

標本抽出と中心極限定理、信頼区間の推定：資料

芳賀昭弘*

1 標本（サンプル）抽出

母集団（大元の集団／調べたい集団）を全てを調べるのが困難であるとき、その一部だけを調べて元の母集団の性質が推定できると便利ことがあります。例えば製品の破壊強度の検査を実施する際、全ての製品で破壊強度検査を行ってしまえば、売るための製品がなくなってしまいます！そのため、製品を一部だけ取り出して、破壊強度検査を行った結果から製品全体の破壊強度を推論します。統計解析では、母集団から一部を取り出し、その一部（標本）を観察することで元の母集団の性質を推論する、ということを行うことがあり、これを推論統計学と呼んでいます。母集団から標本（サンプル）を取り出すことを標本抽出と言います。^{*1}

例を幾つか眺めることで、母集団と標本抽出、統計解析の考え方にもう少し触れてみたいと思います。例えば一ヶ月あたりの徳島県民のすだちの消費量の平均値を知っていたとします。この場合、母集団は徳島県民（の各人の一ヶ月あたりのすだちの消費量）です。さて、出身地が徳島県外の人でも徳島県に移ってしまったら同じ消費量になるのか、それともならないのか、という疑問を持った人が現れ、出身地が徳島県外の何人かの一ヶ月あたりのすだちの消費量を調べてみたところ、徳島県民全体の消費量より少し少ない、という結果が得られました。この何人かの徳島県外出身者が標本です。さて、出身地が徳島県外の人でもすだちの消費量は本当に少ないのか、それとも少ししか違いがないので差がないのか？推論するにはどうすればよいのでしょうか？（徳島県出身者のすだちの消費量も別途調べた場合には、解析方法が変わります！その場合はどのような解析が必要でしょうか？）… 標本が母集団の性質に一致しているか（適合性検定）、2つの母

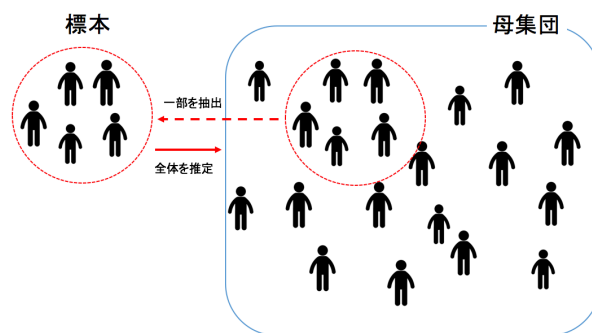


図1 標本抽出と母集団の推定

* Electronic address: haga@tokushima-u.ac.jp

^{*1} 母集団や標本を調べるということはどういうことでしょうか？何をどのように調べるのでしょうか？調査方法は貴方が何を知りた
いかに拠って様々な種類があります。貴方が知りたいことについて、問題として適切に設定されて初めて調べること・調べる方法
が決まります。問題設定は、貴方自身が行わなければなりません。それが他の学問と大きく異なっており、そうした経験がま
だあまり多くない学び始めの学生にとって、統計学が難しく感じる要因になっているのかもしれない。

集団同士の性質が一致しているか（独立性検定）ということを調べるときなどに統計学は力を発揮します。上のような問題に対して少しでも定量的な回答を用意したい場合には統計学の知識が必須です。そして、このような統計解析の方法をマスターしておくことで研究を行う上で大きなアドバンテージになります。また統計学は、卒業研究のみならず就職後の様々な業務においても（もしかしたら遊びや趣味のおいても）必ず役に立ちますので、今はわからないことだらけかもしれませんが頑張ってマスターしましょう。

2 大数の法則と中心極限定理

標本抽出で母集団の性質を知る、もしくは標本同士を調べることが有効であることを教えてくれる非常に重要な定理があります。大数の法則と中心極限定理です；

1. 標本抽出を繰り返したとき、“サンプルの平均値”の平均は母集団の平均に収束する。
2. 標本抽出を繰り返したとき、“サンプルの平均値”の標準偏差（標本誤差とも言われる）は‘母集団の標準偏差’/ \sqrt{n} に収束する。ここで n は抽出するときのサンプルの数（サンプルサイズという）である。
3. 母集団の分布が正規分布なら、標本抽出を繰り返したときの“サンプルの平均値”の分布も正規分布になる。サンプルサイズが十分に大きいとき（ $n \sim 30$ 程度と言われている）、母集団の分布の形によらず、標本抽出を繰り返したときの“サンプルの平均値”の分布は正規分布になる。

（中心極限定理の証明を、この資料の最後に載せています）

上記の項目は、統計解析を行うことを前提に何かの研究を実施する上で都合の良い性質を与えてくれます。すなわち、どの程度の個数の標本を抽出すれば、母集団の性質を推定したり異なる母集団同士を比べたりすることができるのか、ということを事前に検討をつけることができるということを意味します。これは「研究計画」を立てるのに役立ちます。もちろん、「研究計画」が立てられるような研究ばかりではなく、標本の個数が限られてしまうケースも多くありますが、そのような限定された場合でも推定能力を示した上で、何かしらの推論を行うことが可能です。それが数学の力であり、統計解析の真髄です。また、限られた標本を上手く使って、母集団の性質を調べるシミュレーション方法があります。乱数を使った方法がその代表例です（後期の演習では、実際にプログラムを作ります）。

3 信頼区間の推定

それでは、標本抽出によって得られたたった1回のデータの平均値 \bar{x} から、母集団の平均値（真の平均値）を推定してみようことを考えてみましょう。中心極限定理から、 \bar{x} は真の平均値 μ の周りに図のようにガウス分布するはずです。ここで n は標本抽出した数のこと（サンプル数）で、十分に大きな数とします（故に母集団の標準偏差 σ はこの標本の標準偏差で代用可能とする）。 \bar{x} はこのような確率分布を持って様々な値を取ることができます。1回だけのサンプリングなので、たまたまこの95%の区間から外れるような \bar{x} の値を取ることもあるでしょう。このとき、 \bar{x} は真の平均値 μ から $1.96\sigma/\sqrt{n}$ 以上異なってしまうことを意味します。よっ

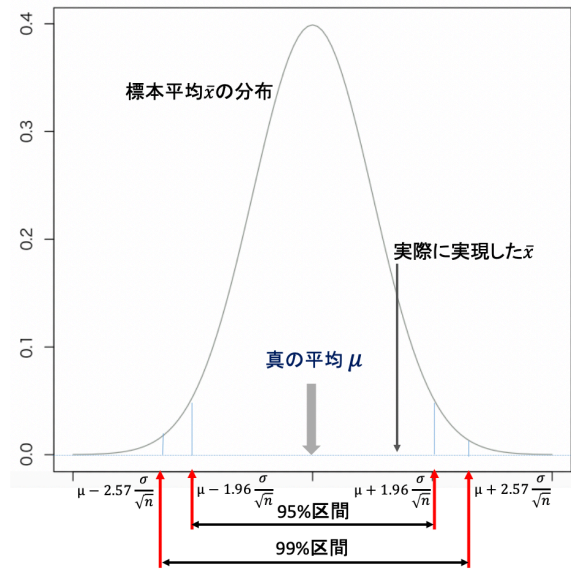
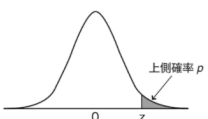


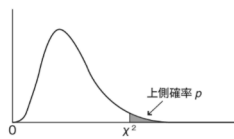
図2 標本抽出されたデータの平均 \bar{x} の分布

III χ^2 分布表 (表頭の上側確率と表側の自由度に対応する χ^2 値)

I 標準正規(z)分布表 (表頭表側の z 値に対応する上側確率)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0706	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367



df \ P	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.169	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156

図3 分布表の見方

て、たった1回のデータの平均値 \bar{x} から真の平均値 μ を推定するのであれば、95% の信頼区間として、

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \quad (1)$$

の範囲に真の平均値 μ があることになるでしょう (n 個のデータをサンプルして平均を求める作業を 100 回行くと、95 回はこの範囲に真の平均値 μ が来るでしょう)。教科書の第 4 章の例題や章末問題を行なうことで、理解の定着を図ってください。

4 付録：中心極限定理の証明

以下で定義する積率母関数

$$\phi(\theta) = \int e^{\theta z} p(z) dz \quad (2)$$

を使って証明する。ここで、 $p(z)$ は任意の確率分布である。今、平均 μ 、分散 σ^2 の母集団（これは正規分布とは限らないとする）から n 個の標本 X_1, X_2, \dots, X_n が独立に得られているとする*2。証明の戦略として、この n 個の標本の平均の確率分布の積率母関数が、平均 μ 、標準偏差 σ/\sqrt{n} の正規分布の積率母関数と一致するということを示す。

ここで、確率変数

$$Z_n = \frac{1}{n} \left(\frac{X_1 - \mu}{\sigma/\sqrt{n}} + \frac{X_2 - \mu}{\sigma/\sqrt{n}} + \dots + \frac{X_n - \mu}{\sigma/\sqrt{n}} \right) \quad (3)$$

の積率母関数を $\phi_n(\theta)$ とすると、

$$\begin{aligned} \phi_n(\theta) &= \int e^{\theta Z_n} p(Z_n) dZ_n \\ &= \int e^{\theta \frac{1}{n} \left(\frac{X_1 - \mu}{\sigma/\sqrt{n}} + \frac{X_2 - \mu}{\sigma/\sqrt{n}} + \dots + \frac{X_n - \mu}{\sigma/\sqrt{n}} \right)} p(Z_n) dZ_n \\ &= \prod_{i=1}^n \int e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}} p(Z_n) dZ_n \end{aligned} \quad (4)$$

独立同分布なので、

$$\begin{aligned} \phi_n(\theta) &= \prod_{i=1}^n \left(\int e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}} p(X_i) dX_i \right) \\ &= \left(\int e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}} p(X_i) dX_i \right)^n \end{aligned} \quad (5)$$

となる。つまり、平均 μ 、分散 σ^2 に従う確率変数 X_i (i は $1 \sim n$ で同じなのでどれでもいい) についての $e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}}$ の期待値の n 乗となる。 $e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}}$ はテイラー展開により、

$$e^{\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma}} = 1 + \frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma} + \frac{1}{2!} \left(\frac{\theta}{\sqrt{n}} \frac{X_i - \mu}{\sigma} \right)^2 + \dots \quad (6)$$

この期待値をとると、 X_i の平均は μ なので右辺第二項は 0、 $\left(\frac{X_i - \mu}{\sigma} \right)^2$ の期待値も 1 となるので、右辺第三項は $\frac{\theta^2}{2n}$ となる。結果、

$$\phi_n(\theta) = \left(1 + \frac{\theta^2}{2n} + \dots \right)^n \quad (7)$$

n が十分大きい時、

$$\phi_n(\theta) \sim (e^{\frac{\theta^2}{2n}})^n = e^{\frac{\theta^2}{2}} \quad (8)$$

*2 標本 X_1, X_2, \dots, X_n は同じ分布から独立に得られているので独立同分布 (i.i.d) などと呼ばれる

一方、正規分布の積率母関数は、

$$\begin{aligned}\phi_{gauss}(\theta) &= \int e^{\theta x} p(x) dx \\ &= \int e^{\theta x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx\end{aligned}\quad (9)$$

変数変換 $u = (x - \mu)/\sigma$ をすると、

$$\begin{aligned}\phi_{gauss}(\theta) &= e^{\theta\mu + \frac{\sigma^2\theta^2}{2}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(u-\sigma\theta)^2}{2}} du \\ &= e^{\theta\mu + \frac{\sigma^2\theta^2}{2}}\end{aligned}\quad (10)$$

式 (8) と比べると、 $\mu = 0$ および $\sigma^2 = 1$ 、つまりここで用いた正規分布の中でも特に標準正規分布 $\mathcal{N}(x|0, 1)$ であれば、その積率母関数が式 (8) と一致することになる。よって、確率変数

$$Z_n = \frac{1}{n} \left(\frac{X_1 - \mu}{\sigma/\sqrt{n}} + \frac{X_2 - \mu}{\sigma/\sqrt{n}} + \cdots + \frac{X_n - \mu}{\sigma/\sqrt{n}} \right) \quad (11)$$

つまり、標準偏差 σ/\sqrt{n} で平均 μ を持つサンプル平均は、標準正規分布 $\mathcal{N}(x|0, 1)$ に従う。わかりやすく言うと、母集団から n 個のサンプルを取ってきてその平均を取ると言う作業を何回も繰り返すと、その平均の確率分布は、

$$p(x) = \frac{1}{\sqrt{2\pi(\sigma/\sqrt{n})^2}} e^{-\frac{(x-\mu)^2}{2(\sigma/\sqrt{n})^2}} \quad (12)$$

という、平均 μ 、標準偏差 σ/\sqrt{n} の正規分布となる。