

MAT220 第14回講義：仮説検定入門

目次

1 はじめに	2
1.1 統計的検定の非公式な構造	3
1.2 正規分布の簡単な復習	3
2 仮説検定	4
2.1 統計的検定の一般的な構成	4
2.2 P値	5
2.3 有意水準	5
2.4 さまざまな種類の検定	6
2.5 第1種の過誤と第2種の過誤	8
3 標本検定：2枚のコイン	9
3.1 怪しいコイン	9
3.2 修道女たちのコイン	10
3.3 検定の図示	11

最近の講義では、標本データを用いて母数がだいたいどこにあるのかを推測する「推定」について学んできました。ここからは、主張の検定を行うことで、推測統計の議論を続けていきます。母数を推定する代わりに、母数の真の値に関する主張を評価します。このような手続きを仮説検定といい、推測統計における強力な手法です。今後4回の講義を通して、この手法を詳しく学びます。今回は、仮説検定の全体的な構造だけを紹介します。

1 はじめに

散歩をしている途中で、2つのグループに出会ったとします。一方は危険で怪しそうな人々のグループで、もう一方は信頼できそうな修道女と子どもたちのグループです。



怪しそうなグループ



信頼できそうなグループ

まず、怪しそうなグループのところへ行きます。彼らはあなたにコインを渡し、次のゲームを提案します。

コインを何度も投げ続けてください。表が出るたびに、私たちはあなたに\$2を支払います。しかし、裏が出るたびに、あなたは私たちに\$1を支払わなければなりません。

最初は気前のよいゲームに思えたので、あなたは参加することにします。しかし、30回投げたところ、表はわずか3回しか出ず、裏が27回出ました。最終的にあなたは\$21を支払うことになる、その場を後にします。

不審に思ったあなたは、いくつか計算を試みることにします。コイン投げは二項実験なので、次の公式を使います。

$$P(X = 3) = C_{30,3}(0.5)^3(0.5)^{27} \approx 3.78 \times 10^{-6},$$

ここでは、コインが公平であり、 $p = q = 0.5$ であると仮定しています。この計算から、30回中3回しか表が出ない確率は非常に小さいことが分かります。そのため、おそらくこのコインには仕掛けがあり、 p は50%よりかなり低いと結論づけれます。

次に、信頼できそうなグループのところへ行き、話しかけます。不思議なことに、こちらのグループもまったく同じゲームを提案します。

コインを何度も投げ続けてください。表が出るたびに、私たちはあなたに\$2を支払います。しかし、裏が出るたびに、あなたは私たちに\$1を支払わなければなりません。

あまり気は進みませんが、もう一度30回コインを投げることにします。今回は、表が12回、裏が18回出たので、\$6を得ることができました。再びコインが公平であると仮定すると、

$$P(X = 12) = C_{30,12}(0.5)^{12}(0.5)^{18} = C_{30,12}(0.5)^{30} \approx 0.081.$$

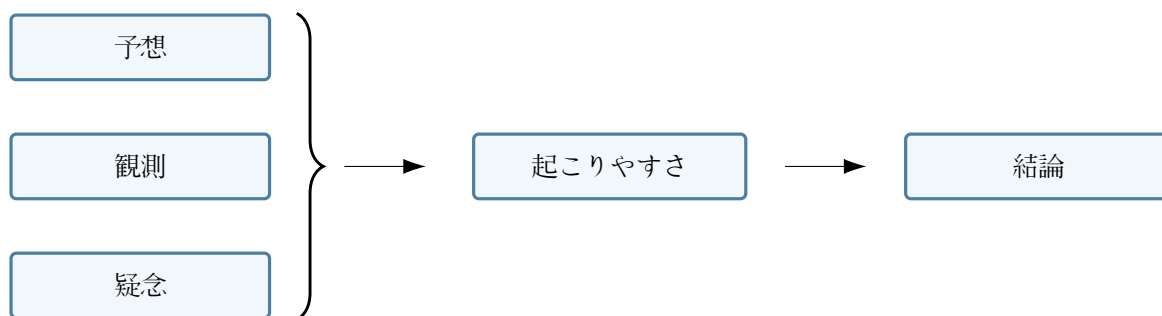
実際、30回中15回表が出る確率でさえ約14%にすぎないことを考えると、この結果はそれほど不自然ではありません。この情報だけでは、2枚目のコインが不公平であると考えer十分な理由はありません。

1.1 統計的検定の非公式な構造

ここまでで、仮説検定と呼ばれる非常に重要な統計的手続きの基本的な考え方を見てきました。どちらの場合も、ある仕組みが提示され、観測した結果に基づいて「公平なコインを使っている」という仮定に疑問をもちました。また、どちらの場合にも何らかの確率計算を行い、異なる2つの結論を得ました。

1. 怪しいコインの場合、観測された結果は極めて起こりにくいと判断しました。これは $p = 0.5$ という予想に疑いを生じさせ、公平なコインであるという考えを棄却することにつながりました。一方で、
2. 修道女たちのコインの場合、観測された結果は $p = 0.5$ という確率と矛盾していなかったため、コインの公平性を棄却する理由はありませんでした。

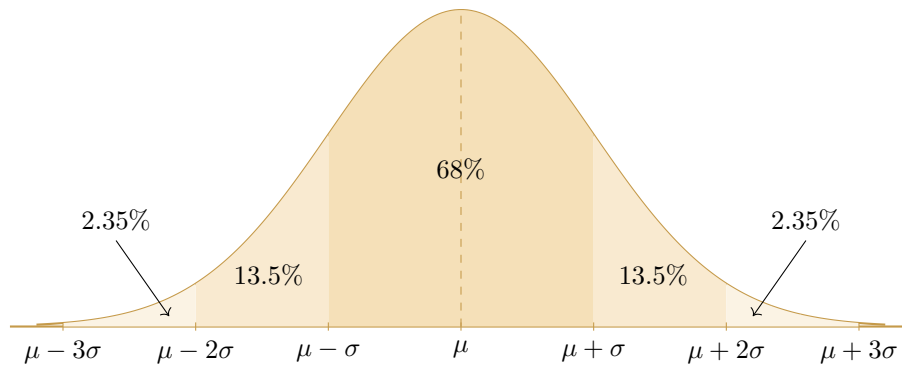
多くの仮説検定は、この一般的なパターンに従います。まず予想があり、次に疑いを生じさせる観測を行い、その後、最初の仮定を棄却することがどの程度妥当であるかを判断するための確率計算を行います。この概念的な流れは、次のようにまとめられます。



今後数回の講義では、この構造をさまざまな状況で詳しく調べます。今回は、この一般的な構造を説明することにします。そのために、まず正規分布の基本的な性質を復習します。

1.2 正規分布の簡単な復習

正規分布は、母平均 μ と母分散 σ^2 という2つの母数によって定まることを思い出してください。 $X \sim \mathcal{N}(\mu, \sigma^2)$ と書くとき、分布の x 軸は確率変数 X が取り得る値を表します。これは連続確率変数なので、 X がある範囲に入る確率は、グラフの下面積によって与えられます。したがって、グラフの形は、 X がその近くに現れる起こりやすさを表しています。正規分布は中心付近で最大となり、裾に向かうにつれて急速に小さくなるため、無作為に選ばれた X の値は、おそらく μ の近くに現れると考えられます。実際、第10回講義で学んだように、 μ から1標準偏差以内に入る確率は約68%です。

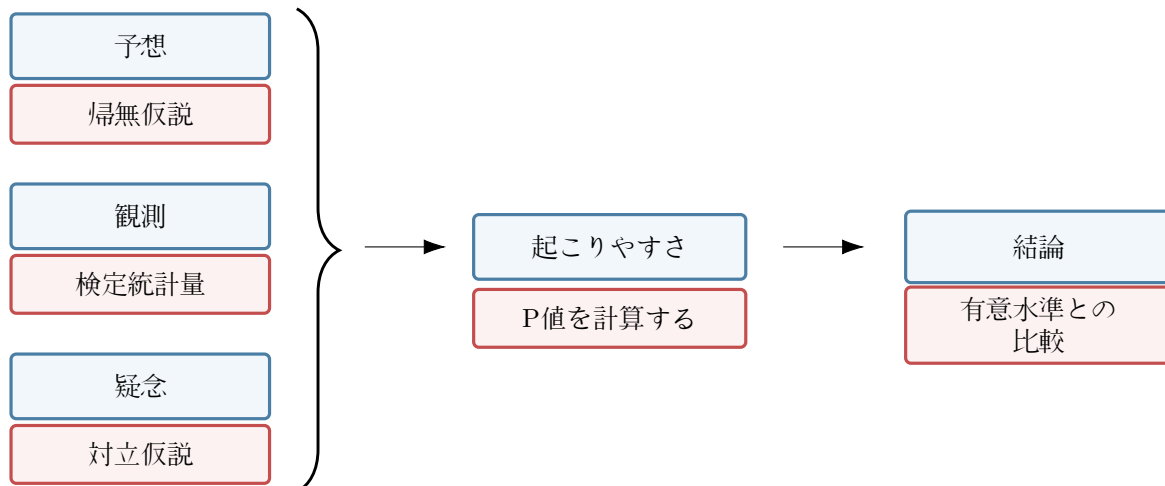


言い換えると、 X の値を何度も無作為に選ぶと、その値は約68%の確率で区間 $[\mu - \sigma, \mu + \sigma]$ に入ります。一方、無作為に選ばれた X の値が中心から遠く離れることは、非常に起こりにくいと考えられます。この観察が、仮説検定の考え方の基礎となります。

2 仮説検定

2.1 統計的検定の一般的な構成

はじめに、統計的検定を行うための一般的な流れを説明しました。以下は、同じ流れに正しい専門用語を付け加えたものです。



これらの用語を一つずつ説明します。

重要な用語

- 帰無仮説 (H_0) : 検定したい主張。
- 対立仮説 (H_1) : 私たちの疑いを表す、帰無仮説と競合する主張。
- 検定統計量 : 標本データから得られる数値的な情報。
- P値 : H_0 が真であると仮定したときに、観測結果と同じか、それ以上に極端な結果が得られる確率。
- 有意水準 (α) : 何を「起こりにくすぎる」とみなすかを定める確率の基準値。

2.2 P値

上の定義によれば、仮説検定におけるP値は、ある結果が起こる確率だけを計算するものではありません。P値は、帰無仮説が真であると仮定したうえで、観測された標本データと同じか、それより悪い結果が得られる確率を計算します。第1節では、2枚のコインについて二項確率を計算し、この考え方の原型を見ました。実際には、これらの確率計算は、まだ仮説検定のP値そのものではありません。P値を求めるには、さらに詳しい計算が必要です。

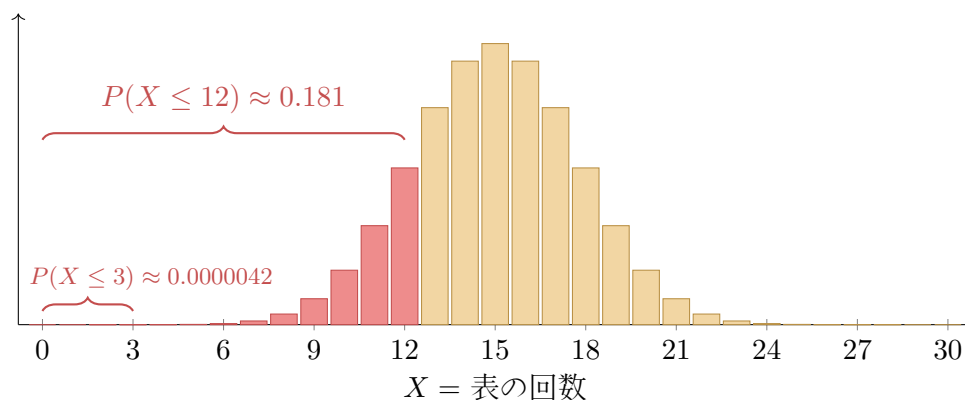
怪しいコインについて、実際に観測された結果は $X = 3$ 、つまり表が3回でした。この観測に対するP値を求めるときには、コインが公平、つまり $p = 0.5$ であると仮定し、表が3回以下となる確率を計算します。

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \approx 0.0000042.$$

2枚目のコインについては、 $X = 12$ が観測されました。この場合も、P値は表が12回以下となる確率を計算することで求められます。

$$P(X \leq 12) = P(X = 0) + P(X = 1) + \dots + P(X = 12) \approx 0.181.$$

図で表すと、 $n = 30$ 、 $p = 0.5$ の二項分布の左側の裾に含まれる確率を計算しています。



2.3 有意水準

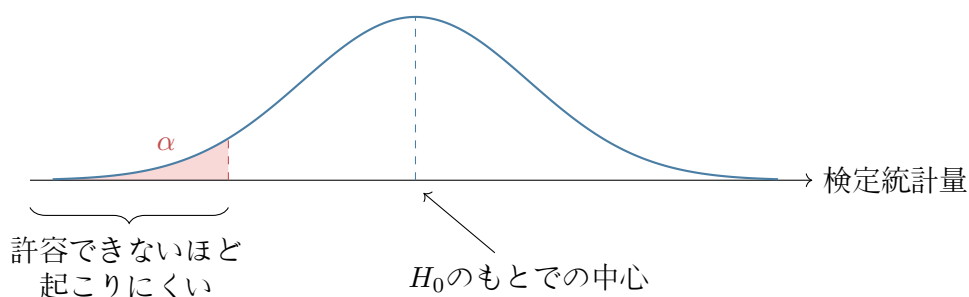
仮説検定では、帰無仮説が実際に真であると仮定し、その仮定からどのような結果が生じるかを調べます。第1.2節で述べたように、通常は、無作為に選ばれた検定統計量は、その確率分布

の中心に比較的近いところに現れると考えられます。特に、その検定統計量が正規分布に従う場合はそうです。もし裾のかなり遠いところにある結果が観測されたなら、その結果が元のモデルのもとで生じる確率は非常に低いことになります。

もちろん、無作為性の性質上、偶然に非常に極端な結果が得られることは常に可能です。本当に考えるべき問いは、次のとおりです。

帰無仮説に何か問題があると考え始めるには、どの程度運が悪ければよいのでしょうか。

有意水準 α は、「運が悪すぎる」と判断するための基準値を与えることで、この問いに答えます。



標本データが得られた後、P値は、データの真の中心が H_0 によって与えられていると仮定したときに、この標本またはそれより悪い結果が得られる確率を表します。この確率が許容できないほど低い、つまり α より小さい場合、 H_0 が主張する母数の値はおそらく誤っていると結論づけます。最終的な判断は、次のようにまとめられます。

判定規則

- 計算されたP値が α 以下である場合、その観測結果は H_0 のもとでは起こりにくすぎるため、 H_0 を棄却します。
- 計算されたP値が α より大きい場合、その観測結果は十分に起こりにくいとはいえないため、 H_0 を棄却しません。

実際には、検定を始める前に有意水準を選びます。よく用いられる値は、

$$\alpha = 0.01, \quad \alpha = 0.02, \quad \alpha = 0.05$$

です。重要なのは、「 H_0 を棄却しない」ということは、 H_0 が真であると証明したことを意味しないという点です。これは単に、標本が H_0 に反する十分に強い証拠を与えていないということです。

2.4 さまざまな種類の検定

対立仮説 H_1 は、常に帰無仮説と反対の主張です。実際、「等しい」という主張に反対する方法はいくつかあります。通常、帰無仮説は、母数の真の値がある特定の主張された値に等しいと述べます。この主張が誤っている可能性には、次の3つがあります。

- 真の値が、主張された値より小さい。
- 真の値が、主張された値より大きい。
- 真の値が、主張された値と単に異なる。

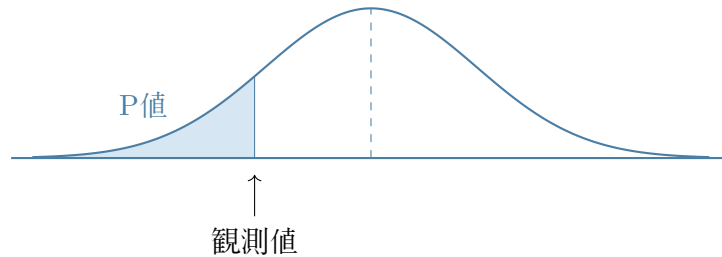
これらをそれぞれ、左側検定、右側検定、両側検定といいます。ここからは、対立仮説によって、与えられた検定の正しいP値の計算方法がどのように決まるかを示します。

2.4.1 左側検定

左側検定では、真の値が帰無仮説によって主張された値より小さいかどうかを調べます。形式的には、検定の最初に次のように書きます。

$$H_0: \text{真の値} = \text{主張された値}, \quad H_1: \text{真の値} < \text{主張された値}.$$

名前が示すとおり、左側検定では、観測データのP値を求めるために、観測値の左側の面積を計算します。

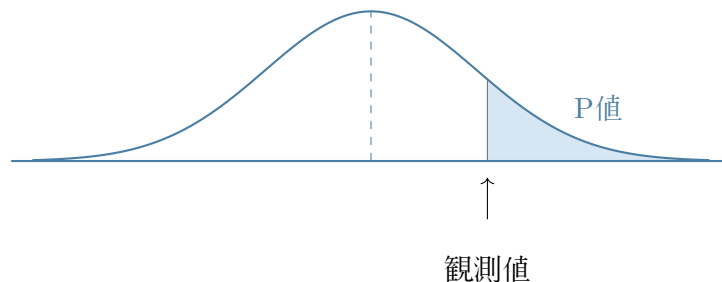


2.4.2 右側検定

右側検定では、対立仮説は、真の値が帰無仮説によって主張された値より大きいかどうかを問います。

$$H_0: \text{真の値} = \text{主張された値}, \quad H_1: \text{真の値} > \text{主張された値}.$$

この場合、P値は、観測された標本の検定統計量より右側の面積です。

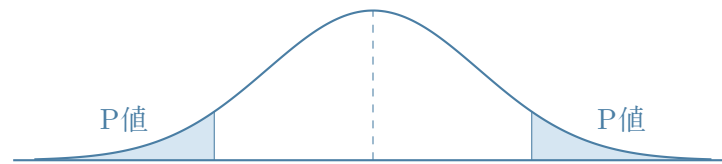


2.4.3 両側検定

両側検定では、真の値が帰無仮説によって主張された値と異なるかどうかを調べます。

$$H_0: \text{真の値} = \text{主張された値}, \quad H_1: \text{真の値} \neq \text{主張された値}.$$

2つの数 a と b が互いに異なるなら、 $a < b$ または $b < a$ のどちらかであることを思い出してください。したがって、両側検定では、両方の裾の確率を同時に合計してP値を求めます。



2.5 第1種の過誤と第2種の過誤

仮説検定には、その性質上、常にある程度の不確実性が含まれます。有意水準より小さいP値が得られたとしても、それによって帰無仮説が偽であることが証明されるわけではありません。これは、標本が帰無仮説に反する証拠を与えていることを意味するだけです。同様に、帰無仮説を棄却しなかったとしても、帰無仮説が本当に真であることが証明されたわけではありません。単に、帰無仮説を棄却するための十分な証拠がないということです。

このような不確実性があるため、偶然に誤った結論を出してしまう可能性が常にあります。仮説検定を行うときに起こり得る誤りには、主に2つの種類があります。これらを第1種の過誤と第2種の過誤といいます。

第1種の過誤と第2種の過誤

- 第1種の過誤は、 H_0 が実際には真であるにもかかわらず、 H_0 を棄却したときに起こります。
- 第2種の過誤は、 H_0 が実際には偽であるにもかかわらず、 H_0 を棄却しなかったときに起こります。

第1種の過誤は、最初の標本で統計的に非常に起こりにくい結果が偶然得られた場合に起こる可能性があります。たとえば、コインを30回投げ、裏が27回、表が3回出るとは、厳密には不可能ではありません。しかし、これは極めて起こりにくいため、そのコインが公平であるという主張を棄却せざるを得ません。ただし、その判断をしたとしても、単に非常に起こりにくい結果が偶然得られただけではないという保証はありません。このような状況で帰無仮説を棄却することは、第1種の過誤の例となる可能性があります。

第2種の過誤は、帰無仮説と矛盾しない標本データが得られたときに起こります。このような場合、帰無仮説が実際には偽であったとしても、それを棄却する十分な理由がありません。これは、標本から得られたP値が α より大きい場合に起こる可能性があります。これを説明するために、修道女たちのコインでは、30回中12回表が出たことを思い出してください。この観測結果は公平なコインと十分に矛盾しないため、コインが公平であるという主張を棄却する理由はないと結論づけました。しかし、実際にはそのコインの確率が $P(H) = 0.4$ だったとしたらどうでしょうか。この標本データからは、公平なコインである可能性を排除できませんが、実際にはその主張が偽である可能性があります。

仮説検定から生じる可能性のある結果は、次の4つです。

	H_0 を棄却しない	H_0 を棄却する
H_0 が真	正しい判断	第1種の過誤
H_0 が偽	第2種の過誤	正しい判断

有意水準 α は、第1種の過誤を犯す確率を制御します。たとえば、 $\alpha = 0.05$ を用いることは、検定を繰り返し行ったときに、真である帰無仮説を約5%の割合で棄却するように検定が設計されていることを意味します。

3 標本検定：2枚のコイン

ここからは、コイン投げの例を正式な仮説検定としてどのように表すかを示します。後のために、以前の講義で真の二項確率 p を推定するときに扱った確率変数 $\hat{P}$ を用いて、実験を正規近似の形で表します。以下では、標本実験に対応するP値をそのまま示し、P値の実際の計算は次の講義に回します。ここでは、確率変数 $\hat{P}$ が、次の母数をもつ近似的な正規分布に従うことを思い出せば十分です。

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

重要なのは、真の二項確率 p がこの分布の中心に位置するということです。したがって、帰無仮説は、この正規分布の中心に関する主張として図示できます。観測された標本データは、最終的にこの分布の x 軸上の点として現れます。言い換えると、上の正規分布は、私たちの「運」をどのように整理すればよいかを示しています。

観測データを図示する前に、まず検定を正しく表す方法を示します。

3.1 怪しいコイン

最初のコインに戻ります。

ステップ1：仮説を立てる。

まず、コインが公平であると仮定します。しかし、表が出る確率は実際には0.5より小さいのではないかと疑っています。したがって、帰無仮説と対立仮説は、それぞれ

$$H_0 : p = 0.5, \quad \text{および} \quad H_1 : p < 0.5$$

となります。対立仮説を見ると、これは左側検定であることが分かります。

ステップ2：検定統計量を記録する。

標本データでは、 $n = 30$ 回のコイン投げで、表が $r = 3$ 回観測されました。したがって、標本比率は

$$\hat{p} = \frac{r}{n} = \frac{3}{30} = 0.1$$

となります。

ステップ3：P値を計算する。

通常であれば、ここでP値を直接計算します。しかし、ここでは正規近似による計算結果だけを示します。

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

したがって、

$$P(\hat{P} \leq 0.1) \approx 0.0000134.$$

左側検定を行っているため、観測された $\hat{p}$ より左側の値を計算していることに注意してください。

ステップ4：有意水準と比較する。

有意水準 $\alpha = 0.05$ を用いると、明らかに

$$0.0000059 < 0.05$$

です。したがって、帰無仮説 H_0 を棄却します。

解釈

30回のコイン投げで表が3回しか出なかったという標本データは、コインが公平であるという仮説に反する非常に強い証拠を与えています。むしろ、このデータは、表が出る確率が0.5より小さいという疑いを支持しています。

3.2 修道女たちのコイン

次に、2枚目のコインについて同じ検定を行います。

ステップ1：仮説を立てる。

再び、このコインが表の出にくいコインであるかどうかを検定します。

$$H_0 : p = 0.5, \quad \text{および} \quad H_1 : p < 0.5.$$

ステップ2：検定統計量を記録する。

この場合、 $n = 30$ 回のコイン投げで、表が $r = 12$ 回観測されたので、

$$\hat{p} = \frac{r}{n} = \frac{12}{30} = 0.4$$

となります。

ステップ3：P値を計算する。

帰無仮説のもとで連続性補正を行うと、P値はおよそ

$$P(X \leq 12) \approx P(Z \leq -0.913) \approx 0.181$$

となります。

ステップ4：有意水準と比較する。

再び有意水準 $\alpha = 0.05$ を用いると、

$$0.181 > 0.05$$

となります。したがって、 H_0 を棄却しません。

解釈

30回のコイン投げで表が12回出たという標本は、表が出る真の確率が0.5より小さいと結論づけるための十分に強い証拠を与えていません。これはコインが公平であることを証明するものではありません。この観測結果が公平なコインと十分に矛盾していないということを意味するだけです。

3.3 検定の図示

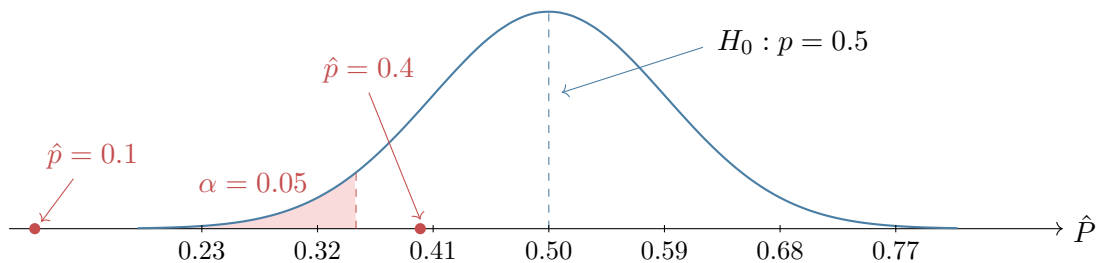
帰無仮説のもとで、観測された標本比率の標本分布は近似的に

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right)$$

となります。この分布の標準偏差は

$$\sqrt{\frac{(0.5)(0.5)}{30}} \approx 0.091$$

です。怪しいコインでは $\hat{p} = 0.1$ が得られました。これは左側の裾の非常に遠いところにあります。修道女たちのコインでは $\hat{p} = 0.4$ が得られました。これは中心より下にありますが、特に極端な値ではありません。



図から分かるように、修道女たちのコインに対応する左側P値の面積は、基準値 $\alpha = 0.05$ より大きくなります。したがって、そのコインが公平であるという仮説を棄却する十分な理由はありません。一方、怪しいコインの場合、観測された標本比率 $\hat{p} = 0.1$ は分布の左側の裾の非常に遠いところにあります。つまり、最初に標本データを取ったときに、このような結果が得られることは極めて起こりにくいということです。0.1より左側の面積は0.05よりはるかに小さいため、怪しいコインが公平であるという仮説を棄却しました。