

$\chi^2$  分布と推定・検定

第 9 回の講義で、正規母集団の母分散がわかっているものと仮定して、母平均の区間推定を説明しました。しかし、母平均がわからないのに母分散がわかっているというのは変な話で、母平均がわからないのなら母分散もわからないはずで、そこで、母平均のかわりにその点推定量として標本平均を用いるように、母分散の点推定量として不偏分散を用いて、区間推定や検定を行う方法が必要です。そこで今日は、不偏分散と母分散の関係の手がかりになる  $\chi^2$  (カイ 2 乗) 分布と、それを用いて母分散の区間推定や検定を行う方法を紹介します。さらに  $\chi^2$  分布は、次回説明する、標本平均と不偏分散から母平均の推定・検定を行う方法 ( $t$  分布にもとづく推定・検定) の基礎になります。

また、 $\chi^2$  分布を利用した統計的推測の方法として、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** をとりあげます。

 $\chi^2$  分布

問題 1 : ある正規母集団は、母分散  $\sigma^2 = 15$  であるという。この母集団からサイズ  $n = 10$  の標本をとったとき、「不偏分散  $s^2$  が  $a$  を超える確率が 0.05」であるような  $a$  の値を求めよ。

確率変数  $Z_1, Z_2, \dots, Z_n$  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、

$$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_n^2 \quad (1)$$

がしたがう確率分布を **自由度  $n$  の  $\chi^2$  (カイ 2 乗) 分布** といい、記号  $\chi^2(n)$  で表します<sup>1</sup>。

不偏分散と  $\chi^2$  分布の間には、次のような関係があります。

母平均  $\mu$ , 母分散  $\sigma^2$  の正規分布  $N(\mu, \sigma^2)$  にしたがう母集団から、 $n$  個からなる標本  $X_1, X_2, \dots, X_n$  をとる。その標本平均を  $\bar{X}$  とし、不偏分散を  $s^2$  とするとき、 $(n-1)s^2/\sigma^2$  は自由度  $n-1$  の  $\chi^2$  分布  $\chi^2(n-1)$  にしたがう。

その証明は付録 2 に収録しています。ここでは標本サイズが 2 の場合について上のことを確かめてみましょう。  $n = 2$  ですから、不偏分散  $s^2$  は

$$s^2 = \frac{1}{2-1} \{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\} \quad (2)$$

となります。  $\bar{X} = (X_1 + X_2)/2$  ですから、  $X_1 - \bar{X}$  と  $X_2 - \bar{X}$  は独立ではありません。したがって、このまま  $\chi^2$  分布の定義を使うことはできません。そこで  $\bar{X} = (X_1 + X_2)/2$  という関係を使ってこの式を変形す

<sup>1</sup>詳細は付録 1 をご覧ください。

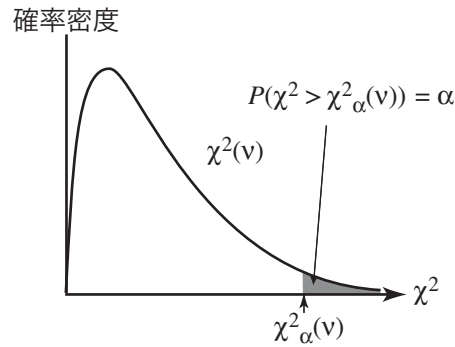


図 1: 自由度  $\nu$  の  $\chi^2$  分布の上側確率  $100\alpha\%$  のパーセント点  $\chi^2_\alpha(\nu)$

ると,

$$\begin{aligned}
 s^2 &= (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 \\
 &= \left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2 \\
 &= \left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2 \\
 &= \frac{(X_1 - X_2)^2}{2}
 \end{aligned} \tag{3}$$

となります。ここで  $X_1 - X_2 = Y$  とおくと、 $s^2 = Y^2/2$  となります。  $X_1, X_2$  は独立で  $N(\mu, \sigma^2)$  にしたがいますから、 $Y = X_1 - X_2$  は平均 0、分散  $2\sigma^2$  の正規分布  $N(0, 2\sigma^2)$  にしたがいます<sup>2</sup>。ですから、 $Y/2\sigma$  は標準正規分布  $N(0, 1)$  にしたがいます。そこで、

$$s^2 = \frac{Y^2}{2} = \frac{1}{2} \left(\frac{Y}{\sqrt{2}\sigma}\right)^2 \cdot 2\sigma^2 = \left(\frac{Y}{\sqrt{2}\sigma}\right)^2 \cdot \sigma^2 \tag{4}$$

すなわち

$$(2-1) \frac{s^2}{\sigma^2} = \left(\frac{Y}{\sqrt{2}\sigma}\right)^2 \tag{5}$$

と変形すると、 $\chi^2$  分布の定義から  $(2-1)s^2/\sigma^2$  が自由度 1 の  $\chi^2$  分布にしたがうことがわかります。もともと 2 つの独立な変数  $X_1, X_2$  があつたのが、不偏分散を求めると独立な変数が  $Y$  だけ 1 つになっていることに注意してください。

自由度  $\nu$  の  $\chi^2$  分布  $\chi^2(\nu)$  にしたがう確率変数  $\chi^2$  がある範囲の値をとる確率を知るには、数表を利用することができます。  $\chi^2$  の数表では、各自由度  $\nu$  (縦軸) と定数  $\alpha$  (横軸) に対して、 $P(\chi^2 > x) = \alpha$  となるような  $x$  が縦  $\nu$ ・横  $\alpha$  の交点の値を読むことで求められます。この  $x$  を **自由度  $\nu$  の  $\chi^2$  分布の上側確率  $100\alpha\%$  のパーセント点** といい  $\chi^2_\alpha(\nu)$  で表します。

問題 1 では、 $P(s^2 > a) = 0.05$  となる  $a$  を求めます。  $s^2 > a$  のとき  $(n-1)s^2/\sigma^2 > (n-1)a/\sigma^2$  です。問題の条件は  $\sigma^2 = 15$  で自由度  $n-1 = 9$  ですから、 $(n-1)a/\sigma^2 = 9a/15 = 0.6a$  です。よって、求める  $a$  は  $P((n-1)s^2/\sigma^2 > 0.6a) = 0.05$  となる  $a$  です。  $(n-1)s^2/\sigma^2$  は自由度  $n-1$  の  $\chi^2$  分布にしたがうので、 $P((n-1)s^2/\sigma^2 > 0.6a) = 0.05$  となる  $0.6a$  は上側確率 5% のパーセント点  $\chi^2_{0.05}(9)$  で、数表から

<sup>2</sup>第 1 2 回の講義を参照。 2 つの独立な確率変数の 和 の分散も、差 の分散も、どちらもそれぞれの分散の 和 になります。

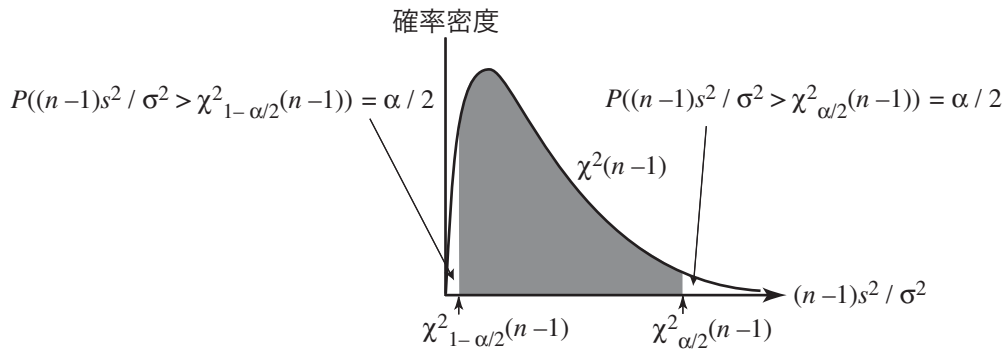


図 2:  $\chi^2$  分布での信頼係数  $1 - \alpha$  信頼区間

$0.6a = 16.9190$ , よって  $a = 28.1983$  となります。母分散が 15 であっても、不偏分散が 28 を超える確率が 5% ほどあることとなります。

## 母分散の区間推定

次の問題を考えてみましょう。

問題 2 : ある試験の受験者から 25 人を抽出して点数を調べたところ、標本平均 53, 不偏分散 48 であった。受験者全体の点数の分布が正規分布であるとして、母分散の 90% 信頼区間を求めよ。

前節で述べた通り、正規母集団について、標本サイズ  $n$ , 母分散  $\sigma^2$ , 不偏分散  $s^2$  とするとき、 $(n-1)s^2/\sigma^2$  は自由度  $n-1$  の  $\chi^2$  分布にしたがいます。よって、パーセント点の定義から

$$P\left(\chi^2_{1-\alpha/2}(n-1) \leq (n-1) \frac{s^2}{\sigma^2} \leq \chi^2_{\alpha/2}(n-1)\right) = 1 - \alpha \quad (6)$$

となります。これを  $\sigma^2$  について解くと、

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right) = 1 - \alpha \quad (7)$$

ですから、信頼係数  $1 - \alpha$  の信頼区間は

$$\left[ \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \right] \quad (8)$$

となります。分母にあるパーセント点は、数表から求めることができます。

この問題例の場合は、不偏分散  $s^2 = 48$  で、自由度  $n-1 = 24$  です。 $\sigma^2$  の信頼係数 90% 信頼区間は

$$\left[ \frac{24 \times 48}{\chi^2_{0.05}(24)}, \frac{24 \times 48}{\chi^2_{0.95}(24)} \right] \quad (9)$$

です。数表から  $\chi_{0.05}^2(24) = 36.4150$ ,  $\chi_{0.95}^2(24) = 13.8484$  ですから,  $\sigma^2$  の 90% 信頼区間は [31.64, 83.19] となります。

## 適合度検定

次のような問題を考えます。

ある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9:3:3:1の比率になっているとすれば, マメの総数は768ですから, それぞれの個数は432, 144, 144, 48になるはずですが。このように, 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  $\chi^2$  (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  $X_1, X_2, \dots, X_k$  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  $p_1, p_2, \dots, p_k$  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	...	$k$	合計
実際に観測された個体数	$X_1$	$X_2$	...	$X_k$	$n$
理論で予想される比率	$p_1$	$p_2$	...	$p_k$	1
理論で予想される個体数	$np_1$	$np_2$	...	$np_k$	$n$

表1: 適合度検定の考え方。  $X_i = np_i$  といえる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  $np_1, np_2, \dots, np_k$  になるはずですが(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

$$T = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \quad (10)$$

を考えます。  $T$  は, 観測値と理論値の差が大きいくほど大きくなります。

$T$  は,  $n$  が大きいとき, 自由度  $k-1$  の  $\chi^2$  分布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3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, ...,  $k$ に属する個体数の比率は  $p_1, p_2, \dots, p_k$  である」で, 対立仮説  $H_1$  は「グループ1, 2, ...,  $k$ に属する個体数の比率は  $p_1, p_2, \dots, p_k$  ではない」です。

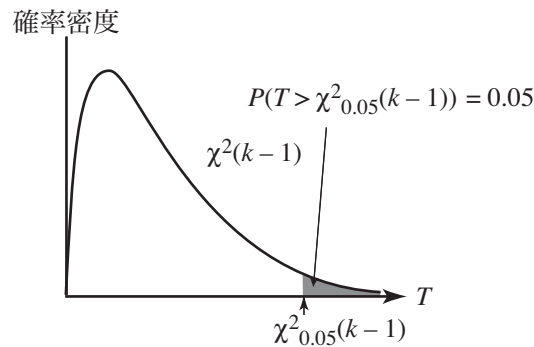


図3: 自由度  $k-1$  の  $\chi^2$  分布の確率密度関数と、上側5パーセント点  $\chi^2_{0.05}(k-1)$

ここで、 $T$  の値が大きいくほど、理論値と観測値とのずれが大きいく、つまり帰無仮説の成立が疑わしく、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  $\chi^2$  分布の確率密度関数のグラフ（ヒストグラム）は、図3のような形をしています。このとき、自由度  $k-1$  の  $\chi^2$  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？ - 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(10)式の  $T$  の値を求めると  $T = (447-432)^2/432 + (131-144)^2/144 + (152-144)^2/144 + (3848)^2/48 = 4.22$  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が。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が。

ところが、ある程度たくさん標本を集めないと、「比率」を推測して検定をしようとは思いません。しかし一方で、第11回の講義で述べたように、標本サイズが大きくと検出力が大きくなります。つまり、棄却域が広くなり、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が、

第11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』というおじさんの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。

---

## 今日の演習

ある機械が製造する製品の寸法は、正規分布にしたがうとします。製品から、10個のデータからなる標本を取り出して測定すると、その寸法 (cm) は次の通りでした。

101, 101, 102, 99, 99, 101, 99, 100, 98, 100

このとき、製品の寸法の母分散の95%信頼区間を求めて下さい。

---

## 付録1： $\chi^2$ 分布の定義と正規分布との関係

$\chi^2$  分布はガンマ分布という確率分布の1種で、その確率密度関数は次の式で表されます。

$$\begin{aligned} f(x) &= \frac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2} \Gamma(\nu/2)} \quad (x \geq 0) \\ &= 0 \quad (x < 0) \end{aligned} \tag{A1}$$

パラメータ  $\nu$  を自由度とよびます。また、 $\Gamma()$  はガンマ関数という関数で、次のように定義されます。

$$\Gamma(\alpha) = \int_{-\infty}^{\infty} x^{\alpha-1} e^{-x} dx \tag{A2}$$

この確率密度関数から、モーメント母関数を求めてみましょう。モーメント母関数の定義により

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \frac{1}{2^{v/2}\Gamma(v/2)} \int_{-\infty}^{\infty} e^{tx} x^{v/2-1} e^{-x/2} dx \\ &= \frac{1}{2^{v/2}\Gamma(v/2)} \int_{-\infty}^{\infty} x^{v/2-1} e^{-\frac{x(1-2t)}{2}} dx \end{aligned} \quad (A3)$$

となります。ここで、 $z = x(1-2t)/2$  という変数変換を行うと、 $dx = 2dz/(1-2t)$  ですから

$$\begin{aligned} M_X(t) &= \frac{1}{2^{v/2}\Gamma(v/2)} \int_{-\infty}^{\infty} \left(\frac{2z}{1-2t}\right)^{v/2-1} e^{-z} \frac{2}{1-2t} dz \\ &= \frac{\left(\frac{2}{1-2t}\right)^{v/2}}{2^{v/2}\Gamma(v/2)} \int_{-\infty}^{\infty} z^{v/2-1} e^{-z} dz \end{aligned} \quad (A4)$$

となり、この式の積分は(A2)式より  $\Gamma(v/2)$  ですから、

$$M_X(t) = \frac{\left(\frac{2}{1-2t}\right)^{v/2}}{2^{v/2}\Gamma(v/2)} \Gamma(v/2) = (1-2t)^{-v/2} \quad (A5)$$

となります。

さて、確率変数  $Z_1, Z_2, \dots, Z_k$  が互いに独立で、それぞれが標準正規分布  $N(0, 1)$  にしたがるとうします。このとき、これらの確率変数の2乗の和  $V = Z_1^2 + Z_2^2 + \dots + Z_k^2$  がしたがる分布を考えてみます。第12回の講義の演習問題で示したように、独立な確率変数の和のモーメント母関数は、それぞれのモーメント母関数の積になるので、

$$M_V(t) = M_{Z_1^2 + Z_2^2 + \dots + Z_k^2}(t) = M_{Z_1^2}(t) \cdot M_{Z_2^2}(t) \cdot \dots \cdot M_{Z_k^2}(t) = \{M_{Z_1^2}(t)\}^k \quad (A6)$$

となります。 $Z_1$  は  $N(0, 1)$  にしたがるので、

$$\begin{aligned} M_{Z_1^2}(t) = E[e^{tZ_1^2}] &= \int_{-\infty}^{\infty} e^{tz^2} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2(1-2t)/2} dz \end{aligned} \quad (A7)$$

となり、ここで  $y = z\sqrt{1-2t}$  という変数変換を行うと、 $dz = dy/\sqrt{1-2t}$  ですから

$$\begin{aligned} M_{Z_1^2}(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2(1-2t)/2(1-2t)} \cdot \frac{1}{\sqrt{1-2t}} dy \\ &= (1-2t)^{-1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \end{aligned} \quad (A8)$$

となりますが、下段の積分の中は標準正規分布の密度関数ですから積分の値は1です。よって

$$M_V(t) = \{M_{Z_1^2}(t)\}^k = (1-2t)^{-k/2} \quad (A9)$$

です。これは、(A5)式で示した通り、自由度  $k$  の  $\chi^2$  分布のモーメント母関数です。つまり、標準正規分布  $N(0, 1)$  にしたがる独立な確率変数  $k$  個の2乗の和  $V = Z_1^2 + Z_2^2 + \dots + Z_k^2$  は自由度  $k$  の  $\chi^2$  分布にしたがるがいます。

## 付録2：不偏分散と $\chi^2$ 分布との関係

母平均 $\mu$ 、母分散 $\sigma^2$ の正規分布 $N(\mu, \sigma^2)$ にしたがう母集団から、 $n$ 個からなる標本 $X_1, X_2, \dots, X_n$ をとります。その標本平均を $\bar{X}$ とし、標本分散（不偏分散） $s^2$ を

$$s^2 = \frac{1}{n-1} \left\{ (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right\} \quad (\text{A10})$$

とします。このとき、つぎの関係がわかります。

$$\begin{aligned} (n-1)s^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n [(x_i - \mu) - (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \cdot n(\bar{X} - \mu) + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned} \quad (\text{A11})$$

上の式の両辺を $\sigma^2$ で割ると

$$\frac{(n-1)s^2}{\sigma^2} + \left( \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \right)^2 = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \quad (\text{A12})$$

が得られます。この関係を $J + K^2 = L$ と表します。ここで、「 $X_i$ が独立で正規分布にしたがうとき、 $s^2$ と $\bar{X}$ は独立である」<sup>3</sup>ことを用いると、 $J$ と $K^2$ は独立となりますから

$$\begin{aligned} M_{J+K^2}(t) &= M_L(t) \\ M_J(t)M_{K^2}(t) &= M_L(t) \\ \text{ゆえに } M_J(t) &= \frac{M_L(t)}{M_{K^2}(t)} \end{aligned} \quad (\text{A13})$$

が得られます。 $X_i$ は $N(\mu, \sigma^2)$ にしたがひ、各々は独立なので、 $L$ は自由度 $n$ の $\chi^2$ 分布にしたがひます。また、 $X$ は $N(\mu, \sigma^2/n)$ にしたがひますから、 $K$ は $N(0, 1)$ にしたがひます。よって、 $K^2$ は自由度1の $\chi^2$ 分布にしたがひます。以上のことから、(A9)式で表される $\chi^2$ 分布のモーメント母関数を使って(A13)式を表すと

$$M_J(t) = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2} \quad (\text{A14})$$

となります。この式は、 $J$ が自由度 $n-1$ の $\chi^2$ 分布にしたがひを表しています。つまり、 $(n-1)s^2/\sigma^2$ は自由度 $n-1$ の $\chi^2$ 分布 $\chi^2(n-1)$ にしたがひます。

<sup>3</sup>この証明は省略します。ホーエル／浅井・村上「入門数理統計学（第4版）」付録6を参照して下さい。



### 付録3： $\chi^2$ 分布と適合度検定について

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(10)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

$$\begin{aligned} T &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} && (A15) \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{((n - X_1) - n(1 - p_1))^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)} \\ &= (X_1 - np_1)^2 \left\{ \frac{(1 - p_1) + p_1}{np_1(1 - p_1)} \right\} = \left\{ \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right\}^2 \end{aligned}$$

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、(A15)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  $\chi^2$  分布にしたがいます。