

確率統計 (3)

概要

1. 統計学の基礎

- 1次元のデータ
- 2次元のデータ
- 相関

2. 確率の基礎

- 確率変数と確率分布
- 中心極限定理
- 独立性

3. 仮説検定

- t検定
- カイ二乗検定

4. 推定

- 点推定・区間推定
- 最尤推定

補助資料：<http://small-island.work/trial/>

ユーザ名：trial

パスワード：trial

仮説検定

確率やデータのばらつきを考慮した議論をし、最終的にある意思決定や断定を行いたい場合に用いる

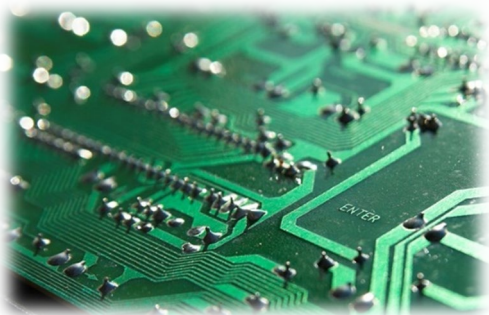
医療データなどの分析

- ・ 薬や治療の効果の有無



自然科学・社会科学

- ・ 発見・現象の検証



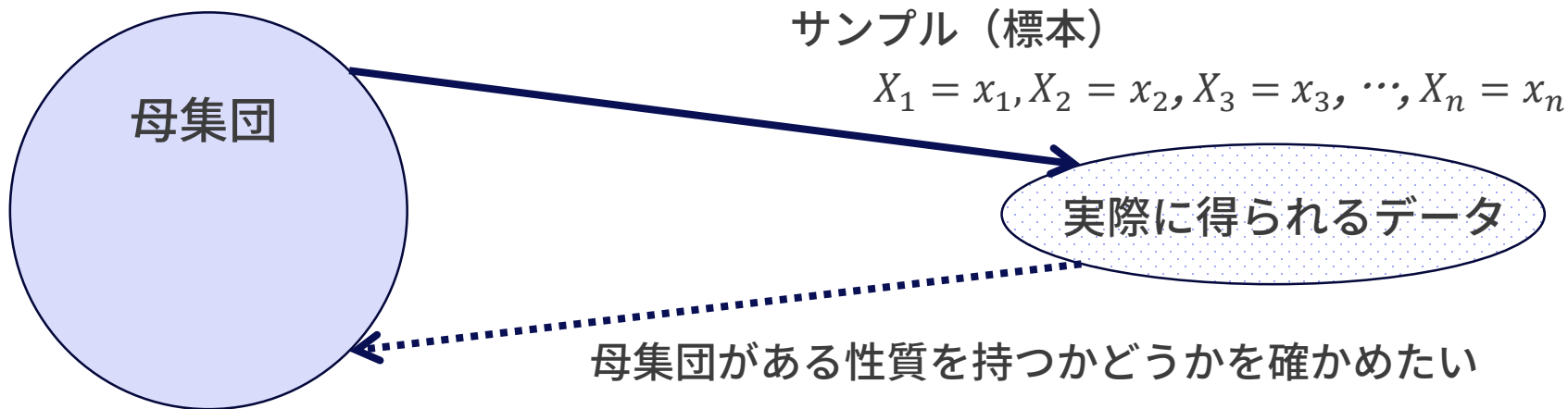
その他の製品開発など

- ・ 製品の保証や検証
- ・ 手法間の性能の比較

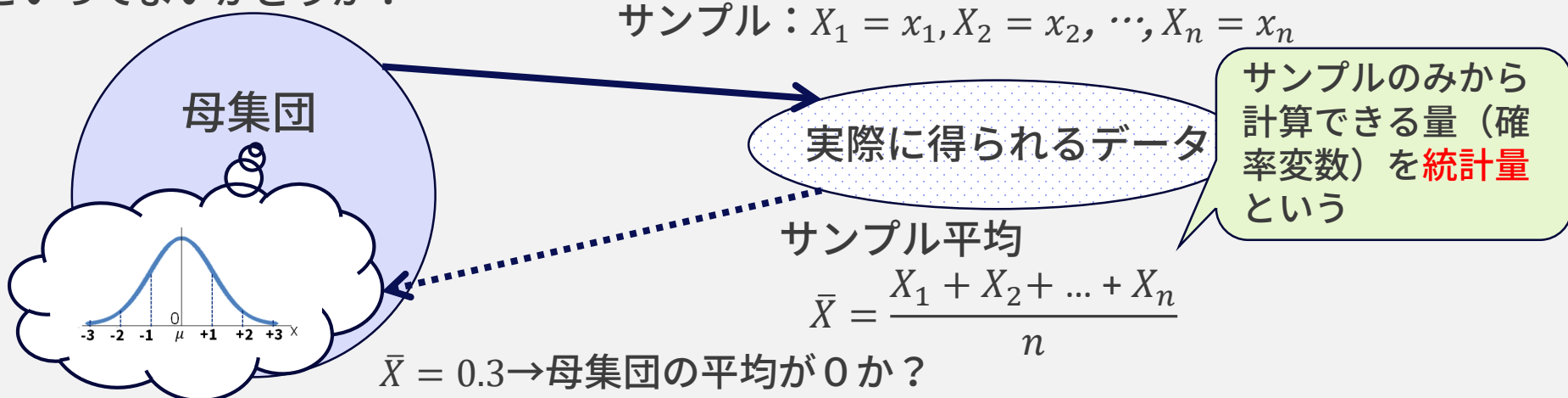


検定をおこなう上で重要な考え方

手元にあるデータは何らかの母集団からサンプルされたものであると考える



例：手元のサンプルの平均がある値（例えば0.3）だった時、母集団の分布が平均0ではないとってよいかどうか？



例：手元のサンプルの平均がある値（例えば0.3）だった時、母集団の分布が平均0ではないとってよいかどうか？

サンプル： $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$

母集団

実際に得られるデータ

- ・実際には分散が仮定できないので後に述べるt検定を利用する
- ・正規分布であることは中心極限定理などから仮定できる場合が多い

サンプル平均

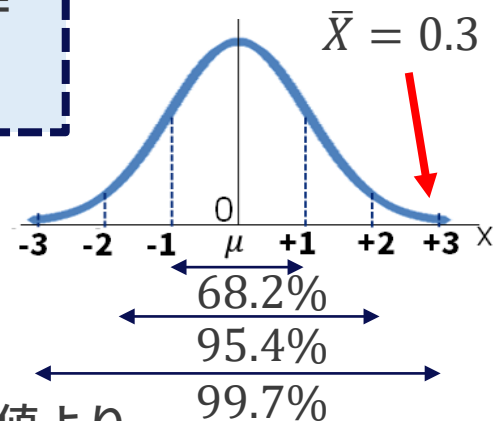
$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

ここでは、母集団は分散1の正規分布に従うことを前提とする

正規分布の基本的な性質

平均 μ 、分散 σ^2 の正規分布から n のサンプルを無作為に抽出する時、サンプル平均 \bar{X} は、 $N(\mu, \frac{\sigma^2}{n})$ に従う

仮に、平均0の分布から抽出されるとすると \bar{X} は、 $N(0, \frac{1}{n})$ に従うはずである。今、 $n = 100$ だったとすると $\sigma = 0.1$ である。今回の $\bar{X} = 0.3$ は 3σ に当たるため、平均0の分布から抽出されたとすると非常に低い確率でしか観測されないことになる。



統計的仮説検定では極端な値が観測される確率が予め決めた閾値より低い場合には仮定を棄却することで結論を出す。

統計的仮説検定の手順

1. 仮説を立てて、その仮説が正しいと仮定する：**帰無仮説**（Null hypothesis） H_0 と呼ぶ

2. 帰無仮説の下で「ある統計量 Y が極端な値で観測される確率を求める。また、観測されたサンプルから統計量の値 y を算出しておく。

つまり、 $P(Y \text{が極端な値} | H_0)$ の分布を計算する
（これが簡単な場合でも複雑な計算になっていることが多いためよく教科書では天下りの与えられる）

3. その確率が十分に（予め決めた閾値 p ：**有意水準**という）小さい範囲（**棄却域**という）に値 y が入っていれば、その仮説を棄却(reject)する

有意水準は慣習的に5%や3%を使うことが多い

4. 帰無仮説は成り立つかどうかを判断する

帰無仮説が棄却された際には、その否定である**対立仮説**（Alternative hypothesis）が採択される

棄却されなかった場合は、「ありえないとは言えない」という消極的な結果となる

統計的仮説検定法の選択

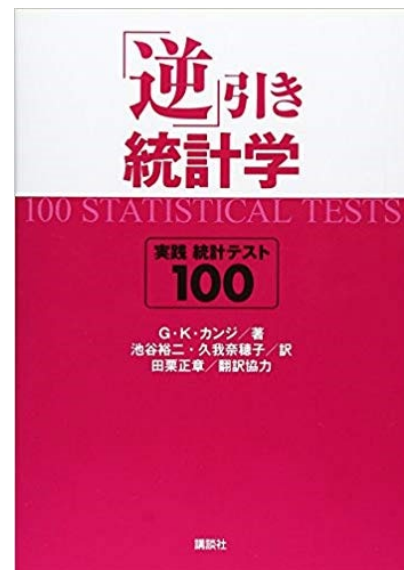
実践上の多くの場合は右のような辞書もしくはweb上に多くの参考資料があるのでそれらを使って検定法を選択するのがよい

特に考慮すべき点としては以下がある

- ・母集団については何がわかっているか？
基本的にわかっていることは使った方がよい結果となる
- ・どのような性質を確かめたいか？
必要以上に多くの知識を得ようとすれば通常多くのデータや仮定を必要とする

しかしこういったノウハウだけでなく、手法を使う上でのよくある注意点（多重検定や仮説の考え方など）や実問題での勘所などをつかむためには、ここまでの基本的な考え方を押さえておく必要がある。

このスライド以降では特に基本的な検定法を紹介して検定に慣れることを目的とする



例えば、仮説検定では「何度もデータを取り直して実行すると間違えた結論となることがある」という注意点があるが、前スライドで述べた通り、検定は希少な事象は起こりにくいことに基づいているため、こう言った事が起こるのは明らかである

t検定

定義
・
性質

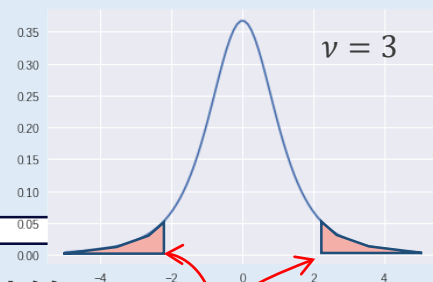
- 帰無仮説が正しいと仮定した場合に、統計量がt分布に従うことを利用する統計学的検定法
- 母集団が正規分布に従うと仮定するパラメトリック検定法

利用
する
分布

数式

t分布
$$p(x; \nu) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} (1 + t^2/\nu)^{-(\nu+1)/2}$$

自由度： ν



特徴

- 検定では0から遠い極端な値が観測される確率を計算する
- 一致するかどうかの検定：両側検定
 - 値の上下が分かっている差があるかどうかだけを検定する場合は片方の領域のみで確率を計算する：片側検定

例

- 母集団の平均値 μ が特定の値である μ_0 であるという帰無仮説
(例) ある集団の平均身長が日本人の平均身長と一致するかを調べたい
- 2組の標本について平均に有意差があるかどうかの検定
(例) 手法Aと手法Bの平均性能に差があるかどうかを調べたい
- 回帰分析の係数に関する検定
(例) 回帰を行った際に各因子が予測結果に寄与しているかを調べたい

t検定の例(1/3)

例:母集団の平均値 μ が特定の値である μ_0 であるという帰無仮説

1. 母集団の平均を μ_0 と仮定する

2. 仮定から、サンプル平均 \bar{X} が抽出される確率を求める

標本数 n 、標本標準偏差 S として

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

と変数変換により、 T を定義すると、

T は自由度 $n - 1$ のt分布に従う (studentの定理)

実際の観測から T を計算した値を t とする

3. その確率が予め決めておいた閾値より小さければ、仮説を棄却する

T がt分布に従うので、t分布での棄却域を設定し、その棄却域に t が入って入れば棄却する

4. 平均値 μ は μ_0 とは異なるかどうかを判断する

t検定の例(2/3)

例:XとYの標本について平均が等しいという帰無仮説

1. XとYの標本が、平均の等しい母集団から抽出されたと仮定する
2. 仮定から、XとYの標本平均 \bar{X} , \bar{Y} が抽出される確率を求める
標本数 n_X, n_Y 、標本分散 v_X, v_Y として

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{v(1/n_X + 1/n_Y)}} \quad v = \frac{(n_X - 1)v_X + (n_Y - 1)v_Y}{n_X + n_Y - 2}$$

と定義すると、

T は自由度 $n_X + n_Y - 2$ のt分布に従う

実際の観測から T を計算した値を t とする

3. T がt分布に従うので、t分布での棄却域に t が入って入れば棄却する
4. 平均値 \bar{x} , \bar{y} とは異なるかどうかを判断する

t検定の例(3/3)：やや発展的

例: XとYの標本 (x, y) の間に $y = ax + b + e$ の関係があるときに、 a が $a = 0$ かという帰無仮説 (e は平均0である分散を持つ正規分布)

($a = 0$ ならばXとYは無相関)

1. XとYの標本から最小二乗法で求めた a, b を \hat{a}, \hat{b} として、 $a = 0$ の母集団から抽出されたと仮定する

2. 仮定から、実際に観測されたXとYの標本平均 \bar{X}, \bar{Y} が抽出される確率を求める

標本数 n 、最小二乗法の誤差の総和を SSE とする。

$$T = \frac{\hat{a}\sqrt{n-2}}{\sqrt{SSE / \sum_{i=1}^n (X_i - \bar{X})^2}}$$

と T を定義すると、 T は自由度 $n - 2$ の t 分布に従う

実際の観測から T を計算した値を t とする

3. T が t 分布に従うので、 t 分布での棄却域に t が入って入れば棄却する

4. 平均値 \bar{x}, \bar{y} とは異なるかどうかを判断する

ピアソンのカイ二乗検定(χ^2 検定)

定義・性質

- 検定統計量が漸近的に χ^2 分布に従うような統計学的検定法
- 「観察された事象の相対的頻度がある頻度分布に従う」という帰無仮説を検定する
- 頻度分布を対象にしており、特定の分布形を仮定しないノンパラメトリック検定

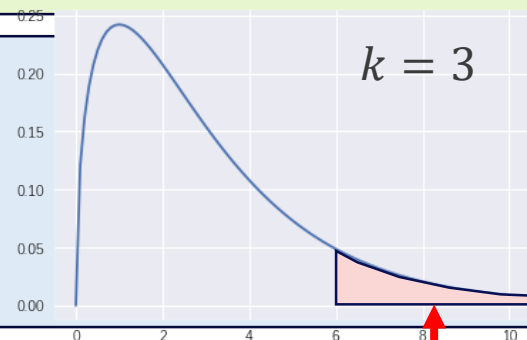
利用する分布

数式

◆ 210pm分布

$$p(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

自由度: k



特徴

独立に標準正規分布に従う k 個の確率変数 X_1, \dots, X_k があるとき

$$Z = \sum_{i=1}^k X_i^2$$

と定義される Z は自由度 k の ◆ 210pm分布に従う

検定では0から遠い極端な値が観測される確率を計算する (片側の場合)

例

- 適合度検定: 観測された度数分布が理論分布と同じかどうかを検定する
(例) サイコロの目が一様かの検定
- 独立性検定: 二つの変数が独立かどうかの検定をする
(例) ある因子とある因子に関連があるかを調べる

適合度検定の例

例:離散の値1~ k をとる確率変数 X の標本がある分布に従うかの検定

1. $X = i$ の確率が p_i であると仮定する

2. 仮定から期待される i の観測回数 E_i を使って、実際に観測された X の標本に関して i が観測された回数を O_i の確率を計算する

標本数 n とすると

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad E_i = np_i$$

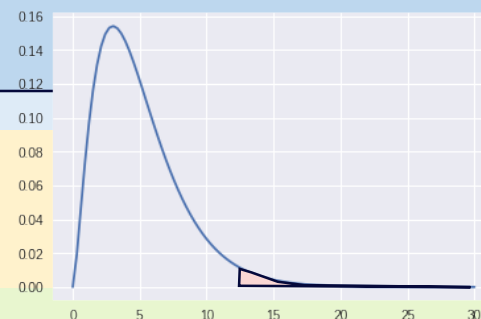
この時、 χ^2 は自由度 $k - 1$ の
b ◆ 21Bpm分布に従う

	1	2	3	4	5	6
観測 O_i	30	25	20	20	15	10
期待 E_i	20	20	20	20	20	20

$$n = 120, p_i = \frac{1}{6}, k = 6$$

$$\chi^2 = 100/20 + 25/20 + 25/20 + 100/20 = 12.5$$

$$p = 0.0285$$



3. χ^2 が b ◆ 21Bpm分布に従うので、 χ^2 を観測から計算した値がカイ二乗分布での棄却域に入っていれば棄却する

4. $X = i$ の確率が p_i の確率分布に従うかどうかを判断する

独立性検定の例(1/2)

例:ある二つの離散の確率変数 X, Y が独立かどうかの検定

1. 確率変数 X, Y (取りうる値の数を n, m とする) が独立であると仮定する

2. 仮定から期待される観測回数と実際に観測された回数の表を作成する

実際に観測された回数の表

$X \setminus Y$	陽性	陰性	
男性	8	42	50
女性	7	93	100
	15	135	

• 陽性・陰性の割合

• 男女の比率

を計算し、独立とした時の確率の表を作成する
(独立なので積)

仮定から期待される観測回数

$X \setminus Y$	陽性	陰性	
男性	5	45	50
女性	10	90	100
	15	135	

$X \setminus Y$	陽性	陰性	
男性	$1/3 \times 1/10$	$1/3 \times 9/10$	$1/3$
女性	$2/3 \times 1/10$	$2/3 \times 9/10$	$2/3$
	$1/10$	$9/10$	

独立性検定の例(2/2)

例:ある二つの離散の確率変数 X, Y が独立かどうかの検定

仮定から期待される観測回数と実際に観測された回数の表から

実際に観測された回数の表を O_i となる確率を計算する

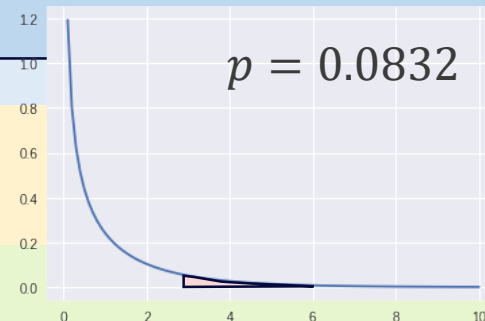
各対応するマスごとに以下を計算し、その総和を χ^2 とすると

$$\frac{(O_i - E_i)^2}{E_i}$$

χ^2 は自由度 $(n - 1)(m - 1)$ のカイ二乗分布に従う

$X \setminus Y$	陽性	陰性	
男性	$(8 - 5)^2 / 5$	$(42 - 45)^2 / 45$	50
女性	$(7 - 10)^2 / 10$	$(93 - 90)^2 / 90$	100
	15	135	

$$\begin{aligned} & \frac{(8-5)^2}{5} + \frac{(42-45)^2}{45} \\ & + \frac{(7-10)^2}{10} + \frac{(93-90)^2}{90} \\ & = 3 \end{aligned}$$



3. χ^2 がカイ二乗分布に従うので、 χ^2 を観測から計算した値が χ^2 分布での棄却域 χ^2 が入って入れば棄却する

4. 二つの確率変数が独立かどうかを判断する

web上の資料の紹介

計算省略部分を補完できる外部資料：

- studentの定理
 - <https://qiita.com/s0sasaki/items/a27e743251218777e343>
- 回帰のt検定
 - http://www2.econ.osaka-u.ac.jp/~tanizaki/class/2018/basic_econome/02.pdf
- 回帰のt検定（その他の検定の資料も含む）
 - <https://www.math.kyoto-u.ac.jp/~ichiro/lectures/2012st.pdf>
- 分割表のカイ二乗検定
 - <http://joe.bayesnet.org/wp-content/uploads/2017/05/2017-5-25-1.pdf>