

MAT220 第17回講義：独立な母集団の検定

目次

1 対応のあるデータの復習	2
2 独立な母集団	3
2.1 仮定の変更	3
2.2 標本平均の差	3
2.3 $\bar{X} - \bar{Y}$ の標本分布の構成	4
3 独立な平均に対する仮説検定	6
3.1 帰無仮説と対立仮説	6
3.2 σ が既知の場合と未知の場合の検定統計量	6
4 2つの二項確率の検定	7
4.1 1標本比率の復習	7
4.2 2つの独立な標本比率の差	8
4.3 成功確率の差に対する仮説検定	8
5 例：オレンジ色のスキットルズ	9
6 仮説検定のまとめ	11

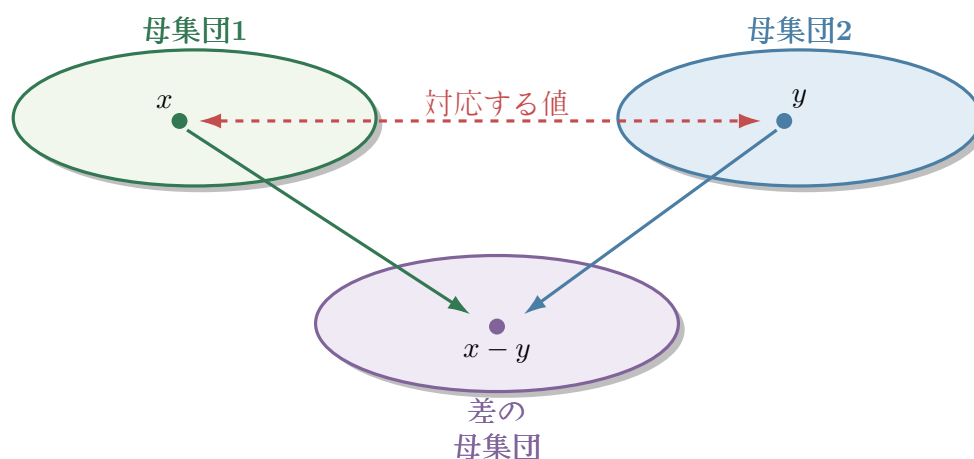
第16回講義では、観測値が自然に対応づけられる2つの母集団を比較しました。この対応づけにより、各ペアの中で差を取り、新しい差の母集団 $D = X - Y$ を作り、問題を1標本 t 検定に帰着させることができました。本講義では、この仮定を取り除き、代わりに母集団が独立であると仮定します。これから見るように、この新しい仮定の下でも、再び2つの母集団について同時に仮説検定を行うことができます。ただし今回は、差 $d = x - y$ を作ることはできません。そのため、この便利な方法を失ったことに別の方法で対処する必要があります。最終的には、新しい手法を用いて母平均の差を検定し、さらに2つの独立な二項実験の成功確率の差も検定します。

1 対応のあるデータの復習

本講義でも引き続き、2つの異なる母集団の母数を比較する仮説検定について考えます。2つの母集団が与えられたとき、中心となる問題は次のとおりです。

この2つの母集団は、平均的に同じだろうか？

原理的には、2つの母集団の真の平均を比較し、それらが等しいかどうかを検定することで、この問いに答えることができます。このような検定を行う過程では、2つの母集団から得られた標本データを組み合わせた標本分布を作る必要があります。第16回講義では、一般にこれは難しい作業であることを確認しました。そこで、母集団が自然に対応づけられていると仮定することで、問題を簡単にしました。一方の母集団の各観測値に、他方の母集団の中のただ1つの対応する観測値が存在するため、この仮定によって、対応する値の差をすべて集めた第3の母集団を作ることができました。



記号では、この「差の母集団」を文字 D で表しました。形式的には、 D は2つの母集団を定義するために用いた確率変数の差 $X - Y$ にほかなりません。

期待値と分散に関する標準的な事実を用いると、この確率変数 D の平均と分散は次のようになります。

$$\mu_D = \mu_X - \mu_Y, \quad \sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2 \text{Cov}(X, Y).$$

平均は簡単な形になりますが、分散には共分散と呼ばれる追加の項が含まれます。大まかに言えば、この共分散の項は、対応する値がどのように一緒に変動するかを測るものです。一般に σ_D は未知であるため、差の標準偏差を s_D で推定し、スチューデントの t 分布を使用する必要があります。その結果、次の検定統計量が得られます。

対応のあるデータの検定統計量

対応のあるデータでは、まず差 $d_i = x_i - y_i$ を計算します。次に、

$$t = \frac{\bar{d} - 0}{s_D / \sqrt{n}}, \quad df = n - 1$$

を使用します。

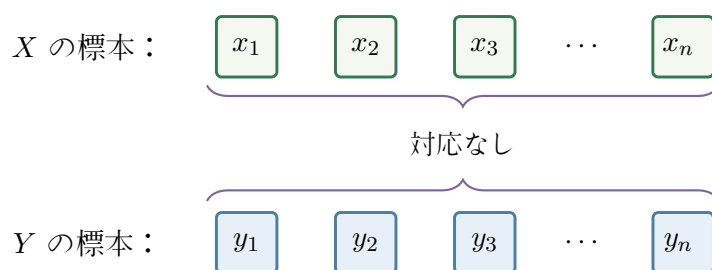
2 独立な母集団

2.1 仮定の変更

2つの母集団を含む仮説検定を行うとき、適切な標本分布を記述しようとする、大きな技術的問題が生じます。前回の講義では、問題を簡単にするため、2つの母集団が自然に対応づけられていると仮定しました。これにより、母集団 D から標本を取り、それに対応する X と Y の標本に分解することで、 $\bar{D}$ の標本分布を記述することができました。

本講義では、母集団が対応づけられているとは仮定しません。代わりに、 X と Y が独立である、すなわち両者の間に確率的な依存関係がないと仮定します。実際、この用語は、この授業の前半で確率を学んだときにも登場しました。2つの事象は、一方の発生が他方の確率を変化させないとき、「独立」と呼ばれました。同様に、一方を知っても他方の確率分布が変化しないとき、 X と Y は独立であると言います。数学的には、 X と Y が独立であるという仮定は、両者の間に相互作用がないことを意味するため、共分散 $\text{Cov}(X, Y)$ は 0 になります。

これは有望に見えますが、対応のあるデータという仮定を緩めたことで、失ったものもあることを認めなければなりません。 X と Y の観測値が対応づけられていたときには、 $D = X - Y$ に従う第3の母集団を作ることができ、そのデータ点は $d = x - y$ という形をしていました。これにより、母集団 D から標本を取り、それを X と Y の個別の標本に分解することができました。 X と Y が対応づけられているという仮定を取り除くと、この方法は使用できなくなります。言い換えると、 X と Y から同じ大きさの2つの単純無作為標本を選んだとしても、それらを自然に組み合わせて差の標本を作る方法はありません。



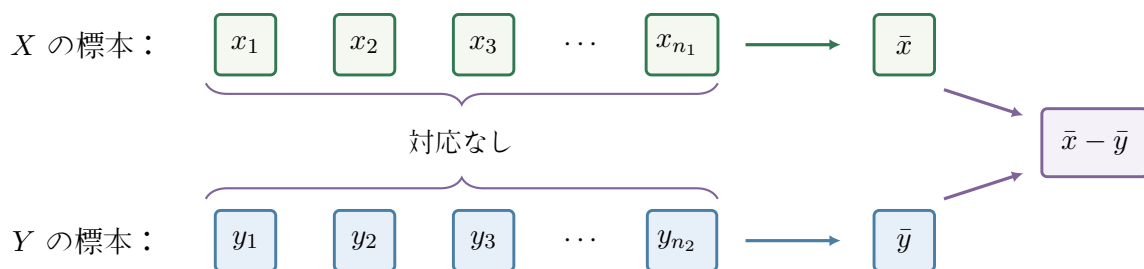
したがって、差 $\mu_X - \mu_Y$ を検定するための個々の差の標本を自然に作る方法はありません。

2.2 標本平均の差

X と Y から、それぞれ独立な無作為標本を取ります。標本は独立なので、2つの標本の大きさを同じにする必要はありません。そこで、 X から取った標本の大きさを n_1 、 Y から取った標本の大きさを n_2 とします。このとき、標本平均 $\bar{x}$ と $\bar{y}$ は個別に計算できます。

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_{n_1}}{n_1} \quad \text{および} \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_{n_2}}{n_2}.$$

最終的な目標は、差の母集団 D を直接用いることなく、 $\mu_X - \mu_Y$ を中心とする標本分布を記述することです。利用できる標本データから、 $\bar{x} - \bar{y}$ という量を考えることで、平均の差を作ることができます。



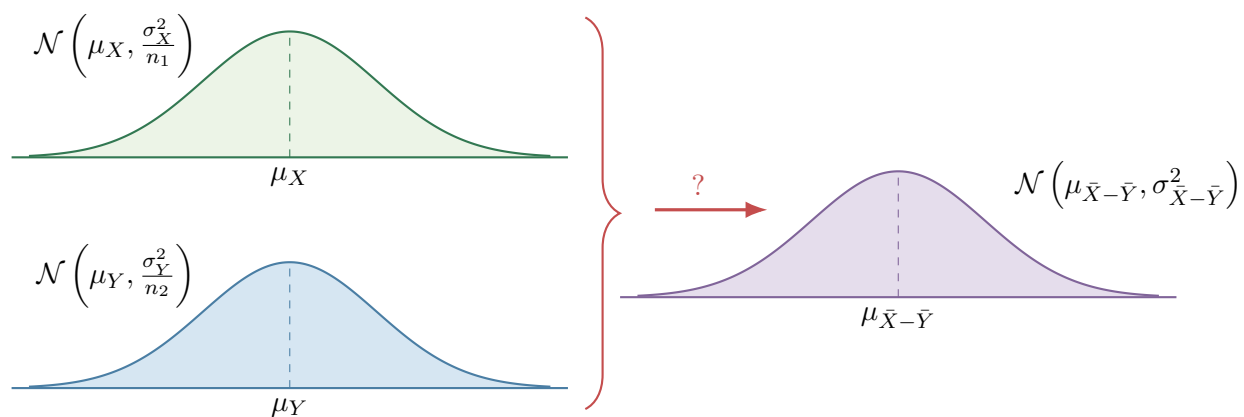
この特定の観測値 $\bar{x} - \bar{y}$ が得られる可能性を評価して仮説検定を行うためには、無作為に得られ得るすべての差 $\bar{x} - \bar{y}$ を集めた確率分布を考える必要があります。そこで、確率変数 $\bar{X} - \bar{Y}$ を調べます。

標本が十分に大きい場合、中心極限定理により、 $\bar{X}$ と $\bar{Y}$ の個別の標本分布は、それぞれ次の母数をもつ正規分布で近似されます。

$$\bar{X} \approx \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n_1}\right) \quad \text{および} \quad \bar{Y} \approx \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right).$$

対応のあるデータの場合とは異なり、2つの標本の大きさは同じでなくてもよいこと、すなわち $n_1 = n_2$ である必要はないことを強調しておきます。

ここから、この2つの正規分布を用いて、確率変数 $\bar{X} - \bar{Y}$ の全体的な分布を記述します。模式的には、次のものを求めています。



2.3 $\bar{X} - \bar{Y}$ の標本分布の構成

前回の講義では、2つの確率変数の差の期待値は、それぞれの確率変数の期待値の差に等しいことを確認しました。 $\bar{X} - \bar{Y}$ について記号で表すと、

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y})$$

となります。 $E(\bar{X})$ は単に $\bar{X}$ の平均、すなわち μ_X であり、 $\bar{Y}$ についても同様です。したがって、

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

と結論できます。都合のよいことに、これは対応のあるデータの場合に現れた中心と同じです。2つの標本平均の差の期待値は、2つの母平均の差になります。

分散についても、第16回講義と同じ事実を再び用いることができ、

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2 \text{Cov}(\bar{X}, \bar{Y})$$

となります。ここで、 $\text{Cov}(\bar{X}, \bar{Y})$ は $\bar{X}$ と $\bar{Y}$ の共分散です。ここで重要な簡単化が起こります。標本は独立であると仮定されているため、 $\bar{X}$ と $\bar{Y}$ も独立であり、この追加の共分散の項は 0 になります。

$$\text{Cov}(\bar{X}, \bar{Y}) = 0.$$

したがって、上の分散の公式は次のように簡単化できます。

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}).$$

$\bar{X}$ の分散は、その正規分布の第2母数、すなわち $\bar{X}$ の標準誤差を2乗した値 $\frac{\sigma_X^2}{n_1}$ です。同様に、 $\bar{Y}$ の分散は $\frac{\sigma_Y^2}{n_2}$ です。したがって、独立な母集団について、 $\bar{X} - \bar{Y}$ の分布の分散は次のようになります。

$$\sigma_{\bar{X} - \bar{Y}}^2 = \text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}.$$

以上をまとめると、 $\bar{X} - \bar{Y}$ を近似的に記述する正規分布を完全に表すことができます。

独立な平均の差の標本分布

独立な標本について、

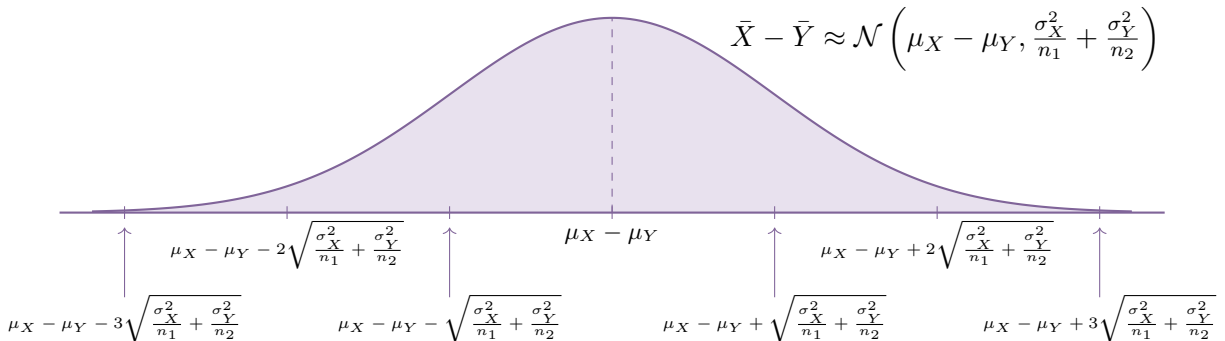
$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right)$$

となります。

この正規分布の中心は $\mu_X - \mu_Y$ であり、次の分散の平方根である標準誤差を単位として目盛りを数えます。

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}.$$

幾何学的には、次の図が得られます。



これは複雑に見えますが、重要なのは、完全に扱うことができるという点です。つまり、必要な確率を計算するために使用できる公式が存在します。

3 独立な平均に対する仮説検定

3.1 帰無仮説と対立仮説

前回の講義と同様に、2つの独立な母集団が平均的に等しいかどうかを検定するとき、帰無仮説は

$$H_0: \mu_X - \mu_Y = 0$$

となります。

次に、 X と Y から、大きさが異なる可能性のある2つの独立な標本を取ります。帰無仮説を仮定すると、標本分布は

$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(0, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right)$$

となります。 X と Y から取った2つの標本には、それぞれ標本平均 $\bar{x}$ と $\bar{y}$ があり、その差 $\bar{x} - \bar{y}$ は、上で記述した正規分布上のどこかに位置します。これまでに見てきた他の仮説検定と同様に、観測値 $\bar{x} - \bar{y}$ が中心から離れるほど、P値は小さくなります。

第16回講義の対応のある母集団の場合と同様に、 H_0 をそれに反する主張に対して検定する場合、具体的な疑いに応じて異なる対立仮説を使用できます。

$$H_1: \begin{cases} \mu_X - \mu_Y < 0, & X \text{ の母集団の平均が小さいことを表す,} \\ \mu_X - \mu_Y > 0, & X \text{ の母集団の平均が大きいことを表す,} \\ \mu_X - \mu_Y \neq 0, & \text{単に2つの平均が異なることを表す.} \end{cases}$$

向きの選択

引き算の順序は、対立仮説と一致していなければなりません。標本の差として $\bar{x} - \bar{y}$ を使用する場合、

- $\mu_X < \mu_Y$ は $H_1: \mu_X - \mu_Y < 0$ となります。
- $\mu_X > \mu_Y$ は $H_1: \mu_X - \mu_Y > 0$ となります。

3.2 σ が既知の場合と未知の場合の検定統計量

2つの母標準偏差 σ_X と σ_Y がどちらも既知であるとします。この場合、 $\bar{X} - \bar{Y}$ の分布の標準誤差は

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

です。観測された差を z スコアに変換するとき、模式的には次の商を取ります。

$$z = \frac{\text{観測値} - \text{平均}}{\text{標準誤差}}.$$

ここでは、帰無仮説 $\mu_X - \mu_Y = 0$ を仮定しています。標準誤差は上で与えられており、利用可能な情報から直接計算できます。したがって、観測された差 $\bar{x} - \bar{y}$ は、次の公式によって z ス

コアに変換されます。

$$z = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

標準偏差が既知の場合の独立な平均

σ_X と σ_Y が既知であるときに $H_0: \mu_X - \mu_Y = 0$ を検定するには、次の公式を使用します。

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

次に、 σ_X または σ_Y のいずれかが未知であるとしてします。この場合、前の公式に代入する値がありません。そこで、次善の方法として、両方の母標準偏差をそれぞれの標本推定値で置き換え、近似的なスチューデントの t 分布を用います。

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \quad \xrightarrow{\substack{\sigma_X \text{ と } \sigma_Y \text{ を} \\ s_X \text{ と } s_Y \text{ に置き換える}}} \quad t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}.$$

σ_X または σ_Y が未知 スチューデントの t 分布、
 $df = \min(n_1 - 1, n_2 - 1)$

この授業では、慣例として、利用可能な2つの自由度のうち小さい方を使用します。

標準偏差が未知の場合の独立な平均

いずれかの母標準偏差が未知であるときに $H_0: \mu_X - \mu_Y = 0$ を検定するには、次の公式を使用します。

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

ここで、 $df = \min(n_1 - 1, n_2 - 1)$ です。

その後、対応する z 表または t 表を用いて、適切なP値を計算できます。

4 2つの二項確率の検定

4.1 1標本比率の復習

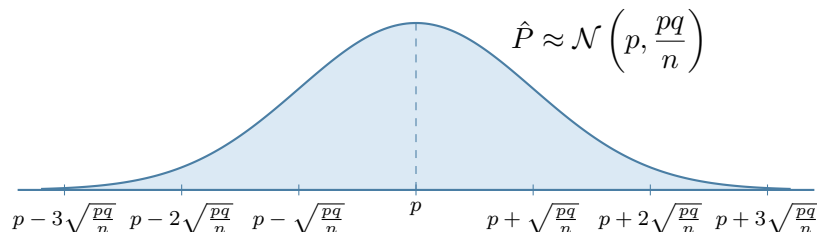
成功確率が p である1つの二項実験を考えます。 n 回の試行を行い、 r 回の成功を観測した場合、この標本データから p の推定値を計算できます。この推定値は標本比率と呼ばれ、次のように簡単に計算できます。

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

第13回講義と第15回講義で確認したように、この状況では、中心極限定理によって次の正規近

似が得られます。

$$\hat{P} \approx \mathcal{N}\left(p, \frac{pq}{n}\right), \quad q = 1 - p.$$



このような状況を扱うときには、正規近似が妥当となるほど標本が十分に大きいことを必ず確認しなければなりません。この授業では、 $np > 5$ および $nq > 5$ という基準を使用します。

4.2 2つの独立な標本比率の差

ここで、2つの二項実験があるとします。

- 実験1の真の成功確率を p_1 、標本の大きさを n_1 、標本比率を $\hat{p}_1$ とします。
- 実験2の真の成功確率を p_2 、標本の大きさを n_2 、標本比率を $\hat{p}_2$ とします。

個別には、

$$\hat{P}_1 \approx \mathcal{N}\left(p_1, \frac{p_1 q_1}{n_1}\right) \quad \text{および} \quad \hat{P}_2 \approx \mathcal{N}\left(p_2, \frac{p_2 q_2}{n_2}\right)$$

となります。ここで、 $q_1 = 1 - p_1$ および $q_2 = 1 - p_2$ です。標本比率は標本平均と同じ役割を果たすため、それらの差を取って新しい変数 $\hat{P}_1 - \hat{P}_2$ を作ることができます。この2つの実験が独立であると仮定すると、共分散の項は再び 0 になり、次の形の正規分布が得られます。

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

これを動機として、2つの実験の成功確率が等しいかどうかを検定する方法を説明します。

4.3 成功確率の差に対する仮説検定

2つの成功確率が等しいかどうかを検定するため、次の帰無仮説から始めます。

$$H_0 : p_1 - p_2 = 0.$$

この仮定の下では、上で述べた正規分布の中心は 0 になります。

$$\mathcal{N}\left(0, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

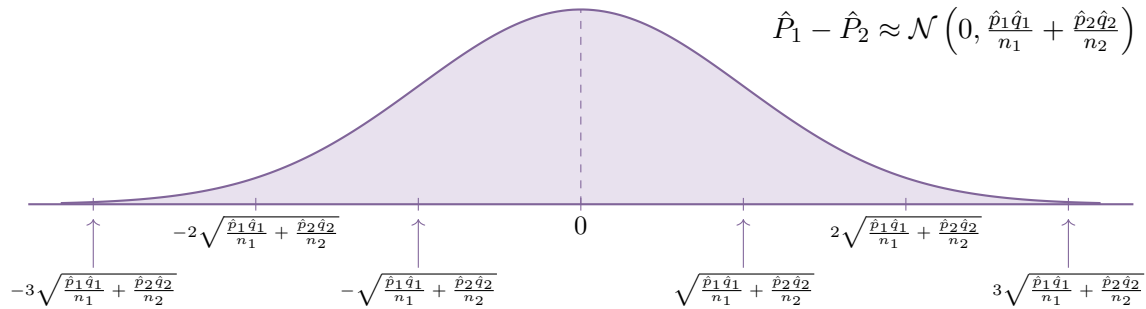
しかし、実際には p_i と q_i の値が分からないため、この分布を検定に直接使用することはできません。なぜでしょうか。帰無仮説が述べているのは差 $p_1 - p_2$ についてだけであり、それぞれの正確な値についてではないからです。したがって、 p_1 と p_2 を未知として扱い、代わりに標本比率を用いてその値を推定する必要があります。つまり、正規分布は次のようになります。

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(0, \frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right),$$

ここで、 $\hat{q}_1 = 1 - \hat{p}_1$ および $\hat{q}_2 = 1 - \hat{p}_2$ です。幾何学的には、これは再び 0 を中心とするベル型曲線ですが、今回は標準誤差

$$\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

を単位として目盛りを数えます。



対応する z スコアは、上の標準誤差を単位として中心 0 から目盛りを数えます。したがって、公式は次のようになります。

$$z = \frac{\text{観測値} - \text{平均}}{\text{標準誤差}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}.$$

これが、適切なP値を計算するために使用する公式です。

2つの独立な二項確率の検定

$H_0: p_1 - p_2 = 0$ を検定するには、

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

を使用します。P値は $Z \sim \mathcal{N}(0, 1)$ を用いて計算します。

第4.1節の正規分布の場合と同様に、まず2つの二項実験を正規分布によって意味のある形で近似できることを必ず確認しなければなりません。ここでは、近似を行う前に、次の4つの量

$$n_1 \hat{p}_1, \quad n_1 \hat{q}_1, \quad n_2 \hat{p}_2, \quad n_2 \hat{q}_2$$

がすべて 5 より大きいことを確認する必要があります。

5 例：オレンジ色のスキットルズ

ある教授が、2つのクラスにスキットルズの袋を配ったとします。教授は、クラス1におけるオレンジ色のスキットルズの比率が、クラス2における比率よりも小さいかどうかを検定したいと考えています。

記録されたデータは次のとおりです。

- ・ クラス1では 740 個のスキットルズを数え、そのうち 139 個がオレンジ色でした。
- ・ クラス2では 741 個のスキットルズを数え、そのうち 150 個がオレンジ色でした。

有意水準 $\alpha = 0.05$ を用いて検定を行います。

ステップ1：情報を整理する

重要な値は

$$n_1 = 740, \quad r_1 = 139, \quad n_2 = 741, \quad r_2 = 150, \quad \alpha = 0.05$$

です。 p_1 をクラス1におけるオレンジ色のスキットルズの真の比率、 p_2 をクラス2における真の比率とします。帰無仮説は2つの比率が等しいことを表し、疑いは第1の比率が小さいことを表します。

$$H_0 : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0.$$

これは左片側検定です。

ステップ2：P値を計算する

まず、2つの標本比率を計算します。

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{139}{740} \approx 0.188, \quad \hat{p}_2 = \frac{r_2}{n_2} = \frac{150}{741} \approx 0.202.$$

したがって、

$$\hat{q}_1 = 1 - \hat{p}_1 \approx 0.812, \quad \hat{q}_2 = 1 - \hat{p}_2 \approx 0.798.$$

次に、正規近似の条件を確認します。

$$n_1 \hat{p}_1 \approx 740(0.188) = 139 > 5,$$

$$n_1 \hat{q}_1 \approx 740(0.812) = 601 > 5,$$

$$n_2 \hat{p}_2 \approx 741(0.202) = 150 > 5,$$

$$n_2 \hat{q}_2 \approx 741(0.798) = 591 > 5.$$

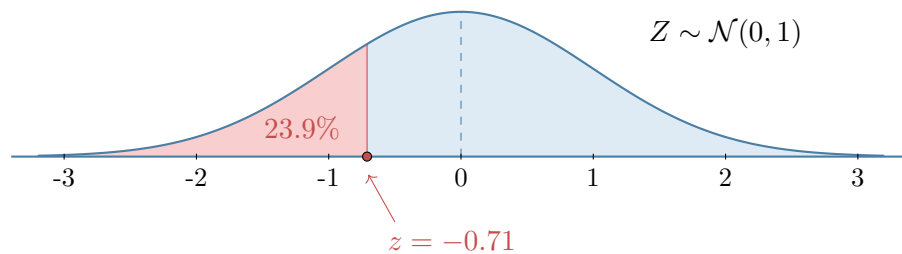
4つの値はすべて 5 より大きいため、正規近似は妥当です。ここで、得られた情報を用いて z スコアを計算できます。

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \approx \frac{0.188 - 0.202}{\sqrt{\frac{(0.188)(0.812)}{740} + \frac{(0.202)(0.798)}{741}}} \approx -0.71.$$

これは左片側検定なので、P値は z 表から直接読み取ることができます。

$$P\text{値} = P(\hat{P}_1 - \hat{P}_2 \leq -0.0146) \approx P(Z \leq -0.71) \approx 0.2389.$$

幾何学的には、左側の裾にある、グラフ全体のおよそ 23.9% に相当する面積を計算したことになります。



この面積は、帰無仮説に基づいて作られた標本分布において、標本データ $\hat{p}_1 - \hat{p}_2$ より左側にある面積と同じです。言葉で表すと、2つのクラスにおけるオレンジ色のスキットルズの比率が本当に同じであるならば、 $\hat{p}_1 < \hat{p}_2$ の方向に少なくとも今回と同じ程度まで離れた差が得られる確率は、約 23.9% です。

ステップ3：P値と α を比較する

比較すると、

$$0.2389 > 0.05$$

です。したがって、 H_0 を棄却しません。

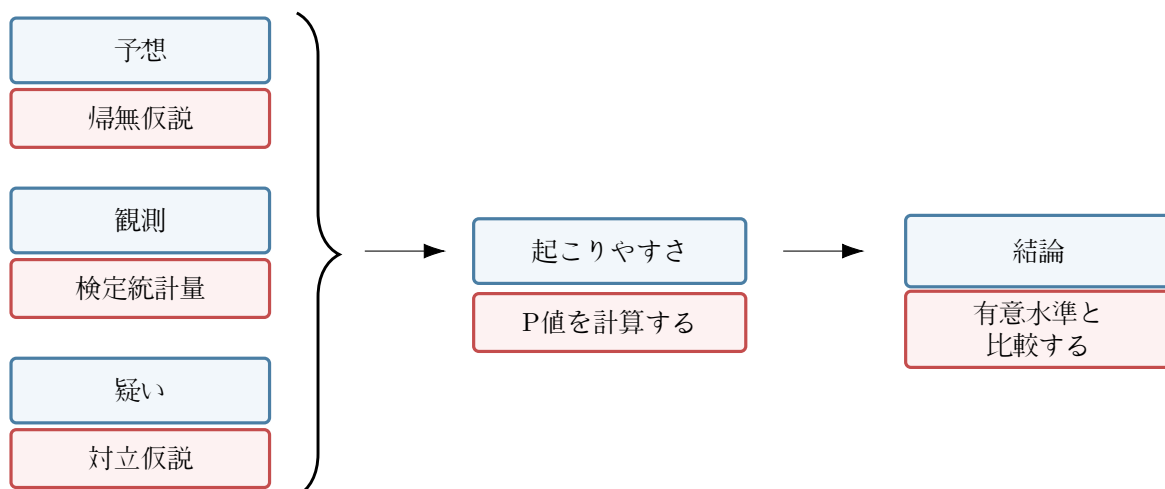
ステップ4：解釈を述べる

解釈

5% の有意水準では、クラス1におけるオレンジ色のスキットルズの比率がクラス2における比率よりも低いと結論づけるための説得力のある証拠は、標本から得られません。

6 仮説検定のまとめ

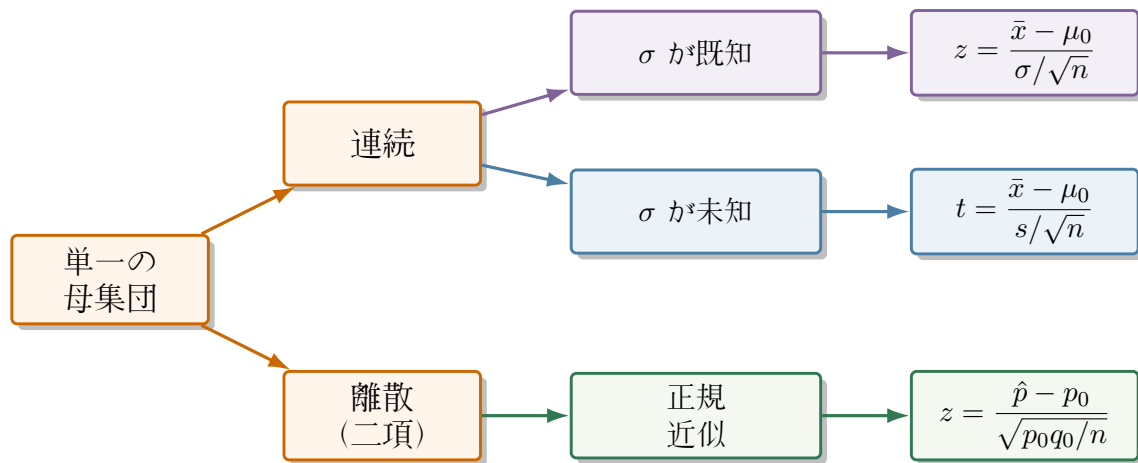
これまでに見てきた仮説検定は、すべて同じ全体的な流れに従います。



これまでに、この授業では7つの主要な仮説検定を学びました。

1. σ が既知の場合の、単一の母平均 μ の検定
2. σ が未知の場合の、単一の母平均 μ の検定
3. 単一の二項成功確率 p の検定
4. 対応のあるデータについての $\mu_X - \mu_Y$ の検定
5. 母標準偏差が既知である独立なデータについての $\mu_X - \mu_Y$ の検定
6. 母標準偏差が未知である独立なデータについての $\mu_X - \mu_Y$ の検定
7. 2つの独立な二項実験についての差 $p_1 - p_2$ の検定

これらは、単一の母集団の検定と、2つの母集団を同時に扱う検定という、2つの大きなグループに分けることができます。単一の母集団の場合、次のフローチャートによって各検定を整理できます。



2つの母集団を同時に検定する場合、次のフローチャートが得られます。

