 pq を持ちます。中心極限定理を適用すると、 $\frac{X}{n}$ は近似的に正規分布に従い、具体的には次のようになります。

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

この情報を使って、もとの変数 X に対する新しい近似を作ることができます。正規分布に関する少し複雑な代数²によって、次の事実が得られます。

$$\frac{X}{n} \approx \mathcal{N}\left(p, \frac{pq}{n}\right) \quad \Rightarrow \quad X \approx \mathcal{N}(np, npq).$$

したがって、 X の正規近似に到達しました。

二項分布の正規近似

X が、試行回数 n 、成功確率 p の二項分布に従い、粗い基準として $np > 5$ かつ $nq > 5$ が成り立つとき、 X は次の正規分布で近似できます。

$$X \approx \mathcal{N}(np, npq).$$

別の書き方をすれば、 $Y \sim \mathcal{N}(np, npq)$ として、 $X \approx Y$ と言います。

¹離散型確率変数について、確率変数の期待値は大まかに言えば

$$E(X) = \sum \text{確率} \times \text{結果}$$

であることを思い出しましょう。

²どうしても書くなら、正規分布を定義する確率密度関数に $\frac{x}{n}$ を代入すると、次のようになる。

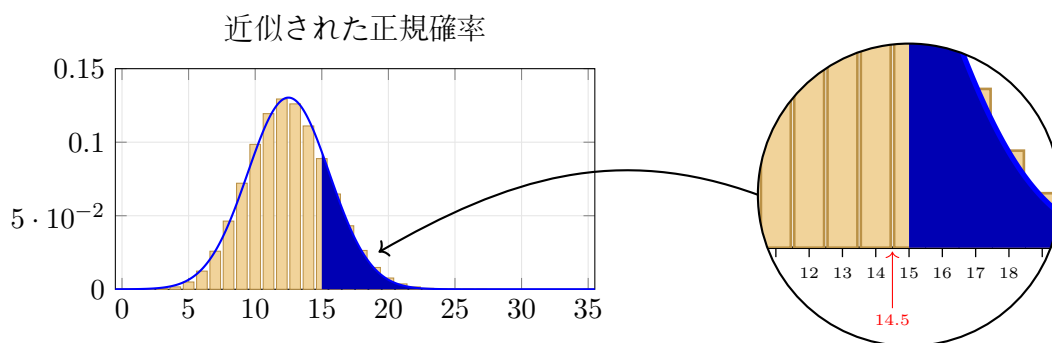
$$f\left(\frac{x}{n}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x}{n} - \mu\right)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - n\mu)^2}{2n^2\sigma^2}\right).$$

これは $\mathcal{N}(n\mu, n^2\sigma^2)$ の定義にほとんど似ていますが、最初の係数だけが異なります。式の前に余分な $\frac{1}{n}$ を付け加えると、 $\frac{1}{n}f\left(\frac{x}{n}\right)$ は $\mathcal{N}(n\mu, n^2\sigma^2)$ の定義と完全に一致します。

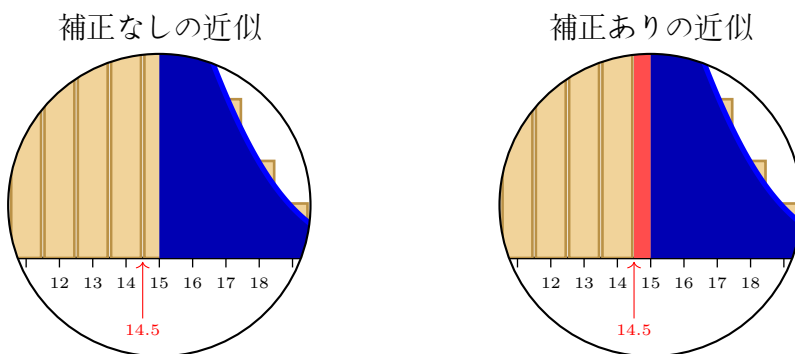
条件 $np > 5$ と $nq > 5$ は、もとの二項分布が十分大きく、かつ偏りすぎていないため、意味のある近似ができるように課されていることに注意してください。例えば、4.1節の図を見れば、 $n = 3, p = 0.25$ の場合の近似は、もとの分布にあまりよく合っていないことが分かります。

4.3 連続性補正

前節によれば、二項分布が十分大きく、かつ偏りすぎていない場合、それを正規分布 $\mathcal{N}(np, npq)$ で近似できます。すると、 z スコアを計算し、面積として確率を求めることができます。これは有用ですが、棒グラフを滑らかな面積で近似するときには、本質的な幾何学的問題があります。もう一度、4.1節の $n = 50, p = 0.25$ の例を考えましょう。図を拡大してみると、近似された連続的な面積が $X = 15$ に対応する棒を半分に切っていることが分かります。



したがって、青い領域に対応する面積は、真の確率より小さくなります。これを修正するには、面積近似を棒の中央からではなく、棒の始まりから数えるようにします。



一般に、これは補正なしの場合よりも近い近似を与えます。この調整を連続性補正と呼びます。

連続性補正の規則

二項確率変数 X を正規確率変数 Y で近似するとき、正規分布の面積が二項分布の棒を完全に覆うように、各境界値を 0.5 だけ動かします。

二項確率		正規近似
$P(X \geq r)$	は	$P(Y \geq r - 0.5)$ になる
$P(X > r)$	は	$P(Y > r + 0.5)$ になる
$P(X \leq r)$	は	$P(Y \leq r + 0.5)$ になる
$P(X < r)$	は	$P(Y < r - 0.5)$ になる
$P(a \leq X \leq b)$	は	$P(a - 0.5 \leq Y \leq b + 0.5)$ になる

例えば、 $P(15 \leq X)$ を近似したい場合は $P(14.5 \leq Y)$ を計算し、 $P(X \leq 12)$ を近似したい場合は $P(Y \leq 12.5)$ を計算します。

5 例題：古着

これから二項近似の例を一つ詳しく見ていきます。私が下北沢へ行き、いい感じの古着を買おうとしているとします。過去の経験から、ある店で買いたいと思う面白い服を見つける確率は、およそ 25% だと分かっています。店同士は独立であり、この確率は店ごとに同じままだと仮定します。私は 25 軒の店に入ります。次の問いに対する近似的な答えを求めます。

25 軒の店のうち、8 軒以上で良い服が見つかる確率はいくらか。

これを二項実験としてモデル化できると仮定して、確率変数を次のように定義します。

X = 良い服が見つかる店の数.

仮定により、変数 X は $n = 25$ 、 $p = 0.25$ 、 $q = 1 - 0.25 = 0.75$ の二項分布に従います。

5.1 方法1：正確な二項計算

今後の参考のため、まずこの結果を正確に計算します。求められているのは $P(X \geq 8)$ です。これは直接次のように計算できます。

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + \cdots + P(X = 25) \\ &= C_{25,8}(0.25)^8(0.75)^{17} + C_{25,9}(0.25)^9(0.75)^{16} + \cdots + C_{25,25}(0.25)^{25}(0.75)^0. \end{aligned}$$

電卓を使うと、

$$P(X \geq 8) \approx 0.27349.$$

5.2 方法2：正規近似

正規近似を使うためには、まず近似条件を確認する必要があります。

$$np = 25(0.25) = 6.25 \quad \text{かつ} \quad nq = 25(0.75) = 18.75.$$

これらはどちらも 5 より大きいので、正規近似は妥当です。 $n = 25$ 、 $p = 0.25$ 、 $q = 0.75$ なので、近似に使う正規分布は

$$X \approx \mathcal{N}(np, npq) = \mathcal{N}(6.25, 4.6875).$$

連続性補正を適用しない場合、 z スコアを計算することで近似確率を求めます。

$$z = \frac{8 - 6.25}{\sqrt{4.6875}} \approx 0.81.$$

したがって、

$$P(X \geq 8) \approx P(Z > 0.81) = 1 - P(Z < 0.81) \approx 1 - 0.79103 = 0.20897.$$

見て分かるように、これは正確な答え 0.27349 より明らかに小さいです。しかし、確率を計算する前に連続性補正を行うと、 $P(X \geq 8)$ を $P(Y \geq 7.5)$ に変えます。この新しい値を z スコアの公式に代入すると、

$$z = \frac{7.5 - 6.25}{2.17} \approx 0.58 \quad \implies \quad P(Y \geq 7.5) \approx P(Z \geq 0.58) \approx 0.28096.$$

これは真の値 0.27349 にかなり近いことに注意してください。