2011 年度前期　統計データ解析Ａ　第１５回

血液型と性格に関連はあるのか — 適合度検定，２標本検定

今回は，各グループに属するデータの数の比率が理論通りかどうかを検定する「適合度検定」と，２つの母集団が同質か異質かを調べる「２標本検定」を調べます。これらの検定は実際のデータ解析でよく用いられる手法ですが，確率の計算はちょっとむずかしいところがありますので，今回の講義ではこれらの手法の考え方を学んでください。

適合度検定

次のような問題を考えます。

> あるエンドウマメの交配実験の結果は，メンデルの法則によれば「黄色・丸」「黄色・しわ」「緑色・丸」「緑色・しわ」の４種類の形質のマメが 9 : 3 : 3 : 1 の割合で現れるはずだという。実際にこの実験を行ったところ，それぞれの形質をもったマメの数は 447, 131, 152, 38 であった。この実験においてメンデルの法則が成り立っているかどうかを有意水準 5% で検定せよ。

この問題で，もしも各形質をもつマメの個数が，正確に理論通りに 9 : 3 : 3 : 1 の比率になっているとすれば，得られたマメの総数は 768 ですから，それぞれの個数は 432, 144, 144, 48 になるはずです。

しかし，実際には，得られたマメは標本でしかありません。たとえ，世界全体で，各形質のマメが理論通りに 9 : 3 : 3 : 1 の比率で実っているとしても，この実験で得られたマメは，各マメの生育具合の違いなどの偶然によって，9 : 3 : 3 : 1 にはならないかもしれません。

そこで，実際に観測されたデータをカテゴリに分類したときの比率と，今仮定している理論（モデル）から予想される比率とのずれが，偶然起こりうる程度のものか，それともモデルの方が間違っていると考えたほうがいいのかを検定するのが，**適合度検定**です。ここでは，よく用いられる χ^2（カイ２乗）適合度検定を紹介します。

上の問題を一般的に考えると，次のようになります：母集団が k 個のグループに分けられるとしましょう。ただし，各グループは，同時に２つ以上のグループに属する個体はない (「排反」といいます) とします。実験によって得た n 個の標本のうち，グループ 1〜k に属するものの数が，おのおの $X_1, X_2, ..., X_k$ であったとき，「各グループに属する個体数（例題でいえば豆の数）の比率が $p_1, p_2, ..., p_k$ になる」というモデルが正しいかどうかを検定します。

グループ名	1	2	$\cdots$	k	合計
実際に観測された個体数	X_1	X_2	$\cdots$	X_k	n
理論で予想される比率	p_1	p_2	$\cdots$	p_k	1
理論で予想される個体数	np_1	np_2	$\cdots$	np_k	n

表 1: 適合度検定の考え方。$X_i = np_i$ とみなしてよいかどうかを検定する。

仮定したモデルによれば，n 個からなる標本に対してはグループ 1〜k に属する個体数は $np_1, np_2, ..., np_k$ になるはずです（表を参照)。ここで，各グループについて「現実に観測された個体数 X_i と，モデルか

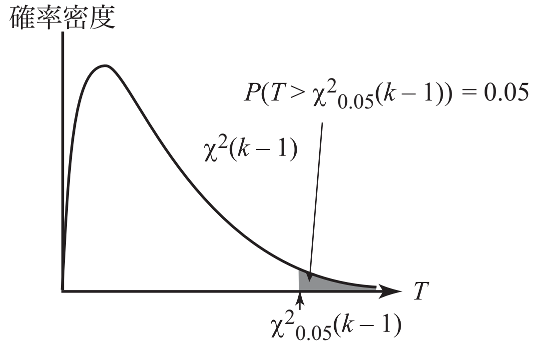

図 1: 自由度 $k-1$ の χ^2 分布の確率密度関数と，上側 5 パーセント点 $\chi^2_{0.05}(k-1)$

ら予想される値 np_i との差」の 2 乗の比率，すなわち $(X_i - np_i)^2/np_i$ を全部のグループに対して合計したもの，つまり

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{1}$$

を考えます。T は，観測値と理論値の差が大きいほど大きくなります。

T は，n が大きいとき，**自由度 $k-1$ の χ^2 分布**という確率分布モデルに近似的にしたがいます。その証明は複雑なのでここでは示しませんが，グループが 2 つ，つまり $k=2$ の場合について，付録 1 に説明を載せておきます。

さて，ここで行う検定では，帰無仮説 H_0 は「グループ $1, 2, \ldots, k$ に属する個体数の比率は $p_1, p_2, \ldots, p_k$ である」で，対立仮説 H_1 は「グループ $1, 2, \ldots, k$ に属する個体数の比率は $p_1, p_2, \ldots, p_k$ ではない」です。ここで，T の値が大きいほど，理論値と観測値とのずれが大きい，つまり帰無仮説の成立が疑わしい，ということなので，「T がそんな大きな値になる確率は，有意水準以下」のとき帰無仮説を棄却します。

自由度 $k-1$ の χ^2 分布の確率密度関数のグラフ（ヒストグラム）は，図 1 のような形をしています。このとき，自由度 $k-1$ の χ^2 分布にしたがう確率変数 T が，$\chi^2_{0.05}(k-1)$ より大きい確率が 5%であるとします（t 分布の場合と同様，「上側 5 パーセント点」といいます）。

したがって，有意水準が 5%のときは，帰無仮説が正しいとしたときに T の値を計算して，それが $\chi^2_{0.05}(k-1)$ よりも大きければ，帰無仮説は棄却されます。

結論は何なのか？ ー適合度検定の問題点

上の例では，帰無仮説 H_0 :「4 つのグループに属する個体数の比率は 9/16, 3/16, 3/16, 1/16 である」，対立仮説 H_1 :「4 つのグループに属する個体数の比率は 9/16, 3/16, 3/16, 1/16 でない」となります。そこで (1) 式の T の値を求めると $T = (447-432)^2/432 + (131-144)^2/144 + (152-144)^2/144 + (38-48)^2/48 = 4.22$ となります。4 つのグループに分かれるので自由度は 3 で，有意水準が 5%のとき，$T > \chi^2_{0.05}(3)$ のとき，H_0 は棄却されます。数表より，$\chi^2_{0.05}(3) = 7.8147$ ですから，$T > \chi^2_{0.05}(3)$ ではなく，H_0 は棄却されません。

これまでに説明した検定の考え方では，帰無仮説が「モデルは実測データに適合している」となっていますから，帰無仮説が棄却されたときは「モデルは実測データに適合していないと言える」と結論さ

れ，棄却されないときは「モデルは実測データに適合していないとまでは言えない」という結論が得られるはずです。したがって，今回の例の結論は「メンデルの法則が成り立っていないとまでは言えない」ということになるはずです。

ところが，適合度検定では中心極限定理を用いています（付録１参照）から，標本サイズが大きくないと理論の正当性が保たれません。しかし一方で，標本サイズが大きいと棄却域が広くなり，帰無仮説が棄却されやすくなります[1]。例えば，母平均の両側検定を考えてみましょう。標本サイズが大きくなると，母平均の信頼区間は狭くなります。両側検定では，帰無仮説で述べた母平均の値が信頼区間に入っていなければ棄却するわけですから，棄却域は広くなります。したがって，標本サイズが大きくなると，理論比率と実際に観察された比率のほんのわずかの違いによっても帰無仮説が棄却され，「モデルは適合しないと言える」という結論が出てしまいます。

現実には，たとえ理論通りの現象が起こっていても，現象には誤差があり，理論通りの比率が実際に観察されることはまずありません。したがって，標本が多ければ，わずかの誤差を検出して帰無仮説が棄却され，どんな場合でも「モデルは適合しない」という結論が出てしまいます。これでは，この検定は役に立ちません。

そこで，適合度検定では，通常の検定とは逆に「標本サイズが大きく，検出力が十分に高いのに帰無仮説が棄却されないときは，『帰無仮説は正しい。モデルは実測データに適合している』あるいは『実測データはモデルと矛盾しない』と言える」という結論を導く，という考え方があります。しかし，標本が十分多くないときには，棄却されないからといって間違った結論を出してしまうことがあります。また，そんなに標本が多ければ，検定などしなくても，モデルによる比率と実測データにおける比率を直接比較するだけでいいはずです。

第８回の講義でくじびきの例を使って説明した，検定のそもそもの考え方を思い出してください。それは

> 「実際にひいた少ない本数のくじが全部はずれだったとき，たったそれだけの数のくじでも
> 『全部はずれ』なんてことは到底（有意水準より小さな確率でしか）起こらない」
> →「『半分の確率で当たる』という台詞を無理に信じるよりも，『半分の確率で当たるなんて
> ウソ』と考えるほうが自然」

というものでした。このような推測をしなければならないのは，実際にひくことのできるくじの本数が少ないからです。もし，くじ箱の中のくじの大半をひくことができるのなら，半分の確率で当たるかどうかは一目瞭然です。

つまり，標本サイズが大きくないと成り立たず，しかも「帰無仮説が棄却されないとき，積極的な結論が得られる」という検定は，上のような検定の考え方や有意水準の意味を逸脱したものです。したがって，統計学者の間では適合度検定の有用性を否定する見解もあり，使用には注意が必要です。

２標本 t 検定

次のような問題[2]を考えてみましょう。

[1] 帰無仮説が棄却される確率を「検出力」といいます。2010 年度後期「情報統計学」第 11 回を参照してください。
[2] これは架空のデータで，実際の調査で得られたものではありません。

日本人全体から，血液型がA型の人 20 人，B型の人 18 人を無作為抽出して，「奔放さ」を評価する性格テストを行いました。その結果，A型のグループでは得点の平均は 65 で不偏分散は 225，B型のグループでは平均 70 で不偏分散 144 でした。日本人全体について，A型の人とB型の人とではこの性格テストの平均点に差があると言えるか，有意水準 5%で検定してください。ただし，A型の人々・B型の人々それぞれの母集団の性格テストの得点の分布は正規分布になると仮定し，また，それぞれの母集団での母分散は等しいとします。

この問題のように，２つの異なった母集団それぞれから標本をとりだして，２つの母集団の比較を行う問題を，２標本問題といいます。上の問題を一般的に考えると次のようになります。

母集団分布が正規分布 $N(\mu_A, \sigma_A^2)$ の母集団 A から，n_A 個からなる標本を，
母集団分布が正規分布 $N(\mu_B, \sigma_B^2)$ の母集団 B から，n_B 個からなる標本を，

それぞれ取り出します。いま，２つの母集団の分散は等しいが未知である，すなわち $\sigma_A^2 = \sigma_B^2 = \sigma^2$ と仮定できるが σ^2 は未知である，とするとき，両母集団の標本平均の差 $\bar{X}_A - \bar{X}_B$ のしたがう分布を考えます。

母集団 A から取り出した標本の標本平均 $\bar{X}_A$ がしたがう分布は $N(\mu_A, \sigma_A^2/n_A)$ で，母集団 B から取り出した標本の標本平均 $\bar{X}_B$ がしたがう分布は $N(\mu_B, \sigma_B^2/n_B)$ です。このとき $\bar{X}_A$ と $\bar{X}_B$ が独立であれば，$\bar{X}_A - \bar{X}_B$ は $N(\mu_A - \mu_B, (\sigma_A^2/n_A) + (\sigma_B^2/n_B))$ にしたがいます[3]。さらに $\sigma_A^2 = \sigma_B^2 = \sigma^2$ ですから，したがう分布は $N(\mu_A - \mu_B, (1/n_A + 1/n_B)\sigma^2)$ となります。よって，

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\sigma^2}} \tag{2}$$

は標準正規分布 $N(0, 1)$ にしたがいます。

ここで，σ^2 は未知ですから，標本から計算できる不偏分散で代用する必要があります。この不偏分散は，母集団 A，母集団 B それぞれからの標本からの不偏分散を混ぜたものになり，この不偏分散を「プールされた不偏分散」といいます。s_A^2 と s_B^2 を母集団 A，母集団 B それぞれからの標本における不偏分散とすると，プールされた不偏分散 s^2 は

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \tag{3}$$

となります（詳しくは付録２を見てください）。このとき，(2) 式の σ^2 を (3) 式の s^2 でおきかえた

$$t = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)s^2}} \tag{4}$$

を**２標本 t 統計量**といい，これは**自由度** $(n_A + n_B - 2)$ **の t 分布**にしたがいます。

２標本 t 統計量を用いると，２つの母集団についての母平均の差に関する検定が行えます。ここでは，A 型の人の平均点と B 型の人の平均点が「偶然とは言えないほどの差があるか」が問題なのであって，どちらが大きいかは問題ではありません。したがって，帰無仮説 $H_0 : \mu_A = \mu_B$，対立仮説 $H_1 : \mu_A \neq \mu_B$ の両側検定を行います。

[3] これを証明するには，多変量確率分布の知識が必要です。2010 年度後期「情報統計学」第 13 回を参照してください。

ここで，有意水準を 5% としましょう。２標本 t 統計量は自由度 $(n_A + n_B - 2)$ の t 分布にしたがうので，第１４回と同様に考えると

$$P\left(-t_{0.025}(n_A + n_B - 2) \leq \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) s^2}} \leq t_{0.025}(n_A + n_B - 2)\right) = 0.95 \tag{5}$$

となることがわかります。

帰無仮説が正しいとしたとき，$\mu_A = \mu_B$ ですから，$\mu_A - \mu_B = 0$ です。このとき，t 統計量は

$$\frac{\bar{X}_A - \bar{X}_B}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) s^2}} \tag{6}$$

となります。(5) 式の通り，t 統計量は「$-t_{0.025}(n_A + n_B - 2)$ 以上，$t_{0.025}(n_A + n_B - 2)$ 以下」の範囲に入っている確率が 95%です。したがって，今得られている標本から計算した標本平均 $\bar{X}_A, \bar{X}_B$ と「プールされた不偏分散」s^2 を使って (6) 式を計算したとき，それが「$-t_{0.025}(n_A + n_B - 2)$ 以上，$t_{0.025}(n_A + n_B - 2)$ 以下」の範囲に入っていなければ，「5%の確率でしか起こらないことが起きている」ことになるので，帰無仮説を棄却します。

ここでは，両母集団の母分散が等しい，すなわち $\sigma_A^2 = \sigma_B^2 = \sigma^2$ という，２標本問題のいちばん簡単な場合を示しましたが，その他さまざまな場合の２標本問題を取り扱う方法が研究されています。

血液型ー性格関連説について

血液型－性格関連説は，なぜか日本では人気のある学説です。現在のところ，「血液型と性格の間に関連がある」という積極的証拠は見つかっておらず，医学者や心理学者は否定的に見ているようですが，一般にはこの説を支持する人も多いようです。

講義のサイトの「統計データ・ツールへのリンク」に，「血液型と性格の関連について」という項があって，この説に関するサイトをいくつか掲載してあります。肯定派と否定派の両方を載せてありますから，その真偽は自分で考えて判断してください。また，これらのサイトの内容を，この講義で得た知識をふまえて読むことによって，統計的推測の実際的応用や，あるいは統計的推測の限界について学ぶことができると思います。

「統計データ解析Ａ」の最後の講義にあたって

わが国や世界の多数の国は，なぜ「民主主義」を採用しているのでしょうか？

今日の演習

上で例として示した，血液型と性格の関連についての問題に答えてください。

付録１：χ^2 分布と適合度検定について

χ^2 分布は，標準正規分布から派生する確率分布のひとつです。確率変数 $Z_1, Z_2, \ldots, Z_n$ が互いに独立で，それぞれが標準正規分布 $N(0,1)$ にしたがうとします。このとき，

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{A1}$$

がしたがう確率分布を「自由度 n の χ^2 分布」といい，$\chi^2(n)$ で表します。

さて，適合度検定において，グループが２つ，つまり $k = 2$ の場合を考えます。このとき，本文 (1) 式の T がどうなるかをみてみましょう。グループが２つしかないので，各グループの個体数 X_1, X_2 の合計は全個体数 n で，また各グループの理論比率 p_1, p_2 の合計は１です。このとき，T は

$$\begin{aligned} T &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \qquad \text{(A2)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{((n - X_1) - n(1 - p_1))^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)} \\ &= (X_1 - np_1)^2 \left\{ \frac{(1 - p_1) + p_1}{np_1(1 - p_1)} \right\} = \left\{ \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right\}^2 \end{aligned}$$

となります。

ここで，p_1 は「ある個体がグループ１に属する確率」と考えられるので，「n 個の個体のうち，グループ１に属している数」を表す X_1 は２項分布 $B(n, p_1)$ にしたがいます。$B(n, p_1)$ の期待値は np_1，分散は $np_1(1 - p_1)$ ですから，中心極限定理により，n が大きいとき，(A3) 式の $\{\}$ 内は標準正規分布にしたがいます。よって，T は自由度１（すなわち $k - 1$）の χ^2 分布にしたがいます。

χ^2 分布は，母分散の推定・検定など，母集団分布が正規分布であるときの，さまざまな統計的推測に現れる確率分布です。また，第 11 回で説明した t 分布は，χ^2 分布からさらに派生する確率分布モデルです。くわしくは，2010 年度後期「情報統計学」第 14, 15 回を参照してください。

付録２：「プールされた不偏分散」について

本文 (4) 式の，「プールされた不偏分散」s^2 を使った t 統計量の式を，以下のように変形します。

$$
\begin{aligned}
t &= \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) s^2}} && \text{(A3)} \\
&= \frac{\dfrac{\left\{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)\right\}}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \sigma^2}}}{\dfrac{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) s^2}}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \sigma^2}}} \\
&= \frac{\dfrac{\left\{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)\right\}}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \sigma^2}}}{\sqrt{s^2/\sigma^2}}
\end{aligned}
$$

ところで，母集団 A，母集団 B それぞれの標本における不偏分散 s_A^2, s_B^2 について，χ^2 分布と不偏分散の関係[4]から

$$(n_A - 1)\frac{s_A^2}{\sigma^2} \sim \chi^2(n_A - 1), \quad (n_B - 1)\frac{s_B^2}{\sigma^2} \sim \chi^2(n_B - 1) \qquad \text{(A4)}$$

という関係があります。付録 1 で述べた χ^2 分布の定義のとおり，「自由度 n の χ^2 分布」は「それぞれが標準正規分布にしたがう n 個の独立な確率変数の合計」がしたがう分布ですから，(A4) 式の２つの式は次のように合計することができます。

$$
\begin{aligned}
(n_A - 1)\frac{s_A^2}{\sigma^2} + (n_B - 1)\frac{s_B^2}{\sigma^2} &\sim \chi^2((n_A - 1) + (n_B - 1)) && \text{(A5)} \\
\text{つまり} \quad \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{\sigma^2} &\sim \chi^2(n_A + n_B - 2)
\end{aligned}
$$

そこで，２標本問題における不偏分散 s^2 を，本文 (3) 式のとおり

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \qquad \text{(A6)}$$

と定義すると，(A4) 式の分母は

$$
\begin{aligned}
\sqrt{\frac{s^2}{\sigma^2}} &= \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \cdot \frac{1}{\sigma^2}} && \text{(A7)} \\
&= \sqrt{\frac{1}{n_A + n_B - 2} \cdot \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{\sigma^2}}
\end{aligned}
$$

となり，(A6) 式より，$\sqrt{\ }$の中の後半が自由度 $(n_A + n_B - 2)$ の χ^2 分布にしたがいますから，t 分布の定義[5]から，(A4) 式の t は自由度 $(n_A + n_B - 2)$ の t 分布にしたがいます。よって，プールされた不偏分散を本文 (3) 式の s^2 で定義します。

[4] 2010 年度後期「情報統計学」第 13 回で説明しています。講義録にウェブページからリンクしています。
[5] 2010 年度後期「情報統計学」第 14 回で説明しています。講義録にウェブページからリンクしています。