1-6- 4　推測統計学

推測統計の目的

- n 個のデータ $x_1, x_2, \dots, x_n$ が与えられているとします。

- 推測統計学では、データがある確率変数の「観測値」として得られているとモデル化し、その確率変数の従う確率分布に関する情報をデータから抽出することを目指します。

- より正確には、 n 個の確率変数 $X_1, X_2, \dots, X_n$があって、観測データ $x_1, x_2, \dots, x_n$ が標本点ωが得られたときの実現値となっていると考えます。すなわち、$x_i = X_i(\omega), i = 1,2, \dots, n$ とモデル化します。

- 最も基本的な状況として、 $X_1, X_2, \dots, X_n$が独立同分布である場合を考えます。これらの確率変数が従う共通の確率分布をQとしたとき、Qに関する情報をデータ$x_1, x_2, \dots, x_n$ から調べることがここでの目標です。
  - Qは母集団分布、あるいは単に母分布と呼ばれることがあります。

例：コイン投げ

- 表が出る確率がわからないコインを何回か投げて、表が出る確率θを調べることを考えます。
- ５回投げた際の結果が、「表、裏、表、表、表」だったとします。
- 表を１、裏を０とそれぞれ記録することにすれば、観測データとして
$$(x_1, x_2, x_3, x_4, x_5) = (1,0,1,1,1)$$
が得られたことになります。
- 推測統計では、これらを「確率θで１、確率$1-\theta$で０をとる５つの独立な確率変数X_1, X_2, X_3, X_4, X_5の観測値」としてモデル化します。
- このとき、母分布Qはベルヌーイ分布Bn$(1,\theta)$となります(67~69ページ参照)。

推測統計の基本概念

- 推定：データをもとに、母分布Qを特徴づけるパラメータ（上の例ではコインの表が出る確率 θ）の値を求めることが目標です。
  - 点推定：パラメータをある１つの値で指定する推定方式です。
    例）コインの表が出る確率θの推定値は0.8である
  - 区間推定：ある確率以上で真のパラメータを含むことが保証される範囲を求める推定方式です。
    例）コインの表が出る確率θは0.7~0.9の範囲に含まれる確率は95%以上である

- 仮説検定：母分布Qに関する仮説がデータと整合的か否かを確率計算で判断することが目標です。
  - 例）コインの表が出る確率θが0.5であると言えるか否かをデータから判断する

平均の点推定

- 確率分布を特徴づける最も基本的な特徴量は平均（期待値）と分散・標準偏差です。

- ここではまず、母分布 Q の平均 μ の点推定について考えます。
  - $X_1, X_2, \dots, X_n$はいずれもQに従っていますので、すべての$i = 1,2,\dots,n$について $\mu = E[X_i]$ です。

- μ の値を指定する「自然な」方法として、データの平均

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

が考えつきます。

平均の点推定

- $\bar{x}$ は、次の確率変数の観測値となっています。

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- このように、推測統計の枠組みでは、点推定によって指定する値は、一般に何らかの確率変数の観測値となっています。
  - 前者を(点)推定値、後者を(点)推定量と区別することがあります。
  - 今の場合、$\bar{x}$ が推定値で、$\bar{X}$ が推定量です。$\bar{X}$ は標本平均と呼ばれることがあります。

- 推測統計では、推定量の確率的な性質を調べることで、点推定の「よさ」を判断します。

一致性と不偏性

- 標本平均 $\bar{X}$は、点推定量として次の 2 つのよい性質を持っています。

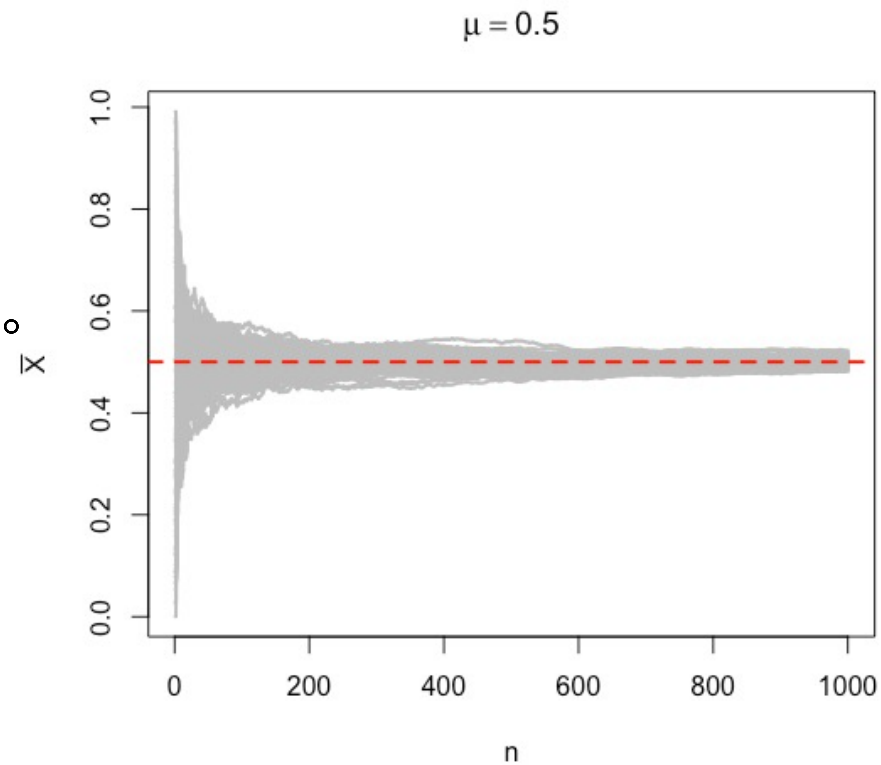
① 一致性：データ数nが大きくなるに従って、$\bar{X}$の値が推定したい真のパラメーターμに近づいていく（大数の法則による）

② 不偏性：確率変数 $\bar{X}$ の期待値が、推定したい真のパラメーターμに一致する：$E[\bar{X}] = \mu$

一致性

- 母分布がベルヌーイ分布Bɴ(1,0.5)の場合に、「標本平均 $\bar{X}$をサンプル数を1から1000まで増やしながら計算し、その推移をプロットする」という試行を100回繰り返して得られたグラフを右図に示します。この場合の平均は$\mu = 0.5$です。

- 100回のどの試行に対するグラフも、サンプル数nが大きくなるに従って、最終的に$\mu = 0.5$に近づいていく様が見てとれます。



不偏性

- 一般に、未知パラメータθを推定量 $\hat{\theta}$ で推定した際の推定誤差$\hat{\theta}-\theta$は、以下のように「確率誤差」と「バイアス」に分解できます：

$$\hat{\theta}-\theta=\underbrace{\left(\hat{\theta}-E\left[\hat{\theta}\right]\right)}_{\text{確率誤差}}+\underbrace{\left(E\left[\hat{\theta}\right]-\theta\right)}_{\text{バイアス}}$$

  - 確率誤差は、ランダムな要因で発生する誤差で、完全に制御することはできませんが、確率分布の計算で見積もることが可能です(後述)。
  - バイアスは、推定値を一定方向に偏らせる、非ランダムな誤差です。

- 推定量の不偏性とは、バイアスが0であることを意味します。

$$E\left[\hat{\theta}\right]-\theta=0 \Leftrightarrow E\left[\hat{\theta}\right]=\theta$$

分散の点推定

- 次に、母分布 Q の分散 σ^2 の点推定について考えます。
  - $X_1, X_2, \ldots, X_n$はいずれもQに従っていますので、すべての$i = 1,2, \ldots, n$について $\sigma^2 = V[X_i]$ です。

- 平均の点推定の類推から、σ^2の推定量としては、データの分散に対応する確率変数
$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$
を考えるのが自然なように思えます。

- しかし、S^2は一致性は持ちますが、不偏性を持たないことが、S^2の期待値を計算すると以下のようになることからわかります。
$$E[S^2] = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n}$$

分散の点推定

- 最後の式から、 S^2は平均的には真の分散 σ^2をσ^2/n だけ過小推定することになります。

- 中央の式から、このバイアスを消すにはS^2に$n/(n-1)$をかけた推定量を考えればよいことになります。

$$s^2 = \frac{n}{n-1}S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

  - s^2は一致性も持っています。
  - s^2は標本分散と呼ばれることもありますが、 S^2と区別するために不偏分散と呼ぶことも多いです。

区間推定

- 未知パラメータθを推定量 $\hat{\theta}$ で点推定する際、通常は「推定誤差」 $\hat{\theta} - \theta$が発生します。
- 推定値がどの程度信頼に足るか評価するのは、この推定誤差の大きさを評価する必要があります。
- 推定誤差も確率変数ですので、「推定誤差が必ず収まる範囲」では、しばしば保守的すぎて、実質的に役に立ちません。
- 例）コインの表が出る確率を、10回投げたうちに表が出た割合で推定する場合、たとえ表が出る確率が0.9であっても、10回連続で裏が出て推定値は0となるかもしれないです。この場合、「推定誤差が必ず収まる範囲」は「0以上1未満」となり、役に立ちません。

区間推定

- 上の例では、10回連続で裏が出るのは100億分の1の確率しかないため、推定誤差の大きさが0.9となるような状況はほぼ起きません。

- このことから、実用的な評価を得るには、ほとんど起きない事象を無視すればよいことが示唆されます。

- 上の議論を踏まえて、統計学では以下のようにして推定誤差の大きさを評価します。
  - (小さい)確率 α を 1 つ定めて、α 未満の確率でしか起きない事象は無視することにします($\alpha = 0.05$や$\alpha = 0.01$がよく使われます)。
  - 推定誤差が $1-\alpha$ 以上の確率で収まる範囲を求めます。数式で述べると、

$$P\left(l \leq \hat{\theta} - \theta \leq u\right) \geq 1 - \alpha$$

    を満たすl, uをデータから求めます。

区間推定

- 実際に興味があるのは推定誤差そのものではなく未知パラメータθの値ですので、上の式をθが収まる範囲で述べ直すと、
$$P\left(\hat{\theta} - u \leq \theta \leq \hat{\theta} - l\right) \geq 1 - \alpha$$
となって、「$1 - \alpha$ 以上の確率で θ は$\hat{\theta} - u$ から$\hat{\theta} - l$ の範囲にある」となります。

- 推定誤差は相対誤差$\hat{\theta}/\theta - 1$などで評価する方法もありますので、より一般的な形で、推定誤差を考慮した推定方式を述べると、
$$P(L \leq \theta \leq U) \geq 1 - \alpha$$
を満たすL, Uをデータから求める、ということになります。

- このような推定方式を区間推定と呼びます。「L 以上 U 以下の範囲」という区間は$100(1 - \alpha)\%$信頼区間と呼ばれます。

例：正規母集団の平均の区間推定（分散既知）

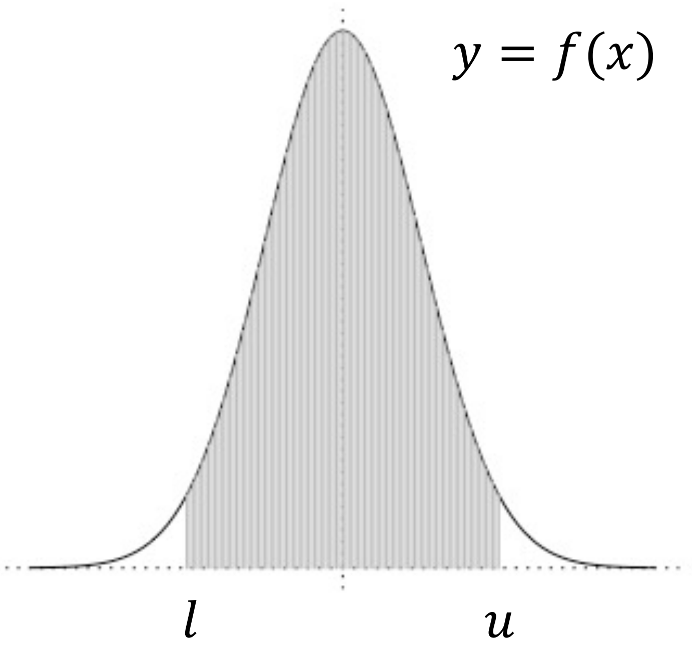
- 母分布が平均 μ 、分散 1の正規分布であるとします(76~78ページ参照)。

- 標本平均 $\bar{X}$ で μ を推定する場合、推定誤差を$\sqrt{n}$ 倍した量 $\sqrt{n}(\bar{X}-\mu)$ は平均0、分散1の正規分布（これを標準正規分布と呼びます）に従うことが知られています。

- 従って、$\sqrt{n}(\bar{X}-\mu)$ が l から u の範囲に含まれる確率は
$$P(l \leq \sqrt{n}(\bar{X}-\mu) \leq u) = \int_l^u f(x)dx$$
のように求めることができます。ここで、$f(x)$は標準正規分布の確率密度関数を表します：
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

例：正規母集団の平均の区間推定（分散既知）

- 上の式は、確率 $P(l \leq \bar{X} - \mu \leq u)$が$y = f(x)$のグラフと$x$軸で l から u の範囲において囲まれる部分の面積と一致します。（下図斜線部）
- 右図の斜線部分の面積が$1 - \alpha$以上になるように、 l, uを決めてやればよいことになります。
- 通常、推定誤差がプラス側に出るのとマイナス側に出るのは等価に見ますので、右図の左右の白抜き部分の面積がそれぞれ$\alpha/2$となるようにl, uを決めます。そのようなl, uの値をそれぞれ標準正規分布の下側100α/2%点、上側100α/2%点と呼びます。



例：正規母集団の平均の区間推定（分散既知）

- 例えば、$\alpha = 0.05$に対応する標準正規分布の下側2.5%点、上側2.5%点は、それぞれおよそ-2、2です。（下表参照）

α	0.1	0.05	0.01
下側$100\alpha/2$%点	-1.64	-1.96	-2.58
上側$100\alpha/2$%点	1.64	1.96	2.58

- 従って、

$$P(-2 \leq \sqrt{n}(\bar{X} - \mu) \leq 2) \approx 0.95$$

となりますから、μの95%信頼区間は、およそ

$$\bar{X} - \frac{2}{\sqrt{n}} \text{ 以上 } \bar{X} + \frac{2}{\sqrt{n}} \text{ 以下の範囲}$$

となります。

例：正規母集団の平均の区間推定（分散既知）

- より一般に、μの$100(1-\alpha)$%信頼区間は、標準正規分布の下側$100\alpha/2$%点、上側$100\alpha/2$%点をそれぞれ l_α, u_αと書くと、
$$\bar{X}-\frac{u_\alpha}{\sqrt{n}} \text{ 以上 } \bar{X}-\frac{l_\alpha}{\sqrt{n}} \text{ 以下の範囲}$$
で与えられます。

- ここまでの議論では母分布の分散を1としていましたが、より一般に分散がσ^2であっても、その値がわかっていれば、$\sqrt{n}(\bar{X}-\mu)/\sigma$ が標準正規分布に従うという事実を利用して、μの信頼区間を構成できます。

- この場合、μの$100(1-\alpha)$%信頼区間は、
$$\bar{X}-\sigma\frac{u_\alpha}{\sqrt{n}} \text{ 以上 } \bar{X}-\sigma\frac{l_\alpha}{\sqrt{n}} \text{ 以下の範囲}$$
で与えられます。

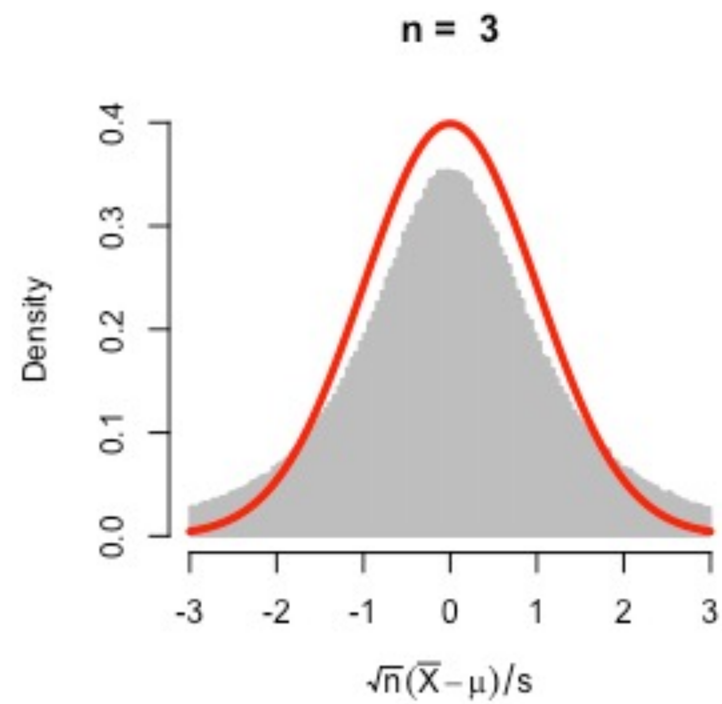
正規母集団の平均の区間推定（分散未知）

- 応用上は、母分布の分散 σ^2 の値がわかっているということはほとんどありません。

- その場合、 σ^2 の値を不偏分散 s^2 で推定するのが自然です。

- しかし、 $\sqrt{n}(\bar{X}-\mu)/\sigma$ の σ を s で置き換えた量 $\sqrt{n}(\bar{X}-\mu)/s$ は標準正規分布に従いません。

- 右図に $n=3$ の場合に $\sqrt{n}(\bar{X}-\mu)/s$ を100万個シミュレーションして描いたヒストグラム（灰）と標準正規分布の確率密度関数（赤）を描画していますが、明らかにずれています。



χ^2分布

- $\sqrt{n}(\bar{X}-\mu)/s$ は自由度$n-1$のt分布と呼ばれる確率分布に従うことが知られています。 t分布を定義するために、まずはχ^2分布と呼ばれる確率分布を定義します。

- $Z_1, Z_2, \ldots, Z_k$を標準正規分布に従う独立なk個の確率変数とします。

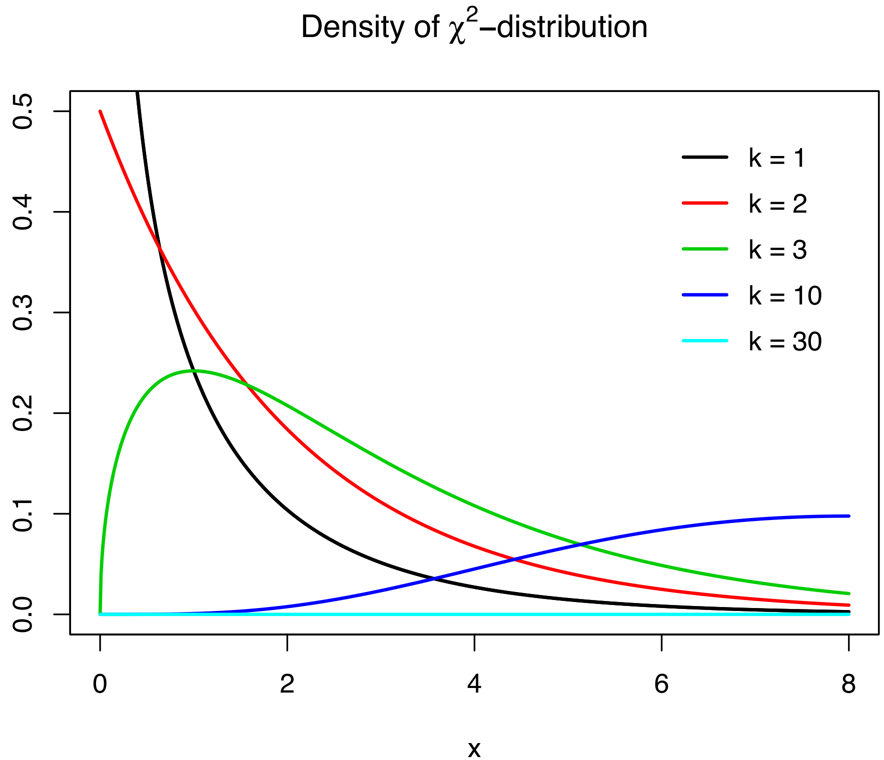
- このとき、確率変数

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

の従う確率分布を自由度kのχ^2分布と呼びます。

χ^2分布

- χ^2分布は連続分布である(確率密度関数を持つ)ことが知られており、確率密度関数は下図のようなグラフを持っています。



t分布

- 自由度kのt分布は次のように定義されます。

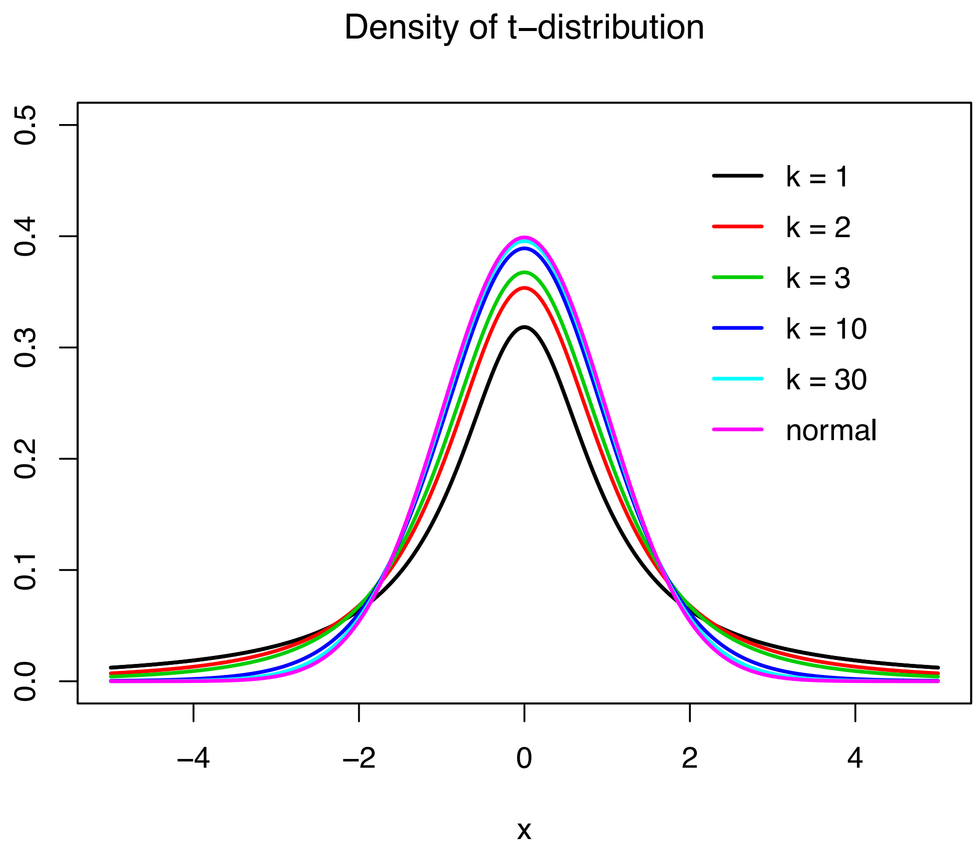
- Zを標準正規分布に従う確率変数、Yを自由度kのχ^2分布に従う確率変数とし、ZとYは独立であるとします。このとき、確率変数

$$t = \frac{Z}{\sqrt{Y/k}}$$

の従う確率分布を自由度kのt分布と呼びます。

t分布

- t分布は連続分布である(確率密度関数を持つ)ことが知られており、確率密度関数は下図のようなグラフを持っています。



正規母集団の平均の区間推定（分散未知）

- 母分布が平均μ、分散σ^2の正規分布ならば、$(n-1)s^2/\sigma^2$は自由度kのχ^2分布に従い、かつ$\bar{X}$と独立であることが知られています。

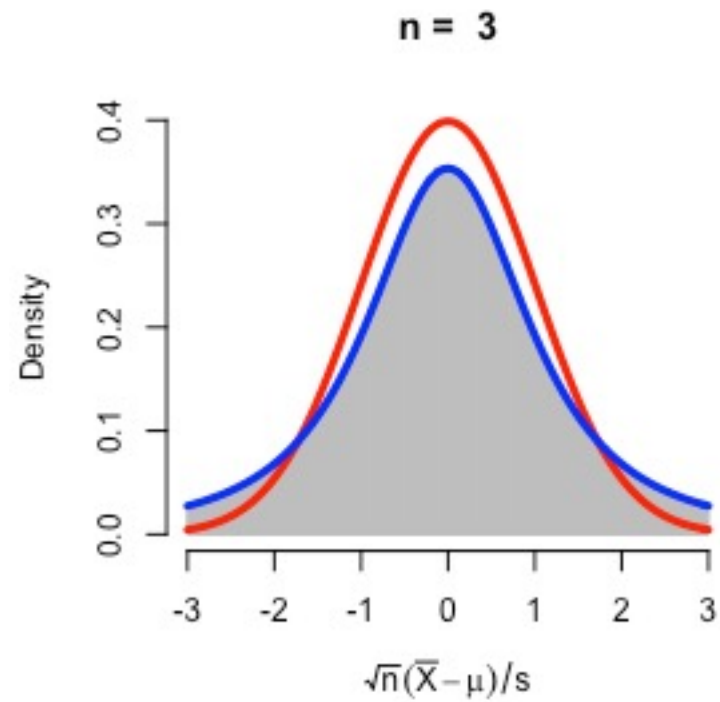
- $\sqrt{n}(\bar{X}-\mu)/\sigma$ は標準正規分布に従い、

$$\frac{\sqrt{n}(\bar{X}-\mu)}{s} = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\dfrac{(n-1)s^2/\sigma^2}{n-1}}}$$

と書き直せるので、t分布の定義から、$\sqrt{n}(\bar{X}-\mu)/s$ は自由度$n-1$のt分布に従うことがわかります。

正規母集団の平均の区間推定（分散未知）

- 先ほどの$n=3$の場合の$\sqrt{n}(\bar{X}-\mu)/s$のヒストグラムに自由度2のt分布の確率密度関数を上書き(青)すると、確かに重なることが確認できます。



- 従って、μの$100(1-\alpha)$%信頼区間は、自由度$n-1$のt分布の下側$100\alpha/2$%点、上側$100\alpha/2$%点をそれぞれtl_α, tu_αと書くと、

$$\bar{X}-s\frac{tu_\alpha}{\sqrt{n}} \text{ 以上 } \bar{X}-s\frac{tl_\alpha}{\sqrt{n}} \text{ 以下の範囲}$$

で与えられます。

仮説検定：考え方

- （統計的）仮説検定は、母分布に対する仮説の真偽を、データの観測値に基づいて検証するための枠組です。

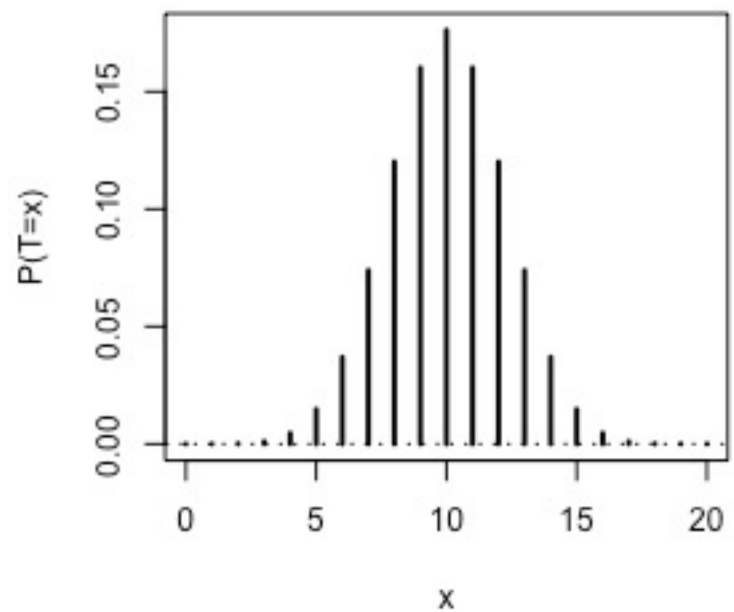
- 例）あるコインの表が出る確率が0.5であるか否か検証したいとします。コインを20回投げた結果は以下の通りでした。

1	2	3	4	5	6	7	8	9	10
表	表	裏	表	表	裏	表	裏	表	表
11	12	13	14	15	16	17	18	19	20
表	表	表	裏	裏	表	表	表	表	表

- この結果から、「このコインの表が出る確率は0.5である」という主張の信憑性を定量的に判断するにはどうすればよいでしょうか？

仮説検定：考え方

- 仮に、「このコインの表が出る確率は0.5である」という主張が正しかったとしましょう。
- この場合、コインを20回投げた際に表が出る総数をTとすると、Tは二項分布BN(20,0.5)に従う確率変数です。
- 従って、Tの期待値は10であり、Tがxという値をとる確率$P(T=x)$も右図のように計算できます。



仮説検定：考え方

- 特に、 Tの観測値が期待値10からいくつかずれる確率を具体的に計算することができます。

- 下表に、$k = 1,2, \dots, 6$に対して、Tの観測値が10から$\pm k$以上大きくずれる確率を示します。

k	1	2	3	4	5	6
確率	0.824	0.503	0.263	0.115	0.041	0.012

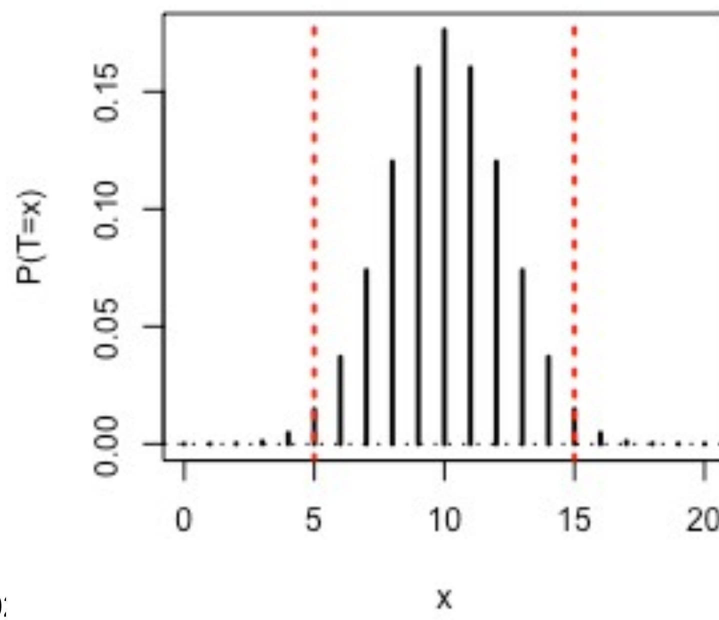
- ところで、今の例では、実際のデータから計算されるTの観測値は15ですので、「コインの表が出る確率は0.5である」という主張を信じる限り、上の議論からこのような結果が得られる確率は5%以下ということになります。

仮説検定：考え方

- 5%という確率を非常に小さいと考えると、このような結果が得られたのは、『運悪くそうなった』か、『そもそも「コインの表が出る確率は0.5である」という仮説が誤っている』のどちらかということになります。

- 運のせいにするのは非科学的なので、「コインの表が出る確率は0.5である」という仮説が誤っている、と判断するのが妥当だろうということになります。

赤点線から外側(赤点線含む)の値が観測される確率は5%以下です。



仮説検定：手順のまとめ

以上の議論をまとめると、以下のようになります。

1. 検証したい仮説 H_0 を設定します。
   - 上の例では「コインの表が出る確率は0.5である」が H_0 にあたります。
2. 「適当な」統計量 T を定め、仮説 H_0 の下で T が値をとる確率が低い領域 R を決めます。
   - 上の例では、 T は「コインを20回投げた際に表が出る総数」で、 R は「 T の値が、H_0 が正しい場合の期待値10から±4より大きくずれる」となります。(5%を「低い確率」と考えています。)
3. T の実際の値 t_0 をデータから計算します。
   - 上の例では $t_0 = 15$ です。
4. t_0 が R に含まれる場合、H_0 は誤りだと判断します。
   t_0 が R に含まれない場合、H_0 はデータとは「矛盾しない」と判断します。

仮説検定：用語

- 手順１で設定する仮説 H_0 を帰無仮説と呼びます。
- 手順２で定める「適当な」統計量 T を検定統計量と呼びます。
- 手順２の「低い確率」を有意水準と呼びます。
- 手順２で決める領域 R を棄却域と呼びます。
- 手順４で
  - H_0 は誤りだと判断することを、H_0 を棄却すると言います。
  - H_0 はデータとは「矛盾しない」と判断することを、H_0 を受容すると言います。

仮説検定：用語

- 検定結果には２種類の誤った判断がありえます（下表参照）
  - 両者の発生はトレードオフの関係にあります。

- 仮説検定の方針：
  - 第一種の過誤が起きる確率を有意水準以下に抑えた下で、第二種の過誤が起きる確率をできる限り小さくすることを目標とします。

	帰無仮説が偽	帰無仮説が真
帰無仮説を棄却	正しい判断	第一種の過誤
帰無仮説を受容	第二種の過誤	正しい判断

対立仮説

- 検定統計量と棄却域は、上の目標を達成できるように定める必要があります。
  1. 第一種の過誤が起きる確率を有意水準以下となるように定める必要があります。このためには、帰無仮説が正しい場合の検定統計量の確率分布（帰無分布）の評価が必要です。
  2. 第二種の過誤が起きる確率をできる限り小さくする必要があります。このためには、帰無仮説が誤りの場合に検定統計量が棄却しやすくなるように棄却域を設定する必要があります。

- 後者のためには、「帰無仮説が誤りの場合」に起こりうるシナリオを想定する必要があり、これを対立仮説と呼びます。
  - 上の例では、「コインの表が出る確率は0.5ではない」(帰無仮説の否定)が対立仮説でした。

対立仮説

- 仮説検定では、検定結果が積極的に支持されるのは、帰無仮説が棄却された場合だけです。
  - 「帰無仮説が正しい場合に誤って棄却してしまう」場合はまれであることは保証されていますが(有意水準以下の確率でしか起きない)、「帰無仮説が誤りだが受容してしまう」確率が小さいかはわからないからです。
  - この理由から、「立証したい主張」を対立仮説にとることが多いです。

- 何らかの理由で「帰無仮説が誤りの場合」に起こりうるシナリオに制限を加えることができる場合、その情報を対立仮説に反映させることで、第二種の過誤が起きる確率を小さくできることがあります。

対立仮説

- 例）先ほどのコイン投げの例で、検証の目的が「コインの表が出る確率が不正に高くないか確認する」ことであったとしましょう。
  - 例えば、コインを投げて表が出たら相手が勝利するようなゲームで、コインを用意したのが相手だった場合を想定します。
- この場合、「表が出る確率が0.5より小さいケース」は起きるシナリオとして想定する必要がありません。
  - もしそうであったとしても、こちらは不利益を被らないので、帰無仮説を棄却して「このコインは不正である」と主張する必要はないからです。
- 従って、対立仮説としては「コインの表が出る確率は0.5より大きい」をとればよいことになります。

対立仮説

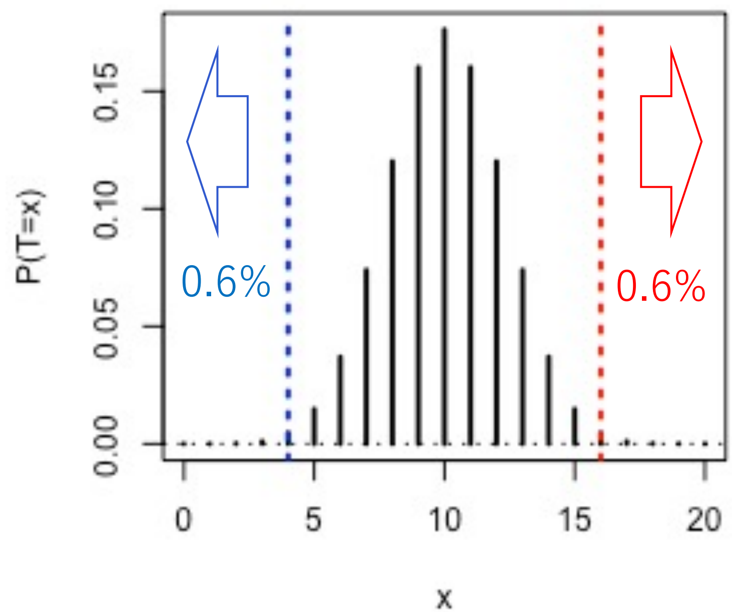
- この場合、対立仮説の下では、検定統計量 T (表の総数)は大きい値をとることが想定されるので、帰無仮説の下での期待値10より大きく外れた場合だけ帰無仮説を棄却すればよいことになります。

- 従って、棄却域は「T の値が$10 + k$ 以上となる」という形の領域に設定すればよいことになります。
  - kは帰無仮説の下でT の値が$10 + k$ 以上となる確率が有意水準以下となるように定めます。

対立仮説

- 例えば有意水準が1%の場合、棄却域は「T の値が16以上となる」です。
  - この場合、Tの観測値が16であれば帰無仮説を棄却できます。
  - 対立仮説が「コインの表が出る確率は0.5ではない」の場合、帰無仮説の下で「T の値が10から±6以上大きく外れる」確率は1.2%程度あるので、Tの観測値が16では帰無仮説を棄却できません。

赤点線から右側(赤点線含む)の値が観測される確率は0.6%程度ですが、赤・青点線の外側(点線含む)に値が観測される確率は1.2%程度です。



片側検定・両側検定

- 先の例のように、対立仮説の設定の仕方は、多くの場合棄却域の形状に影響を与えます。
  - 棄却域が左右両側にある検定は、両側検定と呼ばれます。（上の例では、対立仮説が「コインの表が出る確率は0.5ではない」の場合）
  - 棄却域が右側のみにある検定は、右片側検定と呼ばれます。（上の例では、対立仮説が「コインの表が出る確率は0.5より大きい」の場合）
  - 棄却域が左側のみにある検定は、左片側検定と呼ばれます。（上の例では、対立仮説が「コインの表が出る確率は0.5より小さい」の場合）

P値

- 検定統計量の値 t_0 がデータから計算されたとき、多くの場合、「帰無仮説を棄却できる有意水準の最小値」が計算できます。この値をP値と呼びます。
  - コイン投げの例では、「帰無仮説の下で、T が 10から$\pm|\mathrm{t}_0 - 10|$以上ずれる確率」がP値です。
  - 上の例の $t_0 = 15$ の場合は、P値 = 0.041 です。

- 検定のP値が計算できると、有意水準 α に対して以下の関係が成立します：

$$\text{P値} < \alpha \Leftrightarrow \text{有意水準}\ \alpha\ \text{で帰無仮説は棄却される}$$

- P値を使うと、有意水準を決めずに検定結果を要約できるため、「有意水準を調整して検定結果を操作する」といった不正を防げます。そのため、通常は検定結果だけでなくP値も報告します。