８章 標準正規分布から導かれる分布

正規分布は統計学の基礎であると言いましたが、これから学ぶ推定や検定の分野では、標準正規分布から導かれるいくつかの分布が重要な役割を演じます。この章ではこれらの分布について簡単に説明します。なぜこれらの分布が必要なのかについては、もう少し後になって考えてみましょう。

8.1 χ^2分布

この表題のχ^2はカイ 2 乗と読みます。互いに独立な確率変数X_iがそれぞれ標準正規分布に従うとき、$\chi^2 = \sum_{i=1}^{n} X_i^2$で与えられる新しい確率変数$\chi^2$は、$\chi^2$分布と呼ばれる確率分布に従います。$\chi^2$分布は自由度と呼ばれるパラメータを持っています。自由度は標準正規分布に従う確率変数の 2 乗をいくつ足し上げたかによって与えられるもので、文字通り自由に値のとれる確率変数の数というように考えればよいでしょう。この場合n個の足し上げを行っていますので、自由度nのχ^2分布となり、これをχ_n^2分布とも表わします。以上のことを改めてまとめておきましょう。

$$X_i \sim N(0,1) \text{分布のとき、} \chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \text{分布} \qquad (8.1)$$

次に実際にχ^2分布の確率密度関数の形についてグラフを見てみましょう。図 8-1 は自由度が 1, 2, 3, 4 のときの確率密度関数です。自由度 3 以上でピークができますが、自由度が増すごとにこのピークが右へずれていきます。

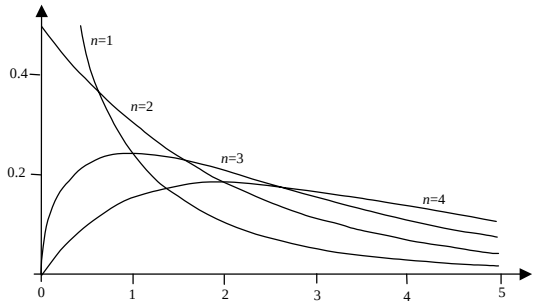

図 8-1　自由度 1,2,3,4 のχ^2分布の確率密度関数

Excel には、χ^2分布の確率とχ^2値の関係を与える関数があります。自由度をnとして、図 8-2 の確率pとχ^2値xとの関係は、以下で与えられます。

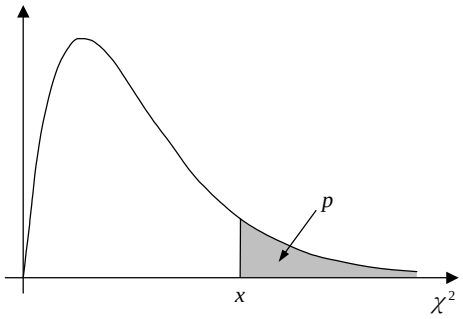

図 8-2 χ^2 分布の確率

$$p = chidist(x, n), \quad x = chiinv(p, n)$$

ここに、*chi* はカイと読み、*dist* は distribution で分布、*inv* は inverse で逆という意味です。

例として、以下のような値を求めてみましょう。簡単に求められますので、説明の必要はないと思われます。

自由度 5, χ^2 値 10 のときの上側確率　　$chidist(10, 5) = 0.075235$

自由度 10, 上側確率 0.05 のときの χ^2 値　　$chiinv(0.05, 10) = 18.30703$

ここに上側確率とは、確率変数の値が、ある値より大きくなる確率のことを言います。説明は後に譲りますが、χ^2 分布は分散に関係した分布です。

8.2 F 分布

分散についての推定や検定でよく利用される分布として、F 分布と呼ばれるものがあります。これは、χ^2 分布に従う 2 つの確率変数を自由度で割ったものの比として表わされます。

2 つの独立な確率変数 χ_1^2 と χ_2^2 が、それぞれ自由度 n_1 と n_2 の χ^2 分布をするとき、その比率 $F = \dfrac{\chi_1^2/n_1}{\chi_2^2/n_2}$ は自由度 n_1, n_2 の F 分布に従います。このように F 分布は自由度を 2 つ持っている確率分布です。自由度 n_1, n_2 の F 分布は F_{n_1,n_2} 分布とも表現することができます。以上をまとめて書いておきましょう。

$$\chi_1^2 \sim \chi_{n_1}^2 \text{ 分布}, \quad \chi_2^2 \sim \chi_{n_2}^2 \text{ 分布のとき、 } F = \frac{\chi_1^2/n_1}{\chi_2^2/n_2} \sim F_{n_1, n_2} \text{ 分布} \qquad (8.2)$$

自由度 2,4、4,8、8,16、8,4 の F 分布について、確率密度関数の具体的な形をグラフ

に表わすと図 8-3 のようになります。

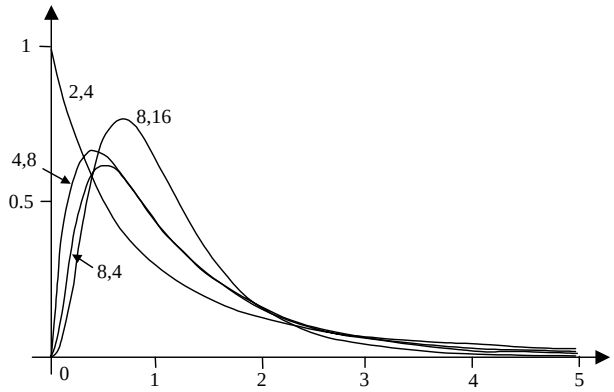

図 8-3　自由度(4,4), (8,8), (16,16)の F 分布の確率密度関数

Excel には、F 分布の確率と F 値の関係を与える関数があります。自由度を n_1，n_2 として、図 8-4 の確率 p と F 値 x との関係は、以下で与えられます。

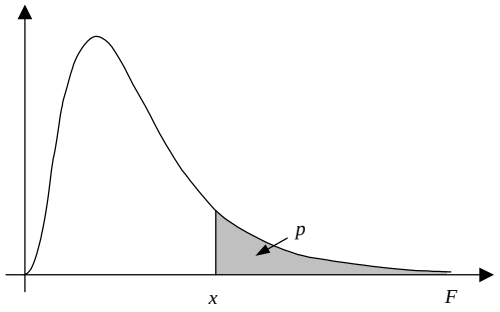

図 8-4 F 分布の確率

$$p = fdist(x, n_1, n_2), \quad x = finv(p, n_1, n_2)$$

例として、以下のような値を求めてみましょう。

自由度(8, 4)，F 値 10 のときの上側確率　　$fdist(10, 8, 4) = 0.020592$

自由度(10, 5)，上側確率 0.05 のときの F 値　　$finv(0.05, 10, 5) = 4.73505$

前節の最後に χ^2 分布は分散に関係した分布であるといいましたが、このことから、F 分布は分散の比に関係した分布であるといえます。

8.3 t 分布

特に平均値の推定や検定で利用される分布として、1 つの自由度を持った t 分布と呼ばれる分布があります。自由度 n の t 分布に従う確率変数 t は、標準正規分布に従う確率変数 X と自由度 n の χ^2 分布に従う確率変数 χ^2 から、$t = X\Big/\sqrt{\chi^2/n}$ のように表わされます。これをまとめて書くと以下のようになります。

$$X \sim N(0,1)\text{分布},\ \ \chi^2 \sim \chi_n^2\text{分布のとき、}\ t = \frac{X}{\sqrt{\chi^2/n}} \sim t_n\text{分布} \qquad (8.3)$$

さて、この変数 t の 2 乗、$t^2 = \dfrac{X^2}{\chi^2/n}$ については、X^2 が自由度 1 の χ^2 分布に従うことから、自由度 $1, n$ の F 分布に従うことが分かります。このように、標準正規分布、χ^2 分布、F 分布、t 分布には相互に関係があります。

自由度 1，2，4 の t 分布について、確率密度関数の具体的な形をグラフに表わすと図 8-5 のようになります。この形は正規分布によく似ていますが、t 分布は自由度が ∞ に近づくと標準正規分布になることが知られています。

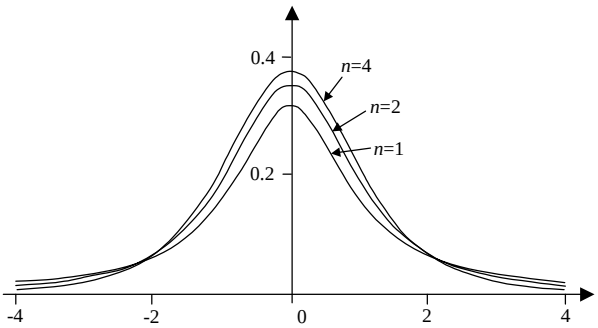

図 8-5　自由度 1, 2, 4 の t 分布の確率密度関数

他の分布と同様、Excel には t 分布の確率と t 値の関係を与える関数があります。自由度を n として、図 8-6 の確率 $p/2$ または p と t 値 x との関係は、以下で与えられます。但し、t 値から確率を求める場合、最後のパラメータによって上側確率 $p/2$ か、両側の確率の合計である p（両側確率）か区別されます。しかし、確率から t 値を求める場合は両側確率だけからしか計算できません。

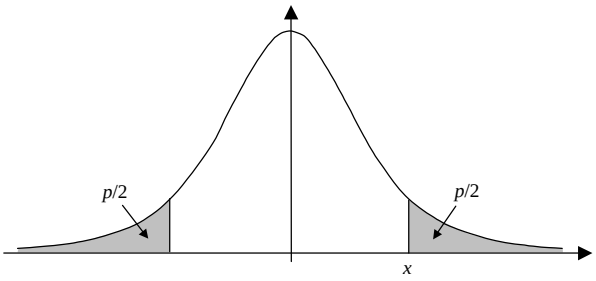

図 8-6　t 分布の確率

上側確率　$p/2 = tdist(x, n, 1)$,

両側確率　$p = tdist(x, n, 2)$,　$x = tinv(p, n)$

例として、以下のような値を求めてみましょう。

自由度 10，t 値 2 のときの上側確率 $p/2$　　$tdist(2, 10, 1) = 0.036694$

自由度 10，t 値 2 のときの両側確率 p　　$tdist(2, 10, 2) = 0.073388$

自由度 10，両側確率 0.05 のときの t 値　　$tinv(0.05, 10) = 2.228139$

最後の t 値は正の値を与えるようになっています。

この章で学んだ分布はこれからずっと利用します。関数名をできるだけ記憶して、確実に計算できるようになっておきましょう。

問題

以下の確率を求めよ。

1) $X \sim \chi_8^2$ 分布のとき、$P(X \geq 5.0)$　　2) $X \sim \chi_4^2$ 分布のとき、$P(X \geq 3.5)$

3) $X \sim F_{3,8}$ 分布のとき、$P(X \geq 3.2)$　　4) $X \sim F_{5,20}$ 分布のとき、$P(X \geq 1.3)$

5) $X \sim t_6$ 分布のとき、$P(X \geq 2.1)$　　6) $X \sim t_{10}$ 分布のとき、$P(|X| \geq 2.0)$

解答

1) 0.757576　　2) 0.477878　　3) 0.083668　　4) 0.303405

5) 0.040239　　6) 0.073388

8.4 分散と χ^2 分布

ここでは今後利用する分散の分布について考えてみます。まず独立な確率変数 X_i について、$X_i \sim N(\mu, \sigma^2)$ 分布を仮定します。これより以下の関係が得られることはすでに話しました。

$$X_i' = \frac{X_i - \mu}{\sigma} \sim N(0,1) \text{ 分布}$$

さらにこれから以下の関係も得ます。

$$\sum_{i=1}^{n} X_i'^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2 \text{ 分布} \qquad (8.4)$$

さて、これらの値は実際に計算できるのでしょうか。平均と分散の値は母集団の値ですので、標本から推測するしかありません。しかし、実測値を利用すると分布自身も変わってくることが知られています。例えば、平均 μ の代わりに標本平均 $\overline{X}$ を使うと、上式の自由度は１つ減ることが知られています。

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \text{ 分布} \qquad (8.5)$$

これは非常に重要な関係です。変数の組合せ $X_i - \overline{X}$ には、以下の制約が付きますので、自由度が自由に動ける変数の数ということであれば直感的には理解できると思います。

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

(8.5) の関係をきちんと見ようと思うと、一般の場合は少し厄介です。ここでは、$n = 2$ の場合について簡単に調べてみようと思います。互いに独立な確率変数 X_1, X_2 が以下の分布に従うとします。

$X_1 \sim N(\mu, \sigma^2)$ 分布, $X_2 \sim N(\mu, \sigma^2)$ 分布

標本平均は $\overline{X} = (X_1 + X_2)/2$ で与えられますので、$n = 2$ の場合の(8.5)式は以下のように変形できます。

$$\sum_{i=1}^{2} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{[X_1 - (X_1 + X_2)/2]^2}{\sigma^2} + \frac{[X_2 - (X_1 + X_2)/2]^2}{\sigma^2}$$

$$= \frac{(X_1 - X_2)^2}{2\sigma^2}$$

今、$Z_1 = (X_1 - X_2)/\sqrt{2}\sigma$ という新しい確率変数を考え、以下の正規分布の性質、

$X_i \sim N(\mu_i, \sigma_i^2)$ 分布ならば、

$$X = c_1 X_1 + c_2 X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2^2\sigma_2^2) \text{ 分布}$$

を利用すると、Z_1 は標準正規分布に従うことが分かります。それ故、χ^2 分布の定義から、上式は自由度 1 の χ^2 分布になります。

$$\sum_{i=1}^{2} \frac{(X_i - \overline{X})^2}{\sigma^2} = Z_1^2 \sim \chi_1^2 \text{ 分布}$$

ここでは確率変数 X_1, X_2 から線形変換で新たに１つの標準正規分布する確率変数 Z_1 が作り出せて、その変数だけで式が表わされることが重要な点です。一般には線形変

換で独立な $n-1$ 個の標準正規分布する変数 Z_i $(i=1,\cdots,n-1)$ が作り出されて、これを用いて上式は以下のように表わされます。

$$\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}=\sum_{i=1}^{n-1}Z_i^2 \sim \chi^2_{n-1}\text{分布}$$

問題

上の関係で $n=3$ の場合、新しい変数を以下のように定義すればよいことを確認せよ。

$$Z_1=\frac{1}{\sqrt{2}\sigma}(X_1-X_2),\quad Z_2=\frac{1}{\sqrt{6}\sigma}(X_1+X_2-2X_3)$$

解答

変数の定義と正規分布の性質から、

$$Z_1 \sim N((\mu-\mu)/\sqrt{2}\sigma,(1+1)\sigma^2/2\sigma^2)\text{分布}=N(0,1)\text{分布}$$

$$Z_2 \sim N((\mu+\mu-2\mu)/\sqrt{6}\sigma,(1+1+4)\sigma^2/6\sigma^2)\text{分布}=N(0,1)\text{分布}$$

であることが分かり、関係式と変数の定義を具体的に展開すると以下となる。

$$\sum_{i=1}^{3}\frac{(X_i-\overline{X})^2}{\sigma^2}=\frac{2}{3\sigma^2}(X_1^2+X_2^2+X_3^2-X_1X_2-X_2X_3-X_3X_1)$$

$$=Z_1^2+Z_2^2$$

最後に、Z_1 と Z_2 の独立性を見るために、共分散 $V(Z_1Z_2)=E(Z_1Z_2)$ を計算する。

$$2\sqrt{3}\sigma^2E(Z_1Z_2)$$

$$=E[(X_1-X_2)(X_1+X_2-2X_3)]$$

$$=E[((X_1-\mu)-(X_2-\mu))((X_1-\mu)+(X_2-\mu)-2(X_3-\mu))]$$

$$=E[(X_1-\mu)^2-(X_2-\mu)^2-2(X_1-\mu)(X_3-\mu)+2(X_2-\mu)(X_3-\mu)]$$

$$=\sigma^2-\sigma^2=0$$

共分散が 0 となるので、変数 Z_1 と Z_2 は独立であり、$Z_1^2+Z_2^2 \sim \chi_2^2$ 分布であることが示された。以上のことから、$n=3$ の場合に (8.5) の関係が成り立つことが分かる。