データ分析基礎
仮説検定

京都大学 国際高等教育院 附属データ科学イノベーション教育研究センター

関戸 啓人（せきど ひろと）

sekido.hiroto.7a@kyoto-u.ac.jp

仮説検定 – 余談

余談

★ **平成24年度全国学力・学習状況調査**

- ★ https://www.nier.go.jp/12chousa/12chousa.htm

★ 中学校の数学Bにて，ヒストグラムの問題が出題されている

★ **概要**

★ スキージャンプの2人の選手A，Bがそれぞれ何回か飛んだときの飛距離の記録のヒストグラムが与えられる

- ★ 問題1：ヒストグラムからそれぞれ何回飛んだのか求めよ
- ★ 問題2：それぞれの選手がもう1回飛んだときに，どちらの選手が飛距離が長いか，（2人のヒストグラムの特徴を比較して）予想せよ

仮説検定 – 例題

検定の例題

★ 例として次の問題を考える

- ★ 友達とサイコロを使うゲームをしているのだが，どうも負けがこんでいる
- ★ そこで，友達が使っているサイコロを1000回振って試したところ，6が207回も出た
- ★ これはおかしい，と友だちに詰め寄ったのだが…
- ★ 友達「6が1/5ぐらいの割合で出てるだけでしょう．1/5も1/6も大差ないし，そんなのよくあることだよ」
- ★ 自分の主張を正当化したい

検定の手順

★ **帰無仮説**として**示したいことの否定**を置き，**対立仮説**として**示したい**ことを置く

★ そして，**「帰無仮説が正しいとしたら，今起こった事象はとてもとても珍しいことである」**ということが示されれば，帰無仮説がおかしいのではないか，つまり，対立仮説が成り立つと結論づける

★ 今の例題では

- ★ **帰無仮説**：このサイコロで6が出る確率は1/6以下である
- ★ **対立仮説**：このサイコロで6が出る確率は1/6より大きい
- ★ 今起こった事象：このサイコロを1000回振ると6が207回出た
  - ★ このように事象から計算される量を**検定統計量**という

検定において計算すること

★ このサイコロを振った時，6が出る確率をpとして，6が207**回以上**出る確率を考える

★ 帰無仮説が正しいとした時，そのような確率が最も大きくなるのは$p = 1/6$のとき

★ $p = 1/6$のとき，このサイコロを1000回振ると6が207**回以上**出る確率がある一定値より小さい場合，帰無仮説を棄却する

★ 一定値としてよく用いられるのは5%，1%など（**危険率**，**有意水準**などという）

★ **帰無仮説が棄却された$\rightarrow$対立仮説が正しいと結論付ける**

★ **帰無仮説が棄却されなかった$\rightarrow$何も主張できない**

確率の計算

★ サイコロを振る試行は独立であると思えるから，N 回振って6が k 回出る確率は

$$B_{N,p}(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

★ よって示すべきことは，$N = 1000, p = 1/6$ で，k が207以上となる確率

$$\sum_{k=207}^{1000} B_{1000,1/6}(k) = \sum_{k=207}^{1000} \binom{1000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{1000-k}$$

がある一定値より小さいことである．例えば，一定値（危険率）は1%としよう．

★ **補足**：このような確率分布を二項分布というので，このような検定を**二項検定**という

検定において計算すること：なぜ以上が必要か

★1から100000の整数が書かれたボール10万個が箱のなかにあって，1個引く

- ★実際引いてみると77463というボールを引いた
- ★これは凄い！　77463というボールは1個しかないから1/100000の確率でしか起こらないことが今起きたぞ！！

★これでは何かがおかしい

検定において計算すること：なぜ以上が必要か

★ 1から100000の整数が書かれたボール10万個が箱のなかにあって，1個引く

★ **77463というボールを引くと予言してする**

★ 実際引いてみると77463というボールを引いた

★ これは凄い！　77463というボールは1個しかないから1/100000の確率でしか起こらないことが今起きたぞ！！

★ 確かに凄い（気がする）

★ **何が起こったら珍しいと思うかは，事前に決めて置かなければならない**

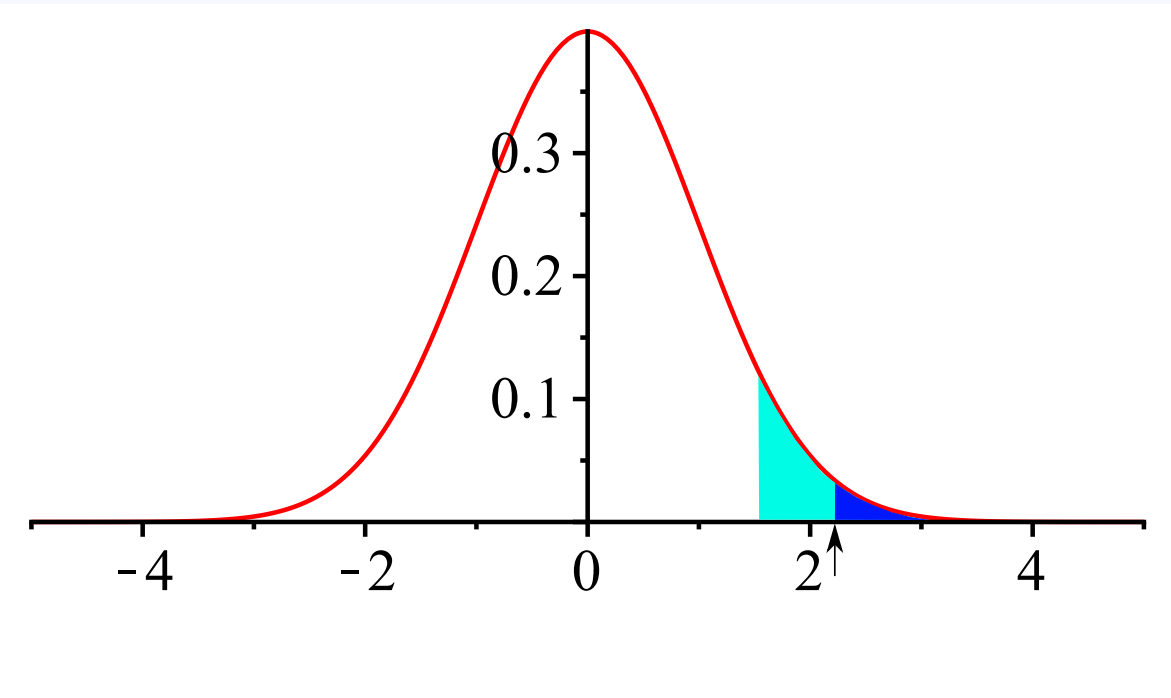
★ k回以上6が出る確率が1%以下となるような最小のkを求めて，k回以上6が出ると珍しいと思う

検定において計算すること：なぜ以上が必要か

★ k 回以上6が出る確率が1%以下となるような最小の k を求めて，k 回以上6が出ると珍しいと思う

★ そのような領域を**危険域**と呼ぶ

★ 実際に m 回6の目が出たとして，m 回以上でる確率が1%以下になることと，m が危険域に含まれることは同値



EXCELを用いた確率の計算

★ 以下の確率をEXCELで計算する

$$\sum_{k=207}^{1000} B_{1000,1/6}(k) = \sum_{k=207}^{1000} \binom{1000}{k} \left(\frac{1}{6}\right)^{k} \left(\frac{5}{6}\right)^{1000-k}$$

★ 方法1（ある程度新しいバージョンでのみ可能，EXCEL2013など）

- ★ セルに`=BINOM.DIST.RANGE(1000,1/6,207,1000)`と入力
  - ★ `=BINOM.DIST.RANGE(n,p,a,b)`で $\sum_{k=a}^{b} B_{n,p}(k)$

★ 方法2（ある程度古いバージョンでも可能）

- ★ セルに`=1-BINOM.DIST(206,1000,1/6,TRUE)`と入力
  - ★ `=BINOM.DIST(n,p,m,TRUE)`で $\sum_{k=0}^{m} B_{n,p}(k)$
  - ★ `=BINOM.DIST(n,p,m,FALSE)`で $B_{n,p}(m)$

Rを用いて二項検定を行う方法

★ binom.test(207, 1000, 1/6, "greater") と入力

★ "greater"：危険域を検定統計量が大きい側に取る

★ 実行結果（改行などは一部変更）

```
        Exact binomial test

data:  207 and 1000
number of successes = 207, number of trials = 1000, p-value = 0.0004981

alternative hypothesis:
true probability of success is greater than 0.1666667

    95 percent confidence interval:
 0.1860848 1.0000000

sample estimates:
probability of success
                 0.207
```

補足：中心極限定理とZ検定

★ 中心極限定理

★ 平均μ，分散σ^2の**独立**で全て同じ分布に従う確率変数
$X_1, X_2, \ldots, X_N, \ldots$に対し，

$$\frac{\sum_{k=1}^{N}(X_i - \mu)}{\sqrt{N}\sigma}$$

はNを大きくすると，平均0，分散1の**正規分布$\mathcal{N}(0,1)$に近づく**

★ 標準正規分布

★ Xが$\mathcal{N}(0,1)$に従うとき，

$$P(X < s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \mathrm{e}^{-x^2/2}\, \mathrm{d}x.$$

★ Z検定

★ **試行回数が大きい場合**は，**近似的に正規分布**を用いて，標本の平均が想定された母集団分布の平均と等しいかどうかを検定することもできる

仮説検定 – 検定の手続きと危険域の設定方法

検定の手続き

★ 主張したいこと「対立仮説」とその否定「帰無仮説」を設定する

★ 帰無仮説が正しいと仮定したら，ほぼ起こりえないようなことを設定する（危険域）

- ★ 帰無仮説が正しいと仮定したら，危険域に設定したことが起こる確率を危険率
- ★ 帰無仮説が正しくないならば，危険に設定したことは起こりそうになるように設定する

★ 実際に実験などをしてみて，危険域に設定したことが起こるかどうかを調査

- ★ 危険域に設定したことが起こったら，帰無仮説を棄却（対立仮説を採択）
- ★ 危険域に設定したことが起こらなかったら，何も言えない

検定のパターン 1

★ 二項検定の例で，検定する際，どのように考えれば良いかという考え方を述べた．しかし，どんな確率変数を考えれば良いか，などは，多少曖昧であり，いろいろ考えられる場合もある．ところが，実際には，検定の理論は大体の場合において確立されており，こういう場合にはこうやるのが「正解」とされるものがある．しばしば以下のように定式化できることがある．

★ 調べたいもの p が大きいほど，検定統計量（確率変数）X は大きい値を取りやすいとする．今，なんらかの試行をし，確率変数 X の値が確定した．

★ 先ほどの例では，p はサイコロを振って 6 の目が出る確率．確率変数 X はサイコロを 1000 回振って 6 が出る回数を表し，実際に試行をし，207 回出た．

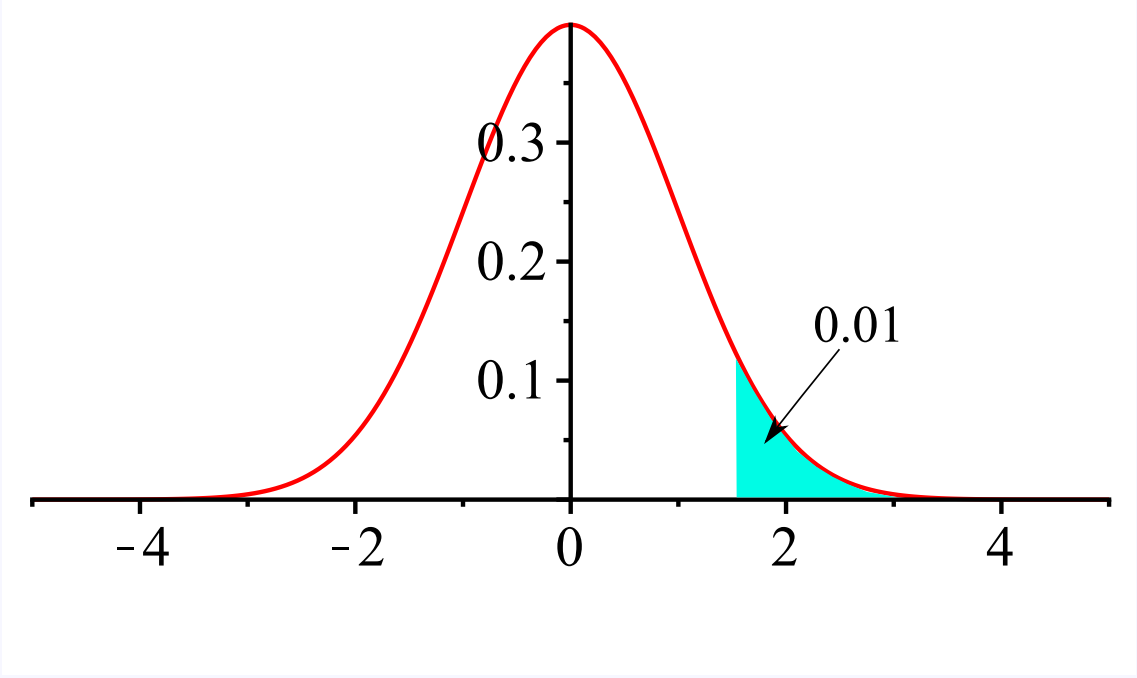
検定のパターン 2

★ 帰無仮説：$p \leq p_0$ (p_0 は定数)

★ 対立仮説：$p > p_0$

★ 有意水準：0.01 （珍しいと思う確率の閾値）

★ 実際に試行して得られた値が，以下の水色の領域にあれば帰無仮説は棄却される．以下のグラフは $p = p_0$ の時の確率変数 X の密度関数



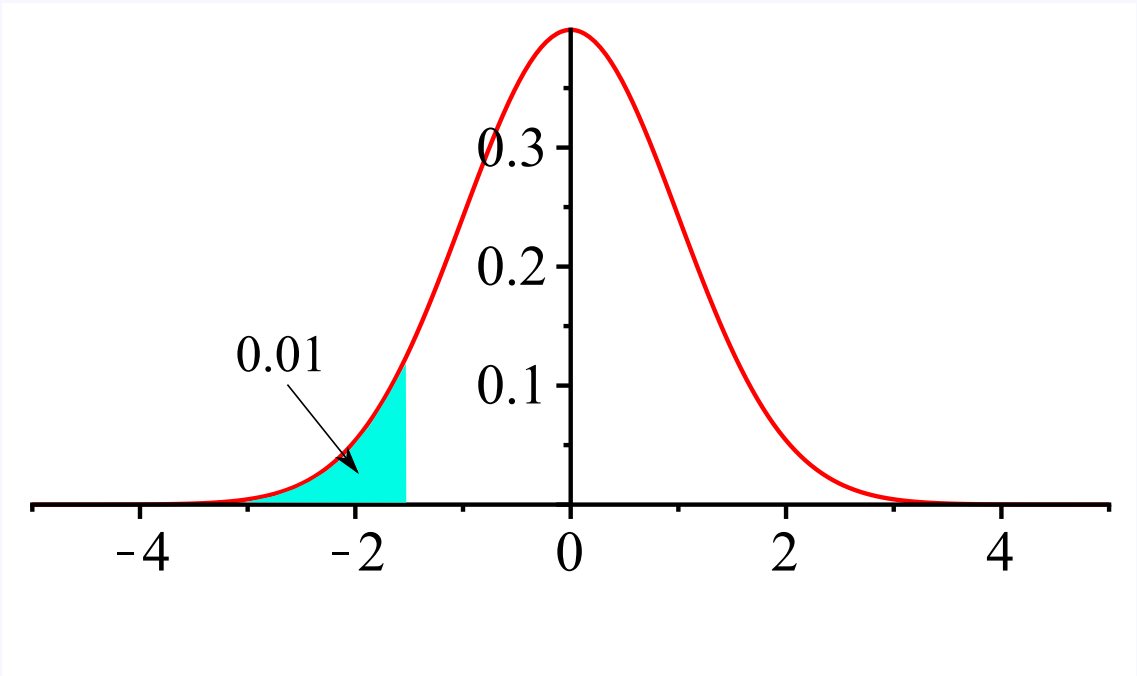
検定のパターン 3

★ 帰無仮説：$p \geq p_0$ (p_0 は定数)

★ 対立仮説：$p < p_0$

★ 有意水準：0.01 （珍しいと思う確率の閾値）

★ 実際に試行して得られた値が，以下の水色の領域にあれば帰無仮説は棄却される．以下のグラフは $p = p_0$ の時の確率変数 X の密度関数



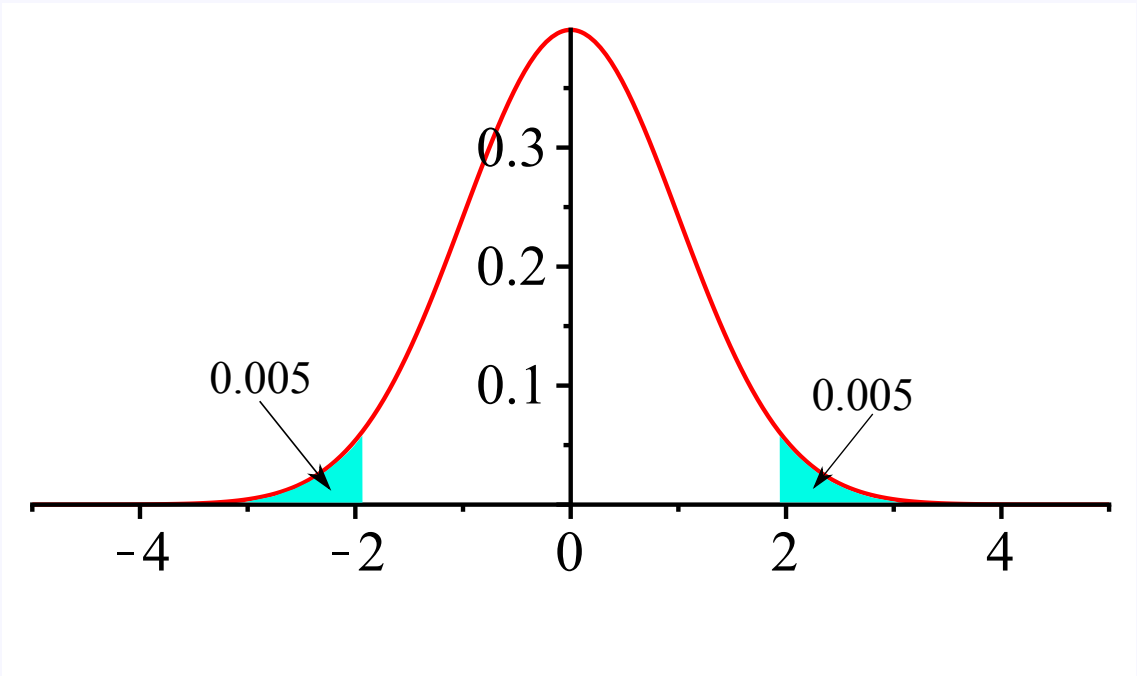
検定のパターン 4

★ 帰無仮説：$p = p_0$ (p_0 は定数)

★ 対立仮説：$p \neq p_0$

★ 有意水準：0.01 （珍しいと思う確率の閾値）

★ 実際に試行して得られた値が，以下の水色の領域にあれば帰無仮説は棄却される．以下のグラフは $p = p_0$ の時の確率変数 X の密度関数



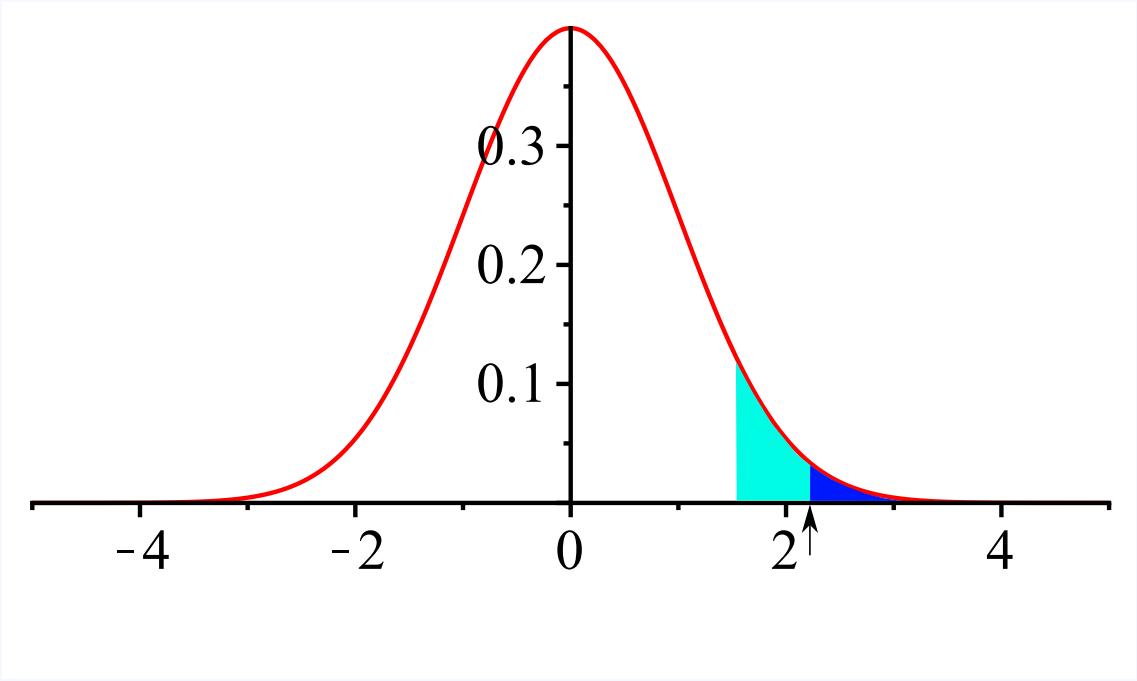
検定のパターン 5

★ 帰無仮説：$p \leq p_0$ (p_0 は定数)

★ 対立仮説：$p > p_0$

★ 有意水準：0.01 （珍しいと思う確率の閾値）

★ 実際に施行して得られた値が，矢印の値だとすると，青色の面積を P 値という．P 値が有意水準より小さければ帰無仮説は棄却される



仮説検定 – 仮説検定の手続き（もう少し厳密に）

検定の手順（ちょっと精密版）

★ 示したい主張を対立仮説と，その否定を帰無仮説と置く

★ 標本から計算可能な着目する確率変数（検定統計量）を設定する．帰無仮説が正しいという仮定のもとでの検定統計量の確率分布を求める

★ 検定統計量が極端な値となるような集合（危険域）を考える．帰無仮説が正しいという仮定のもとで検定統計量が危険域に含まれる確率を危険率という

★ 検定統計量を実際に評価し，危険域に含まれるなら帰無仮説を棄却する．そうでなければ何も主張できない

検定統計量と危険域の設定

- ★ 検定統計量としては，帰無仮説と対立仮説の違いが際立つものを選ぶ
- ★ 危険域としては，帰無仮説が正しいなら起こらなさそうで，対立仮説が正しいなら起こっても不思議でないものを選ぶ
  - ★ 実際には危険率を決めてしまえば，危険域は後は帰無仮説よりオートマチックに決まることが多い
  - ★ 危険率を α，検定統計量を X，帰無仮説を H_0 とすれば，帰無仮説が正しいという条件下でのある条件 A を満たす確率 $P(A|H_0)$ を用いて
    - ★ $P(X > c|H_0) = \alpha$ なる c を用いて $A = [c, \infty)$
    - ★ $P(X < c|H_0) = \alpha$ なる c を用いて $A = (-\infty, c]$
    - ★ $P(X < a|H_0) = P(X > b|H_0) = \alpha/2$ として $A = (-\infty, a] \cup [b, \infty)$

検定の計算方法

★ 危険率を α，危険域を $[c, \infty)$ とする

★ 検定統計量を評価して $X = x$ となったとする

★（危険率 α および X の確率分布がわかっているとする）

- ★ $P(X > x|H_0)$ を計算して α 以下なら帰無仮説を棄却
- ★ $P(X > c|H_0) = \alpha$ なる c を計算して $x \geq c$ なら帰無仮説を棄却

仮説検定 – 検定の例

平均，分散の性質

★ 確率変数 X, Y の平均を $\mathrm{E}[X], \mathrm{E}[Y]$，分散を $\mathrm{V}[X], \mathrm{V}[Y]$ とする

★ $X + Y$ の平均は $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

★ X と Y が独立ならば $X + Y$ の分散は $\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y]$

- ★ 一般的には $\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y] + 2\mathrm{Cov}[X, Y]$
- ★ ここで $\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$ は共分散
- ★ X が大きければ Y も大きくなる傾向があるというなら $\mathrm{Cov}[X, Y] > 0$
- ★ 共分散を -1 以上 1 以下になるように正規化したものは相関係数

$$\frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{V}[X]\mathrm{V}[Y]}}$$

平均，分散に関する検定

★ データ数が十分あるとき，あるいは，正規分布であるとき

★ 平均値がある値か，2群の平均は等しいかの検定

- ★ 分散が既知ならZ検定（正規分布）
- ★ 分散が未知ならt検定（t分布）

★ 分散がある値か，2群の分散は等しいかの検定

- ★ 1群の場合χ^2検定（χ^2分布）
- ★ 2群の場合F検定（F分布）

Rで計算するには

★ 各検定ごとに関数が用意されていることが多い

★ 分布名を`dist`として，以下の様な関数を用いることもできる

- ★ `ddist`：確率密度関数
- ★ `pdist`：累積確率分布関数：$P(X \leq x)$ の値
- ★ `qdist`：α 点：$P(X \leq x) = \alpha$ なる x

★ 分布名：

- ★ `norm`：正規分布
- ★ `chisq`：χ^2 分布
- ★ `t`：t 分布
- ★ `f`：f 分布

中心極限定理

★ $X_1, X_2, X_3, \ldots$ は同分布で独立とする

★ それぞれの平均を m，分散を σ^2 とする（存在を仮定）

★ $Y_n = \dfrac{(X_1 + X_2 + \cdots + X_n) - nm}{\sqrt{n}\sigma}$ とすると $n \to \infty$ の極限で Y_n は平均0分散1の正規分布に近づく

★ 中心極限定理の直感的な解釈：独立な確率変数をたくさん足すと正規分布に近い分布になる

★ それぞれの平均を m，分散を σ^2 の正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$$

★ X, Y をそれぞれ独立で平均を m_X, m_Y，分散を σ_X^2, σ_Y^2 の正規分布とすれば $X + Y$ は平均 $m_X + m_Y$，分散 $\sigma_X^2 + \sigma_Y^2$ の正規分布

χ^2 分布

★ $X_1, X_2, \ldots, X_n$ は独立でそれぞれ標準正規分布（平均0分散1）に従うとする

★ $Z = X_1^2 + X_2^2 + \cdots + X_n^2$ が従う分布を自由度 n の χ^2 分布という

t 分布

★ X は標準正規分布に従い，Z は自由度 n の χ^2 分布に従うとする．また，X と Z が独立とする

★ $T = \dfrac{X}{\sqrt{Z/n}}$ が従う分布を自由度 $n-1$ の t 分布という

- ★ $X_1, X_2, \ldots, X_n$ を独立で平均 m，分散 σ^2 の正規分布に従うとする
- ★ $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ とし，不偏分散 $U^2 = ((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)/(n-1)$ とする
- ★ $T = \dfrac{\overline{X} - m}{U/\sqrt{n}}$ が従う分布も自由度 $n-1$ の t 分布となる

★ t 分布の密度関数はガンマ関数を用いて書くことができ，拡張することで自由度が非整数の場合も考えることができる

★ 自由度 $n \to \infty$ の極限で t 分布は標準正規分布に近づく

F 分布

★ Z_1 を自由度 n_1 の，Z_2 を自由度 n_2 の χ^2 分布に従い，互いに独立とする

★ $F = \dfrac{Z_1/n_1}{Z_2/n_2}$ が従う分布を自由度 (n_1, n_2) の F 分布という

Z検定：平均がある値か（分散既知）

★ $X_1, X_2, \ldots, X_n$ の平均は m_0 に等しいか（小さくないか，大きくないか）を調べる検定

★ ただし，X_k の分散 σ^2 は既知とする（あまり現実的ではない）

★ X_k が正規分布であれば正確な検定で，n が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $Z = \dfrac{\overline{X} - m_0}{\sigma / \sqrt{n}}$ が標準正規分布に従うことから検定を行う

★ この工場で作られているお菓子は内容量の平均が50gで標準偏差は1gとのことだが，本当に平均が50gだろうか

Z検定：平均がある値か（二項分布版）

★ $X_1, X_2, \ldots, X_n$ の平均は m_0 に等しいか（小さくないか，大きくないか）を調べる検定

★ ただし，X_k は0または1（起こらなかったか，起こったか）

★ n が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $Z = \dfrac{\overline{X} - m_0}{\sqrt{m_0(1-m_0)/n}}$ が標準正規分布に従うことから検定を行う

★ このサイコロを振って6の目が出る確率は1/6に等しいだろうか

Z検定：2群の平均が等しいか（分散既知）

★ $X_1, X_2, \ldots, X_{n_1}$ の平均と $Y_1, Y_2, \ldots, Y_{n_2}$ の平均が等しいかを調べる検定

★ ただし，X_k の分散 σ_X^2，Y_k の分散 σ_Y^2 は既知とする（あまり現実的ではない）

★ X_k, Y_k が正規分布であれば正確な検定で，n_1, n_2 が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}}$ が標準正規分布に従うことから検定を行う

★ 工場Aで作られているお菓子と工場Bで作られているお菓子の内容量の平均に差はないだろうか

t 検定：平均がある値か（分散未知）

★ $X_1, X_2, \ldots, X_n$ の平均は m_0 に等しいか（小さくないか，大きくないか）を調べる検定

★ X_k が正規分布であれば正確な検定で，n が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ とする（不偏分散）

★ $T = \dfrac{\overline{X} - m_0}{\sqrt{S^2}/\sqrt{n}}$ が自由度 $n-1$ の t 分布に従うことから検定を行う

★ この工場で作られているお菓子は内容量の平均が50gとのことだが，本当に平均が50gだろうか

t 検定：2群の平均が等しいか(分散未知で等しい)

★ $X_1, X_2, \ldots, X_{n_1}$ の平均と $Y_1, Y_2, \ldots, Y_{n_2}$ の平均が等しいかを調べる検定

★ X_k の分散と Y_k の分散は等しいとする

★ X_k が正規分布であれば正確な検定で，n_1, n_2 が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $S_X^2 = \sum_{i=1}^{n_1}(X_i - \overline{X})^2$，$S_Y^2 = \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$，$S^2 = \dfrac{S_X^2 + S_Y^2}{n_1 + n_2 - 2}$

★ $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S^2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ が自由度 $n_1 + n_2 - 2$ の t 分布に従うことから検定を行う

★ 工場Aで作られているお菓子と工場Bで作られているお菓子の容量は同じだろうか

t 検定：2群の平均が等しいか（分散未知）

★ $X_1, X_2, \ldots, X_{n_1}$ の平均と $Y_1, Y_2, \ldots, Y_{n_2}$ の平均が等しいかを調べる検定

★ X_k, Y_k が正規分布，または，n_1, n_2 が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $S_X^2 = \dfrac{1}{n_1 - 1} \displaystyle\sum_{i=1}^{n_1} (X_i - \overline{X})^2$，$S_Y^2 = \dfrac{1}{n_2 - 1} \displaystyle\sum_{i=1}^{n_2} (Y_i - \overline{Y})^2$

★ $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}$ が自由度 v の t 分布に従うことから検定を行う

★ $v = \dfrac{\left(\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2} \right)^2}{\dfrac{S_X^4}{n_1^2(n_1 - 1)} + \dfrac{S_Y^4}{n_2^2(n_2 - 1)}}$

χ^2 検定：分散がある値か

★ $X_1, X_2, \ldots, X_n$ の分散は σ_0^2 に等しいか（小さくないか，大きくないか）を調べる検定

★ X_k が正規分布であれば正確な検定で，n が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $S = \dfrac{1}{\sigma_0^2} \displaystyle\sum_{i=1}^{n} (\overline{X} - X_i)^2$ が自由度 $n-1$ の χ^2 分布に従うことから検定を行う

★ 工場で3mmのネジを作っている．このネジの標準偏差が0.1mm以上でないことを確認したい

F 検定：2群の分散が等しいか

★ $X_1, X_2, \ldots, X_{n_1}$ の分散と $Y_1, Y_2, \ldots, Y_{n_2}$ の分散が等しいかを調べる検定

★ X_k, Y_k が正規分布であれば正確な検定で，n_1, n_2 が十分大きい（例えば30以上）であれば近似的に正しい検定

★ $S_X^2 = \dfrac{1}{n_1 - 1} \displaystyle\sum_{i=1}^{n_1} (X_i - \overline{X})^2$，$S_Y^2 = \dfrac{1}{n_2 - 1} \displaystyle\sum_{i=1}^{n_2} (Y_i - \overline{Y})^2$ とする

★ $F = S_X^2 / S_Y^2$ が自由度 (n_1, n_2) の F 分布に従うことを用いて検定する

★ 工場Aと工場Bでネジを作っている．精度に差があるだろうか

回帰分析の係数に関する t 検定

★ $Y = aX + b$ というモデルで最小二乗法で回帰分析した結果 $a = \hat{a}, b = \hat{b}$ となった

★ $\varepsilon_i = y_i - \hat{a}x_i - b$：データ i に対する残差 $(1 \leq i \leq n)$ とする

★ $a = a_0$ かどうかを検定したいとする

★ $T = \dfrac{(\hat{a} - a_0)\sqrt{n-2}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}{\displaystyle\sum_{i=1}^{n}\varepsilon_i^2}$ が自由度 $n-2$ の t 分布に従うことを用いて検定する

★ $a_0 = 0$ として検定すれば，X と Y に相関があることが確かめられる

U検定：2群が同分布かの検定（データ数小）

★ $X_1, X_2, \ldots, X_{n_1}$ と $Y_1, Y_2, \ldots, Y_{n_2}$ が与えられた時，X_k と Y_k が同じ分布かを検定したい

★ $X_1, X_2, \ldots, X_{n_1}$ と $Y_1, Y_2, \ldots, Y_{n_2}$ を昇順に並べ替えた時，X_k が何番目に来るかの和の値で持って検定を行う

U検定：2群が同分布かの検定（データ数小）

★例題：同じような体型の人が7人いる.

- ★ダイエット法Aを3人に試してもらうと1週間でそれぞれ1.2kg，1.1kg，0.9kg痩せた.
- ★ダイエット法Bを4人に試してもらうと1週間でそれぞれ1.0kg，0.8kg，0.2kg，0.7kg痩せた.

★ダイエット法Aの方が効果が高いといえるか.

★今，ダイエットの効果があった方から並べるとAABABBBである.

★Aの人が何番目かと言う和を考えると$1+2+4=7$である.

★ところで，ダイエット法の効果に差がないとすると，この和について

- ★6になるのはAAABBBBの1通りのみだから確率1/35
- ★7になるのはAABABBBの1通りのみだから確率1/35
- ★8になるのはABAABBB，AABBABBの2通りのみだから確率2/35

U検定：2群が同分布かの検定（データ数小）

★なので6になる確率と7になる確率の和でP値は2/35となる.

- ★危険率10%の場合はダイエット法Aの方が効果が高いといえる
- ★危険率5%の場合は何も言えない

★Rで計算するには`wilcox.test`を用いる（ウィルコクソンの順位和検定）

★Rでは，各Bについて，そのBの前にAがいくつあるかの和を考えている

χ^2検定：適合度に関する検定（ピアソン）

★ K個に分類された観測データがあり，カテゴリiに属すデータの件数はO_iである

★ 理論的にカテゴリiに属すデータの期待値はE_iである

★ このデータは理論的な期待値と隔たりがあるかどうかを検定する

★ 任意のiに対してE_iがそこそこ大きい場合（5以上）に適用可能

- ★ $X = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$が近似的に自由度$K-1$の$\chi^2$分布に従うことを利用して検定する
- ★ ただし，理論的な期待値を求めるために，パタメータを推定する場合は，その分自由度が小さくなる

★ このサイコロを振った時にでる目の割合は，全ての目で1/6だろうか

χ^2 検定：独立性に関する検定

★ 属性1は1からpのp種類，属性2は1からqのq種類ある

★ 属性1がiで属性2がjのデータがn_{ij}個ある．またデータの総量はN個

★ 属性1と属性2は独立かどうかを調べる検定

★ 任意のi, jに対してn_{ij}がそこそこ大きい場合（6以上）に適用可能

- ★ $n_{i*} = \sum_{j=1}^{q} n_{ij}$，$n_{*j} = \sum_{i=1}^{p} n_{ij}$とする
- ★ $X = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - n_{i*}n_{*j}/N)^2}{n_{i*}n_{*j}/N}$ が近似的に自由度$(p-1)(q-1)$のχ^2分布に従うことを利用して検定する

★ 学生の得意科目は所属高校に関係があるだろうか