データサイエンスのための統計学

二大定理と標本分布論

by 尚晋　(名古屋大学経済学研究科助教)

on 2025 年 6 月 17 日

目次

* 二大定理
* 標本分布
* 正規母集団からの標本分布論
* 2 標本問題

目次

* 二大定理
* 標本分布
* 正規母集団からの標本分布論
* 2 標本問題

» 二大定理　　大数の法則

確率収束 下記の式を満たすなら、$\{X_n\}$ は c に確率収束するという.

$$\lim_{n\to\infty} P(|X_n - c| \leq \varepsilon) = 1, \quad (\forall \varepsilon > 0) \tag{1}$$

或いは $plim_{n\to\infty} X_n = c$ または $X_n \xrightarrow{p} c$ と表す.

大数の法則

無作為標本のサンプルサイズ n が十分大きいならば, 標本平均 $\bar{X}_n = (X_1, X_2, .., X_n)/n$ の確率分布は母集団確率分布の平均 (母平均)μ の近くに集中している. 式は

$$plim_{n\to\infty} \bar{X}_n = \mu \tag{2}$$

或いは、$\bar{X}_n \xrightarrow{p} \mu$ と表し, $\bar{X}_n$ が μ に確率収束するという.

チェビシェフの不等式 $(P(|X - \mu| \geq k\sigma) \leq 1/k^2, \quad (k > 0)$ が成り立つ) より

$$P(|\bar{X}_n - E(\bar{X}_n)| > \varepsilon) \leq V(\bar{X}_n)/\varepsilon^2, \quad (\forall \varepsilon > 0)$$

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} \to 0 \quad (n \to \infty)$$

» 二大定理 中心極限定理

中心極限定理

無作為標本のサンプルサイズ n が十分大きいならば, 標本平均 $\bar{X}_n = (X_1, X_2, .., X_n)/n$ を標準化したもの ($Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$) の確率分布は標準正規分布 $N(0,1)$ に近づいていく. 式は

$$Z_n \xrightarrow{d} N(0,1), \quad (n \to \infty) \tag{3}$$

Z_n が $N(0,1)$ に分布収束するという.

漸近分布 確率変数の分布が分布収束する場合, その収束先を漸近分布という.

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1), \quad (n \to \infty) \tag{4}$$

$\overset{a}{\sim}$ は漸近分布であることを表し, 標準化した標本平均の漸近分布は標準正規分布という.

目次

» 統計的推測:母集団・母数と標本

統計的推測 全体(母集団)から一部(標本)を抽出・分析し、全体の特徴(母数)を推測する手法を、統計的推測と呼ぶ。統計的推測の構成要素:母集団・母数と標本

母集団 分析者が興味のある対象「全体」を母集団。有限個の個体から成る母集団を有限母集団という. 無限個の個体から成る母集団を無限母集団という. 母集団の例:「サラリーマン全体」の年収平均。

■ しかし、母集団全体を把握するのは、ほぼ不可能!例:日本中のサラリーマン全員を調査するのはムリ。
しかし「一部」を観測・分析し、「全体」の特徴を推測するのは?

標本 母数の推測のために母集団から抽出(サンプリング)したデータ, $X_1, X_2, \ldots, X_n$ を標本と呼び, 代表して X_i と表記(n は標本の大きさ、sample size)。例:サラリーマン $n = 1000$ 人に年収をアンケート調査で、標本平均を求める。

» 統計的推測:母集団·母数と標本

標本抽出
(a) 母集団から標本を取り出すことを標本抽出という.
(b) 取り出した個体を母集団に戻しながら繰り返す抽出を復元抽出という.
(c) 取り出した個体を母集団に戻さずに繰り返す抽出を非復元抽出という.
(d) 母集団の各要素が標本に含まれる確率 (抽出率) が等しく、即ち、等確率で取り出される抽出を(単純) 無作為抽出という. 無作為抽出した標本を無作為標本という.
(e) 復元抽出した無作為標本の各個体を確率変数で表すと, それらは独立かつ同一に (independent and identically distributed, iid) 母集団分布にしたがう. この教科書は復元抽出した無作為標本を想定する.

母集団分布と母数
(a) 母集団における各個体の属性値の分布をを母集団分布という。例: 日本人 (母集団) の身長 (属性)
(b) 母集団分布の特性を表す定数を母数 (パラメーター) という.
(c) 母集団分布の平均, 分散を母平均, 母分散という.
(d) 母集団分布がある知られた確率分布であることが, 理論的·経験的にわかっている場合で, 有限個の母数で表せる分布をパラメトリックな分布という.

» 統計量と標本分布

統計量 一般に、母集団の母数の推測に使われる、標本から要約したものを統計量という。標本平均と標本分散は最も重要な二つの統計量である。

標本分布 確率的な標本抽出にともなう統計量の確率分布をその統計量の標本分布という.

標本平均 標本 $X_1, X_2, \ldots, X_n$ から計算された平均を標本平均という。標本 $X_1, X_2, \ldots, X_n$ は母集団 (母平均 μ、母分散 σ^2) に従う独立な確率変数であるが、標本平均 $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ の期待値と分散は

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \sigma^2/n \tag{5}$$

となる. 期待値が母平均 μ に一致する. この性質は不偏性という。$n \to \infty$ の時 $\bar{X}$ の分散は 0 に近づいていき、標本平均 $\bar{X}$ が母平均 μ に集中、確率収束していくのである. この性質は一致性という。

» 統計量と標本分布 (続き)

標本分散 標本 $X_1, X_2, \ldots, X_n$ から計算された分散を標本分散という。

(a) 母平均が既知の場合, 標本 $X_1, X_2, \ldots, X_n$ の標本分散は

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{6}$$

で定義される. $E(\hat{\sigma}^2) = \sigma^2$ の性質を満たす.

(b) 母平均が未知の場合, 標本 $X_1, X_2, \ldots, X_n$ の標本分散は

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{7}$$

で定義される. $E(s^2) = \sigma^2$ の性質を満たす. したがって, 上の式の s^2 を母分散 σ^2 の不偏推定量, あるいは不偏分散 unbiased variance という.

' 不偏でない' 標本分散 ($S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$) は過小評価が起こるので、$S^2$ と s^2 の違いには、要注意。

目次

* 二大定理
* 標本分布
* 正規母集団からの標本分布論
* 2 標本問題

» 正規標本論の導入

* 正規分布は統計学で最も基本的・扱いやすい分布.
* 多くの統計的手法は「正規母集団」を前提とする.

正規標本論

正規母集団から得られる標本 $X_1, X_2, \ldots, X_n$ に基づく統計量(例:平均, 分散)の標本分布を扱う理論.

測定誤差の例:鉛筆の長さを n 回測定 (広い意味では「観測」).

$$X_i = \mu + \varepsilon_i \tag{8}$$

* μ:真の長さ(定数, 未知)
* ε_i:誤差(確率変数, 独立・同一分布 (iid))

» ガウスの誤差理論と正規分布

ガウスの誤差理論より、誤差の仮定

$$\varepsilon_i \sim N(0, \sigma^2) \tag{9}$$

* 誤差は平均 0, 分散 σ^2 の正規分布に従う
* 精度の良い測定であれば σ^2 は小さい

測定値の分布：

$$X_i \sim N(\mu, \sigma^2) \tag{10}$$

* 測定値も正規分布に従う（位置は μ だけ平行移動）
* 全ての X_i は独立で同一分布 (iid)

誤差理論の意義： 測定値 = 真の値 + 誤差というモデルが，正規標本論の出発点．⟶ 推定・検定など，統計的推論の土台を形成．

» 母分散が既知場合の標本平均 の標本分布

* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$：独立・同分布な測定値
* 標本平均 $\bar{X} = \frac{1}{n} \sum X_i$ は次の分布に従う：

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{11}$$

* 標準化すれば：

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{12}$$

意義

σ^2 が既知であれば，$\bar{X}$ の分布を標準正規分布で扱える．

* $\bar{X}$ の標準偏差は $\sigma/\sqrt{n}$ で、n が大きくなるほど，$\bar{X}$ は μ のより正確な推定値となる：
  * 正確さを 2 倍にするには n を 4 倍にする必要がある
* 標本平均 $\bar{X}$ は，分散の小さい，単独の X_i よりも優れた測定値．

» 母分散が既知場合の標本平均の標本分布 例

鉛筆の例で、下記の測定データによって、真の長さが $\mu = 18.0$ という仮定は維持できるか：

* 母平均 $\mu = 18.0$
* 母標準偏差 $\sigma = 0.02$
* $n = 10$ 回の測定, 平均 $\bar{X} = 18.06$

標準化して Z 統計量：

$$Z = \frac{18.06 - 18.0}{0.02/\sqrt{10}} = 9.49$$

解釈：

* 標準正規分布で $Z = 9.49$ は極めて稀（付表より $z_{0.00005} \approx 3.89$）(1 - stats.norm.cdf(9.49)) で値を確認.
* よって, $\mu = 18.0$ という仮定は棄却されるべき.

» 標本分散の標本分布　カイ二乗分布

χ^2 分布：

$Z_1, Z_2, \ldots, Z_k$ を独立な標準正規分布 $N(0,1)$ に従う確率変数とする．いま，

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \tag{13}$$

とすると，確率変数 χ^2 が従う確率分布を，自由度 k の χ^2 分布という．この自由度は，独立な標準正規確率変数の二乗をいくつ加えたかを表す．また，自由度 k のカイ二乗分布を $\chi^2(k)$ で表す．

パーセント点：

* $\chi^2_\alpha(k)$：上側確率 α の点
* 例：$\chi^2_{0.05}(5) = 11.070 \Rightarrow P(\chi^2 > 11.070) = 0.05$

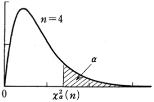
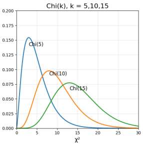


図 10.4　自由度 k の χ^2 分布 $\chi^2(k)$ のパーセント点

χ^2 分布は，密度関数 $f(x)$ が $x>0$ の部分でのみ正の値をとるもので，自由度 k が大きくなるに従い $f(x)$ は右方向へ移動する．$\chi_\alpha{}^2(k)$ は $\mathrm{P}(\chi^2 > \chi_\alpha{}^2(k)) = \alpha$ で定義される．なお，$\chi^2(k)$ はガンマ分布 $Ga(k/2, 1/2)$ のことである．



» 標本分散の標本分布 カイ二乗統計量

カイ二乗統計量

標本分散（不偏分散）は：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ とするとき、カイ二乗統計量は：

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{14}$$

意味：

* 正規母集団から得た標本分散 s^2 は χ^2 分布に従う
* 母分散 σ^2 に関する推測（検定・信頼区間）に用いられる

例題：

* 母分散 $\sigma^2 = 25$, 標本サイズ $n = 10$, $s^2 > 50$ の確率は？

$$P(s^2 > 50) = P\left(\chi^2 > \frac{9 \cdot 50}{25}\right) = P(\chi^2 > 18)$$

$\Rightarrow P = 0.038$ （比較的稀な出来事）

» σ^2 が未知の場合の標本平均の標本分布　t 統計量

* 前節の母分散 σ^2 が既知であるかては現実ではない.
* 現実的には, 母分散の代わりに標本分散 s^2 を用いて t 統計量を使う:

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

* この統計量は標準正規分布には従わず, t 分布に従う.

t 統計量の確率分布の導出の概要:

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

の形に変形できる. 分子 $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$ は標準正規分布 $\mathcal{N}(0,1)$ に, 分母の $\sqrt{(n-1)s^2/\sigma^2}$ 部分は自由度 $n-1$ の χ^2 分布 $\chi^2(n-1)$ に従う. これらの比になっている.

» σ^2 が未知の場合の標本平均の標本分布 t 分布

t 分布 (或いは, スチューデントの t 分布):

* $Z \sim \mathcal{N}(0,1)$（標準正規分布）
* $Y \sim \chi^2(k)$（自由度 k のカイ二乗分布）
* Z と Y は独立

このとき,

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

は自由度 k の t 分布 $t(k)$ に従う.

t 統計量:

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

で定義した t 統計量は自由度 $n-1$ の $t(n-1)$ 分布に従う. $\bar{X}$ の標準偏差 $s/\sqrt{n}$ を, 標本平均の標準誤差(standard error) という.

» σ^2 が未知の場合の標本平均の標本分布 t 分布性質

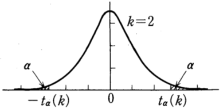
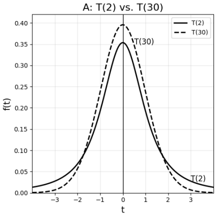


図 10.5　自由度 k の t 分布 $t(k)$ とパーセント点

母平均の仮説検定，推定論などにおいて，実際に用いられるのは，標準正規分布より，この t 分布である．二つは酷似するが，大標本($n\to\infty$)のときは，一致する．$t_\alpha(k)$ は $\mathrm{P}(t>t_\alpha(k))=\alpha$ で定義される．なお，両側(−の側)も入れて α とする流儀もある．



* t 分布は $t=0$ に関して左右対称（偶関数）
* 標準正規分布に似ているが，裾が広い（自由度が小さいほど）
* 自由度 1 の t 分布はコーシー分布
* $k\to\infty$ で $t(k)\to\mathcal{N}(0,1)$
* 大標本 $n\to\infty$ で $s^2\approx\sigma^2$ より，正規分布に近づく

パーセント点：$t_\alpha(k)$ は自由度 k の $100\alpha\%$ 上側確率点

» σ^2 が未知の場合の標本平均の標本分布 例

母平均 $\mu = 25.0$ で、下記の測定データによって、t 統計量を求めよう

* 標本サイズ $n = 7$, 標本平均 $\bar{X} = 25.21$
* 標本分散 $s = 0.715$

t 統計量：

$$t = \frac{25.21 - 25.0}{0.715/\sqrt{7}} = 0.777$$

解釈：

* t 分布表で自由度 6 で、$t = 0.777$ に対応する上側確率は 0.20 から 0.25 の間にある、(scipy.stats.t.sf(0.777,6)=0.23)
* 即ち、片側に 0.777 以上の t 値が出る確率が 23%。

目次

* 二大定理
* 標本分布
* 正規母集団からの標本分布論
* 2 標本問題

» 2 標本問題 導入問題

* 身長の分布を考えた場合, 一定地域で同一学年の男子の身長と女子の身長ではその母集団分布に図 10.6 のように大きな差があると考えられる.
* このような場合, 一つの正規母集団から標本を抽出したと仮定して分析することは適当でない. 男子の身長と女子の身長など明らかに異なる.

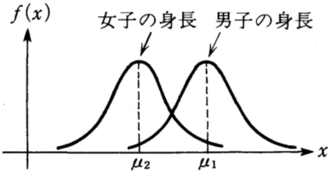


図 10. 6　2 母集団分布の比較

統計的推定, 検定ではもっともよく扱われる問題の一つである.

2 標本問題　2 種の標本による 2 母集団の比較を扱う問題を2 標本問題という.
2 標本問題では, 二つの母集団から別々に標本を抽出したと考える.
ここでは, 大きさ m の第一の標本 $X_1, X_2, \ldots, X_m$ を母集団分布 $N(\mu_1, \sigma_1^2)$ から, 大きさ n の第二の標本 $Y_1, Y_2, \ldots, Y_n$ を母集団分布 $N(\mu_2, \sigma_2^2)$ から独立に抽出した場合の標本分布について説明する.

» 標本平均の差の標本分布 設定

* 2 標本問題では, 二つの母平均 μ_1, μ_2 の差 $\mu_1 - \mu_2$ を分析する問題. 例: 男女の平均賃金の差 $\mu_1 - \mu_2$、男女の格差があれば 0 とならない.
* $\mu_1 - \mu_2$ を分析するには, 対応する標本平均

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

の差 $\bar{X} - \bar{Y}$ を調べればよい. $\bar{X} - \bar{Y}$ 即ち、$\bar{X}, \bar{Y}$ のばらつきを考慮

実際、これらの標本分布はもちろん双方の母分散にも依存するので、ここでは下記の 3 通りに分けて考える.

(a) 母分散 σ_1^2, σ_2^2 が既知のとき
(b) 母分散は未知であるが等しいとき
(c) 母分散が未知であり等しいとは限らないとき

» 標本平均の差の標本分布　(a) 母分散既知の場合

* 正規母集団と仮定しているので、各標本平均の分布：

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$$

* 差の分布：

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

* 標準化：

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0, 1)$$

例：男子の身長 $\mu_1 = 172.3, \sigma_1^2 = 30$, 女子の身長 $\mu_2 = 160.2, \sigma_2^2 = 25$, 男子 10 人, 女子 15 人を抽出したとき, 標本平均の差の標本分布は：
$\bar{X} - \bar{Y} = 12.1, \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = \frac{30}{10} + \frac{25}{15} = 4.667$ より：

$$\bar{X} - \bar{Y} \sim N(12.1, 4.667)$$

» 標本平均の差の標本分布 (b) 母分散未知・等しい

* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ の時、標準化した Z の分布：

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{m} + \frac{1}{n})\sigma^2}} \sim N(0, 1)$$

* しかし、σ^2 は未知なので、代わりに合併分散（pooled variance）を使う：

$$s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}$$

* 2 標本 t 統計量：

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

* この統計量は，次々章で述べる 2 標本 t 検定 に主に用いられる．

» 標本平均の差の標本分布(c) 母分散未知・等しくない

- 母分散が未知であり等しいとは限らないとき, 二つの母散を先にそれぞれ推定する必要があり, 正確な分布は求めにくいが, 近似的に分布を求めるウェルチ (Welch) の近似法 が用いられる.
- Welch の t 統計量:

$$
t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}
$$

- 自由度の近似 (Welch-Satterthwaite):

$$
\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}}
$$

- 最も近い整数 ν^* に従う $t(\nu^*)$ 分布とみなす

» 標本分散の比と *F* 分布

問題の背景

* 標本平均の差の分布を求めるには, 母分散の等しさが重要.
* 標本分散の比 s_1^2/s_2^2 に注目: ≈ 1 なら母分散も等しいと推測可.
* その分布は, F 分布を用いて分析できる.

F 分布:

$U \sim \chi^2(k_1)$, $V \sim \chi^2(k_2)$ が独立のとき, U と V をそれぞれの自由度で割って調整した後にとった比, 即ちフィッシャーの分散比を

$$F = \frac{U/k_1}{V/k_2} \sim F(k_1, k_2)$$

と定義, F が自由度 (k_1, k_2) の F 分布 $F(k_1, k_2)$ に従う.

» 標本分散の比と F 分布 標本分散と F 統計量

標本分散 s_1^2, s_2^2 について:

* $(m-1)s_1^2/\sigma_1^2 \sim \chi^2(m-1)$,
* $(n-1)s_2^2/\sigma_2^2 \sim \chi^2(n-1)$,
* s_1^2 と s_2^2 は独立である.

F 統計量:

したがって, F 分布の定義から F 統計量:

$$F = \frac{\frac{(m-1)s_1^2}{\sigma_1^2}/(m-1)}{\frac{(n-1)s_2^2}{\sigma_2^2}/(n-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

で定義され, F 統計量は自由度 $(m-1, n-1)$ の F 分布 $F(m-1, n-1)$ に従う.

* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ のときの F 分布の形 (よく使う):

$$F = s_1^2/s_2^2 \sim F(m-1, n-1)$$

» 標本散の比と F 分布 　　　F 分布性質

* 自由度 (k_1, k_2) の F 分布 $F(k_1, k_2)$ において, 上側確率が α となる値を上側確率 $100\alpha\%$ のパーセント点といい, $F_\alpha(k_1, k_2)$ と記す (付表 4).
* 性質:

$$F \sim F(k_1, k_2) \Rightarrow \frac{1}{F} \sim F(k_2, k_1)$$

$$F_{1-\alpha}(k_1, k_2) = \frac{1}{F_\alpha(k_2, k_1)}$$

* 例: $F \sim F(3, 5)$ のとき, $\frac{1}{F} \sim F(5, 3)$ に従い, $F_{0.05}(5, 3) = 9.013$ なら,

$$F_{0.95}(3, 5) = 0.110 \quad (\text{逆数関係})$$

» 標本分散の比と F 分布 応用例

例: 母分散が同一の正規母集団から, $m = 10, n = 15$ の標本を抽出したとする. このとき, $F = s_1^2/s_2^2$ の形になる. 二つの標本分散について, 母分散が等しいという仮定のもとで, s_1^2 が s_2^2 の 2 倍以上となる確率を求めよう.

* $P(F > 2) \approx 0.1183$
* (「scipy.stats.f.sf(Fvalue, df1, df2) で計算できる」)

解釈:

* 標本分散が 2 倍違っていても, 母分散が等しい可能性が約 11% ではあるが, 言いうる.