１５章 区間推定

母集団と標本の比較の検定では、母平均や母分散を既知として、標本から得られた値がどの程度の確率で実現されるかを計算し、有意水準と比較するものでした。これには検定統計量の確率分布が利用されました。区間推定は標本から得られた標本平均や標本分散から、この検定統計量の式を利用して、母平均や母分散がどの位の範囲に入るかを推定します。その際、場合によっては外れることもありますので、推定した範囲に入る確率で安全性を示しておく必要があります。この確率を信頼係数と呼び、区間推定はこの値を先に決めて範囲を指定する方法を取ります。通常の場合、信頼係数は 95%か 99%を用います。これらの信頼係数に基づく統計量の範囲を信頼区間と呼びます。

よく利用される区間推定には、母比率、母平均、母分散、母相関係数等の区間推定がありますが、ここでは前者 3 つの場合について見ていこうと思います。

15.1 母比率の区間推定

最初に質的データについて比率の推定の話をします。まず標本アンケート調査等で得られたある意見に対する賛成の比率から母集団の比率の区間推定を行ないます。例を見てみましょう。

例

ある制度についてのアンケート調査をランダムに抽出された 100 人に対して行ったところ、賛成 65 人、反対 35 人であった。母集団の賛成の比率を、信頼係数 95%（有意水準 5% に相当）で推定せよ。また、調査数 1000 人で同じ比率ではどうか。

理論

データ数 n、標本比率 $\hat{p}$ の標本から、母比率 p を信頼係数 $(1-\alpha)\times 100\%$ で推定する。

$z_0 = normsinv(1-\alpha/2)$ として、信頼区間は以下で与えられる。

$$\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 \le p \le \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 \qquad (15.1)$$

解答

データ数は $n=100$、標本比率は $\hat{p} = 65/100 = 0.65$、$\alpha = 0.05$ として以下を得ます。

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.047697, \quad z_0 = normsinv(0.975) = 1.959961$$

これを用いると比率の上限と下限（これを信頼限界といいます）は以下となります。

$$\hat{p}-\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}z_0 = 0.556516 \cong 0.557$$

$$\hat{p}+\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}z_0 = 0.743484 \cong 0.743$$

これから母比率の信頼区間は以下のようになります。

$$0.557 \le p \le 0.743$$

1000 人では、以下のように精度が上がることが分かります。

$$0.620 \cong 0.620438 \le p \le 0.679562 \cong 0.680$$

データ数が多いほど精度が上がるのは直感的に理解できると思います。

解説

ここではこの推定の理論を考えてみましょう。確率変数 X が出現確率 p と 1-p の 2 項分布に従う場合、試行回数を十分大きくすると以下のように正規分布に従うことを 10.2 節で話しました。

$$z = \frac{X-np}{\sqrt{np(1-p)}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \underset{n\to\infty}{\sim} N(0,1) \text{ 分布}$$

ここに、$\hat{p} = X/n$ で、これは標本比率を表わしています。このままでは後の計算が厄介になるので、分母の母比率を標本比率で置換え、以下のような近似を考えます。

$$z \cong \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

この量が近似的に標準正規分布に従うことから、信頼係数 $(1-\alpha)\times 100\%$ の信頼区間は、$z_0 = normsinv(1-\alpha/2)$ として、以下のようになります。

$$-z_0 \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_0$$

これを図で描くと図 15.1 のようになります。統計量 z の信頼区間は網掛けのある $-z_0$ から z_0 の間です。

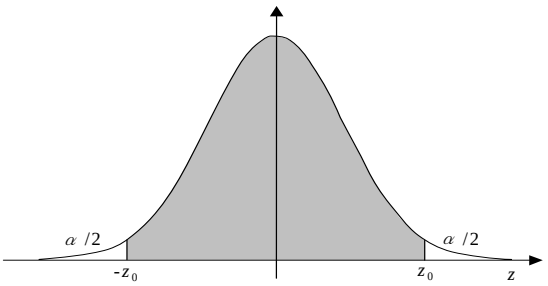

図 15.1 正規分布と信頼区間

p の信頼区間は、まず上式から分母を払い、

$$-\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\cdot z_0 \le \hat{p}-p \le \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\cdot z_0$$

以下の関係を得ます。

$$\hat{p}-\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\cdot z_0 \le p \le \hat{p}+\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\cdot z_0$$

問題

ある 500 人に対する調査で支持 205 人、不支持 295 人という結果を得た。母集団における支持の比率を信頼係数 95%で推定せよ。

解答

$$\hat{p}=0.41,\quad z_0=1.959961,\quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}z_0=0.04311$$

$$\hat{p}-\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}z_0=0.36689\cong 0.367,\quad \hat{p}+\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}z_0=0.45311\cong 0.453$$

以上より、信頼区間は $0.367 \le p \le 0.453$ となる。

15.2 正規母集団の母平均の区間推定

ここでは正規分布する量的データについて、標本平均から母平均を推定する方法を学びます。比率の場合と同様、推定は信頼係数による区間推定で行なわれます。例を見てみましょう。

例

ある標本データから所得について集計したところ以下の結果を得た。母集団は正規分布するとして母平均を信頼係数 95%で推定せよ。

データ数 30，平均 620，標準偏差 90

また、データ数を 100 にすると結果はどう変わるか？

理論

正規分布する母集団から得られた標本より、母平均 μ を信頼係数 $(1-\alpha)\times 100\%$ で推定する。データ数を n，標本平均を $\bar{x}$，不偏分散を u^2，$t_0=tinv(\alpha, n-1)$ として、信頼区間は以下で与えられる。

$$\bar{x} - \frac{u}{\sqrt{n}} t_0 \le \mu \le \bar{x} + \frac{u}{\sqrt{n}} t_0 \qquad (15.2)$$

解答

データから $n = 30$, $\bar{x} = 620$, $u = 90$ となり、信頼係数 95%で t_0 は以下となります。

$$t_0 = tinv(0.05, 29) = 2.045231$$

これらを用いると、次のようになり、

$$\frac{u}{\sqrt{n}} t_0 = 33.60657, \quad \bar{x} - \frac{u}{\sqrt{n}} t_0 = 586.3934 \cong 586, \quad \bar{x} + \frac{u}{\sqrt{n}} t_0 = 653.6066 \cong 654$$

母平均の信頼区間は以下のようになります。

$$586 \le \mu \le 654$$

データ数を 100 にすると、以下のように精度が向上します。

$$602 \cong 602.142 \le \mu \le 637.858 \cong 638$$

解説

この区間推定には、11. 3 節で述べた以下の検定統計量の性質を利用します。

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{u} \sim t_{n-1} \text{分布}$$

これから信頼係数 $(1-\alpha)\times 100\%$ の信頼区間は、$t_0 = tinv(\alpha, n-1)$ として以下の ように与えられます。

$$-t_0 \le \frac{\sqrt{n}(\bar{x} - \mu)}{u} \le t_0$$

これを図に描くと図 15.2 のようになります。統計量 t の信頼区間は $-t_0$ から t_0 の網掛けのある区間です。

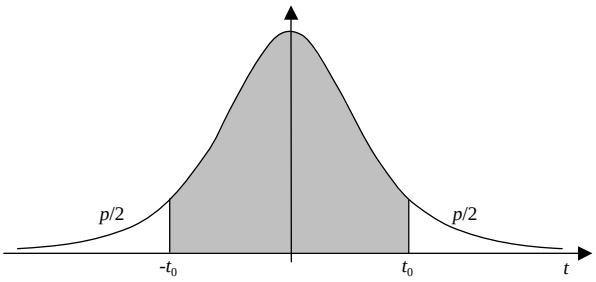

図 15.2 t 分布と信頼区間

上の関係から次のようになり、

$$-\frac{u}{\sqrt{n}} t_0 \le \bar{x} - \mu \le \frac{u}{\sqrt{n}} t_0$$

最終的に以下の信頼区間が求められます。

$$\bar{x} - \frac{u}{\sqrt{n}} t_0 \le \mu \le \bar{x} + \frac{u}{\sqrt{n}} t_0$$

問題

正規分布を仮定して、以下の身長データ(cm)から母平均を信頼係数 95%で推定せよ。

184, 170, 164, 176, 177, 170, 171, 159, 174, 170,

165, 170, 171, 183, 175, 169, 181, 172, 171, 164

解答

$n = 20$, $\bar{x} = 171.8$, $u = 6.379243$, $t_0 = 2.0930$, $\frac{u}{\sqrt{n}} t_0 = 2.985578$ より、

$$\bar{x} - \frac{u}{\sqrt{n}} t_0 = 168.8144 \cong 168.8, \quad \bar{x} + \frac{u}{\sqrt{n}} t_0 = 174.7856 \cong 174.8$$

以上から信頼区間は$168.8144 \le \mu \le 174.7856$ となる。

15.3 正規母集団の母分散の区間推定

ここでは標本の不偏分散から母分散を推定する問題を考えます。母分散の区間推定は分布が χ^2 分布であることから、信頼区間が左右対称ではありません。実際に例を見てみましょう。

例

ある標本データから所得について集計したところ以下のデータを得た。母集団は正規分布するとして母分散を信頼係数 95%で推定せよ。

データ数 30, 平均 620, 不偏分散 8100

理論

正規分布する母集団から得られた標本より、母分散σ^2を信頼係数$(1-\alpha)\times 100\%$ で推定する。

データ数n, 不偏分散u^2, $x_1 = chiinv(1-\alpha/2, n-1)$, $x_2 = chiinv(\alpha/2, n-1)$ として、信頼区間は以下で与えられる。

$$\frac{(n-1)u^2}{x_2} \le \sigma^2 \le \frac{(n-1)u^2}{x_1} \qquad (15.3)$$

解答

データから、それぞれの量は以下のようになります。

$n = 30$, $u^2 = 8100$

$$x_1 = chiinv(0.975, 29) = 16.04705$$
$$x_2 = chiinv(0.025, 29) = 45.72228$$

これを用いて

$$\frac{(n-1)u^2}{x_2} = 5137.539 \cong 5140\ ,\quad \frac{(n-1)u^2}{x_1} = 14638.2 \cong 14640$$

となり、以下の信頼区間を得ます。

$$5140 \le \sigma^2 \le 14640$$

解説

この区間推定には 8.4 節で述べた分散に関する次の性質を利用します。

$$\chi^2 = \frac{(n-1)u^2}{\sigma^2} \sim \chi^2_{n-1} \text{ 分布}$$

統計量 χ^2 について、信頼係数 $(1-\alpha)\times 100\%$ の信頼区間は、下限と上限をそれぞれ、$x_1 = chiinv(1-\alpha/2, n-1)$，$x_2 = chiinv(\alpha/2, n-1)$ として以下のように与えられます。

$$x_1 \le \frac{(n-1)u^2}{\sigma^2} \le x_2$$

これを図で表わすと図 15.3 のようになります。

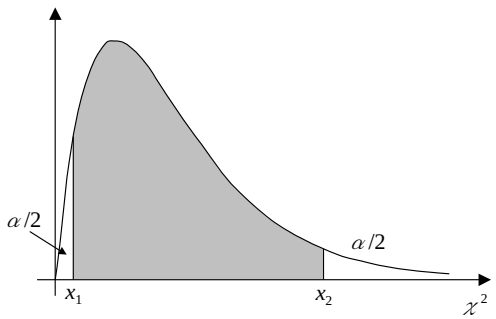

図 15.3 χ^2 分布と信頼区間

これから、分母と分子を逆にし、以下の信頼区間を得ます。

$$\frac{(n-1)u^2}{x_2} \le \sigma^2 \le \frac{(n-1)u^2}{x_1}$$

問題

身長(cm)についての以下の標本データを用いて、母分散を信頼係数 95%で推定せよ。

184, 170, 164, 176, 177, 170, 171, 159, 174, 170,

165, 170, 171, 183, 175, 169, 181, 172, 171, 164　　（データ数 20）

解答

$n = 20$, $u^2 = 40.69474$,

$x_1 = chiinv(0.975,19) = 8.90651$

$x_2 = chiinv(0.025,19) = 32.85234$

$\dfrac{(n-1)u^2}{x_2} = 23.53562 \cong 23.54$, $\dfrac{(n-1)u^2}{x_1} = 86.81286 \cong 86.81$

以上から信頼区間は$23.54 \le \sigma^2 \le 86.81$となる。

１６章 回帰分析

5.4 節で 2 変量についての散布図を描き、直線状に並んでいる度合いで相関係数の値が決まるという話をしました。この章ではデータの並びを近似するこの直線について学びます。この直線は回帰直線と呼ばれ、散布データに最も適合するように引かれています。また回帰直線を表わす回帰式は $y = ax + b$ のように示されますが、この式に意味があるのかどうかという検定や、係数 a, b の値が 0 と異なるかどうかという検定も行なわれます。この 2 変数間の関係を 1 次式のモデルとして考える分析は回帰分析と呼ばれ、様々な分野で頻繁に利用されています。それでは例を見てみましょう。

例

下の表のデータを用いて、身長により体重を推定する式を考えよ。ただし、式は 1 次式（体重 $= a \times$ 身長 $+ b$）と仮定し、その有効性を検討せよ。

身長	169	175	170	179	176	174	173	181	179	178
体重	71	68	67	72	69	80	75	65	74	71
身長	170	180	177	175	172	166	168	173	169	170
体重	62	75	70	70	62	58	60	58	59	73

理論

回帰式の決定

2 変数の関係を、 $y = ax + b$ の直線で表わし、 x を説明変数、 y を目的変数と呼ぶ。図 16.1 のようにデータ点からこの直線へ垂直におろした線の長さの 2 乗の合計が最小となるように回帰係数 a, b を決める。2 変数について、平均 $\bar{x}, \bar{y}$, 標準偏差 u_x, u_y, 相関係数 r とすると回帰係数は以下のように表わされる。

$$a = r\frac{u_y}{u_x}, \quad b = \bar{y} - r\frac{u_y}{u_x}\bar{x} \tag{16.1}$$

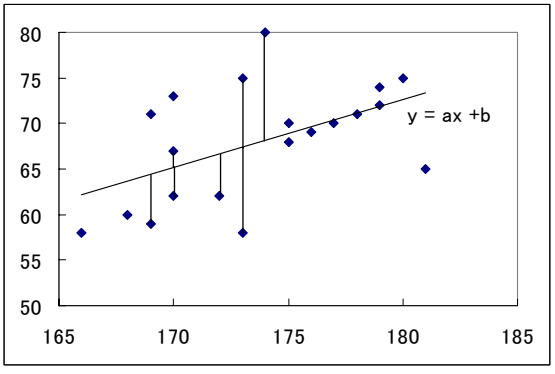

図 16.1 回帰直線とデータとの距離

回帰式の有効性の検討

相関係数 r	目的変数と説明変数の相関係数であると同時に、目的変数の実測値と回帰式による予測値の相関係数
寄与率（決定係数） r^2	目的変数の変動のうち回帰式が説明する割合
回帰式の有効性の検定	回帰式は無意味と考えられる確率で検討する。

解答

データから、以下の集計結果を得ます。

$$\bar{x} = 173.7,\quad \bar{y} = 67.95$$

$$u_x = 4.402153,\quad u_y = 6.378211,\quad r = 0.513047$$

これを用いて回帰係数及び回帰式を求めると、以下となります。

回帰係数　$a = 0.743346 \cong 0.743,\quad b = -61.1692 \cong -61.2$

回帰式　$y = 0.743 \cdot x - 61.2$

また、相関係数と寄与率は以下となります。

相関係数　$r = 0.513047 \cong 0.513$

寄与率　$r^2 = 0.263217 \cong 0.263$

回帰式が説明する割合が 26%くらいですから、余り良い近似とは言えないようです。

解説

一般に目的変数を説明変数で予測する分析は、複数の説明変数を用いることが多く、ここで述べた 1 つの説明変数の場合はむしろ特殊です。このように複数の説明変数の 1 次式で目的変数を予測する分析は重回帰分析と呼ばれ、その中で、説明変数が 1 つの場合を特に回帰分析と呼んでいます。重回帰分析では、目的変数と回帰式による予測値の相関係数を重相関係数、その 2 乗で、目的変数の変動のうち回帰式が説明する部分を寄与率または重決定係数と呼びます。

回帰分析を行うにはここで述べたように計算する他、Excel に含まれている分析ツールを利用することもできます。これを用いると上で与えた統計量以外に、回帰式の有効性の検定や回帰係数の値が 0 か否かの検定も行ってくれます。具体的な実行画面が以下の表です。ここでは重相関や重決定という言葉が使われていますが、このツールは重回帰分析にまで対応しているため、このような表現になっています。

統計の初心者が見る部分は網掛けの部分でよいと思います。特に、有意 F の部分はこの回帰分析が有効であるか否かの検定確率で、値が有意水準より小さいと有効と判断されます。回帰係数の右の方にある P-値は、それぞれの回帰係数の値が 0 か否かの検定確率値です。値が有意水準より小さい場合、0 と異なると判定されます。回帰式で

説明変数の前の係数が 0 の場合、回帰分析自体が意味のないものになってしまいますから、この係数の検定と回帰分析の有効性の検定は同じものであり、検定確率値も 0.020703 と同じ値になっています。一般の重回帰分析ではこのようなことはありません。

表 6.1 Excel の分析ツールを用いた解答例

回帰統計	
重相関 R	0.513047
重決定 R2	0.263217
補正 R2	0.222285
標準誤差	5.624827
観測数	20

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	203.4538	203.4538	6.430541	0.020703
残差	18	569.4962	31.63868		
合計	19	772.95			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-61.1692	50.93303	-1.20097	0.245327	-168.176	45.83721
X 値 1	0.743346	0.293135	2.535851	0.020703	0.127492	1.3592

注）標準誤差：線形回帰式における予測値と実測値とのずれの標準偏差

数学的解説　[Skip OK]

ここでは、回帰式を導いておきましょう。データの個数を n とし、説明変数と目的変数のデータの組を (x_i, y_i) として、実測値 y_i を以下の 1 次式で予測します。

$$Y_i = ax_i + b$$

実測値と予測値の差（図 16.1 で縦線の部分）の 2 乗を S とし、これを最小化します。

$$S = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

そのためにまず S を係数 a と b で微分（偏微分）して 0 と置き、以下の関係を得ます。

$$2\sum_{\lambda=1}^{n} x_i(ax_i + b - y_i) = 0, \quad 2\sum_{i=1}^{n} (ax_i + b - y_i) = 0$$

ここで、x_i と y_i の平均 $\bar{x}$，$\bar{y}$ を用いると、2 番目の式は以下となります。

$$a\bar{x} + b - \bar{y} = 0$$

この式から得られる b を最初の式に代入すると、

$$\sum_{i=1}^{n} x_i[a(x_i-\bar{x})-(y_i-\bar{y})]=0$$

のようになります。さらに、

$$\sum_{i=1}^{n} \bar{x}[a(x_i-\bar{x})-(y_i-\bar{y})]=0$$

の関係を用いると上の式は以下のように変形されます。

$$\sum_{i=1}^{n} (x_i-\bar{x})[a(x_i-\bar{x})-(y_i-\bar{y})]=0$$

x_i と y_i の標準偏差を u_x, u_y、共分散を u_{xy} とするとこの式は

$$au_x^2-u_{xy}=0$$

と表わされ、これから a についての関係を得て、b の値も以下のように求められます。

$$a=\frac{u_{xy}}{u_x^2}=\frac{u_y}{u_x}\cdot\frac{u_{xy}}{u_x u_y}=\frac{u_y}{u_x}r\ ,\qquad b=-a\bar{x}+\bar{y}=-\frac{u_y}{u_x}r\cdot\bar{x}+\bar{y}$$