１１章 母集団と指定値との量的データの検定

11.1 検定手順

前章で質的データの検定手法について説明しましたので、ここからは量的データの検定について話します。量的データの検定は少し分量が多くなりますので、「母集団と指定値との検定」、「対応のない２群間の検定」、「対応のある２群間の検定」と３つに章を分けて話を進めることにします。ここでは、母集団と指定値との検定について説明します。

例えば全国平均が分かっている場合で、ある地域の標本と全国平均を比較するような場合や、理論的に与えた結果を実験結果と比較する場合等がこれに当たります。この検定方法は分布に正規性があるかどうかによって、図 11-1 のように２つに分かれます。

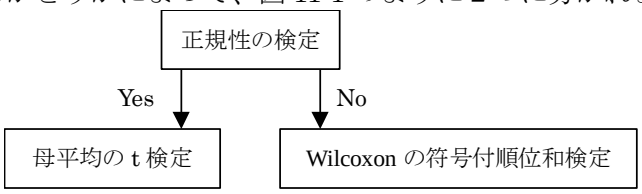

図 11-1 母集団と指定値との量的データの検定手法

そのために、まず得られたデータが正規分布に従うかどうか調べてみる必要があります。そこでこの章では最初にこの正規性の調べ方について説明します。その後、これらの検定手法について解説します。また最後に、母平均と指定値との比較の問題で、有意差を得るために必要なデータ数の求め方について簡単に触れてみます。

11.2 正規性の検定

データが正規分布しているかどうか調べる方法として、4.2 節でヒストグラムを描く方法を学びましたが、これはある程度データ数が多くないと使えません。それではデータ数が少ない場合はどうするのでしょうか。この場合には統計処理用に作られた正規確率紙というものを利用する方法があります。しかし、この正規確率紙を手に入れるのは少々厄介ですので、これに変わる方法を Excel で考えてみます。原理は正規確率紙と同じです。では具体的に例を用いて説明します。

例

以下のデータの正規性を調べよ。

2.5, 2.1, 3.4, 2.8, 4.6, 3.2, 3.8, 4.8, 4.0

解答

Excel を用いた視覚的方法について順を追って説明します。

1) データを入力する（データ数を n とする）。
2) データを小さい順に並べ替える。

これは範囲を指定し、メニュー［データ－並べ替え］で昇順に並べ替えます。

3) データに 1 から番号を振る。

データの左側に 1 から順番にデータの末尾まで数字を振ります。

4) 累積比率を求める。　$p_i = \dfrac{i}{n+1}$　i は番号

先ほど入力した、番号を使って累積比率を計算し、データの横に入力します。

5) 関数 $z_i = normsinv(p_i)$ を用いて座標値 z_i を求める。

累積比率 p_i を用いて、これに相当する正規分布の座標値 z_i を求めます。ここで座標値 z_i と下側確率 p_i の関係は以下の図のようになります。

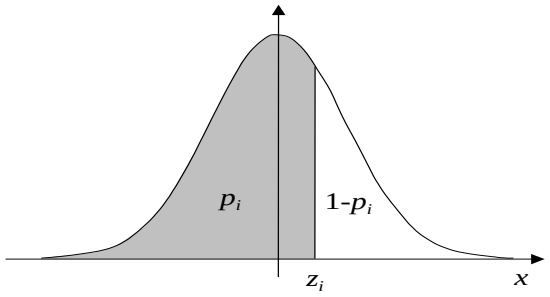

図 11-1 正規分布と確率

$$p_i = normsdist(z_i),\ z_i = normsinv(p_i)$$

6) データと座標値を用いて散布図を描く。

データ x_i（横軸）と上の座標値 z_i（縦軸）を用いて、2 次元の散布図を描きます。

7) グラフに近似直線を加える。

グラフにメニュー［グラフ－近似曲線の追加］を用いて近似直線を加えます。

8) 直線に近く並んでいるようなら正規分布

この直線の近傍に点が散らばっているようなら、正規分布とみなされます。

表 11-1 正規確率紙の方法

番号	データ	累積比率	x 値
1	2.1	0.1	−1.28155
2	2.5	0.2	−0.84162
3	2.8	0.3	−0.52440
4	3.2	0.4	−0.25335
5	3.4	0.5	0.00000
6	3.8	0.6	0.25335
7	4.0	0.7	0.52440
8	4.6	0.8	0.84162
9	4.8	0.9	1.28155

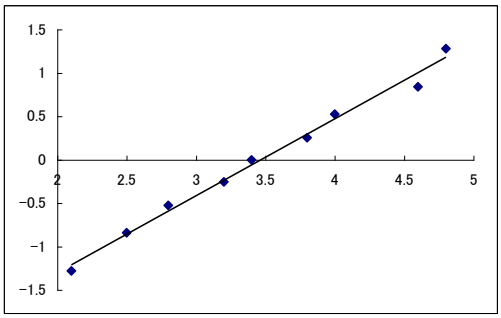

図 11-2 正規確率紙の方法

この例題の場合、データが直線状に並んでいると認められるので、正規分布とみなせます。

解説

ここではなぜこのようなやり方で正規性が示されるのか考えてみます。今確率変数 X が $N(\mu,\sigma^2)$ 分布であるとします。1 つのデータ値を x として、$X \le x$ である確率 p は、

$p = normsdist((x-\mu)/\sigma)$ のように表されます。ここに、変数 $(X-\mu)/\sigma$ は標準正規分布に従い、$normsdist(\ )$は標準正規分布の下側確率を求める Excel 関数でした。この確率は近似的にデータ数で見た累積比率に等しいと考えてみます。

$$p = normsdist\left(\frac{x-\mu}{\sigma}\right) \cong \frac{i}{n+1}$$

ここに n はデータの個数、i は小さいほうから数えたデータ x の番号です。

右辺の近似式から、逆に標準正規分布の座標値を求めて $z = normsinv(i/(n+1))$ とすると、以下のように z は近似的に x の 1 次関数となります。

$$z = normsinv\left(\frac{i}{n+1}\right) \cong \frac{x-\mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$$

このようにデータが正規分布に従うならば、上の z を y 軸に、x を x 軸にして散布図を描くと、データは直線状に並ぶはずです。もし、この直線から外れるような場合があれば、これはデータの正規性に問題があるということです。しかし、確率を $i/(n+1)$ で近似していますので、完全に直線状に並ぶという訳にもいきません。大体直線に並ぶという微妙な基準しかありません。

上の方法は直線に並んでいるという直感的な感覚が頼りでしたから、当然人によって判断基準も変わってきます。そこでこれをはっきりさせるために数値的な方法も考案されています。代表的な方法には、コルモゴロフースミルノフ（Kolmogorov-Smirnov）の正規性の検定やシャピローウィルク（Shapiro-Wilk）の W 統計量を用いた方法等があります。後者の方法で、例題について、正規分布と考えられる確率を求めてみると p < 0.9147（統計ソフト statistica による）となります。また以下の問題にも参考のためにこの確率の値を付記しておきます。グラフを見た場合の基準にしてもらえればと思います。

問題

以下のデータの正規性を調べよ。

507, 491, 421, 493, 415, 640, 464, 602, 530, 395

解答

表 11-2 正規確率紙の方法

番号	データ	累積比率	x 値
1	395	0.090909	-1.33518
2	415	0.181818	-0.90846
3	421	0.272727	-0.60458
4	464	0.363636	-0.34876
5	491	0.454545	-0.11418
6	493	0.545455	0.11419
7	507	0.636364	0.34876
8	530	0.727273	0.60458
9	602	0.818182	0.90846
10	640	0.909091	1.33518

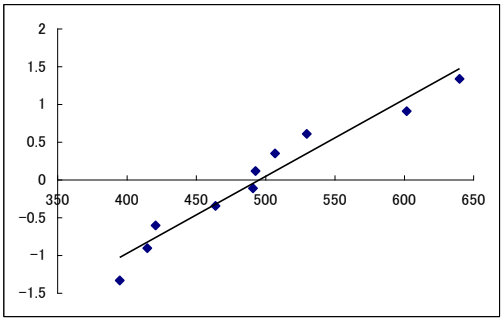

図 11-3 正規確率紙の方法

この場合、ほぼ正規分布とみなせる。(p < 0.5515)

問題

以下のデータの正規性を調べよ。

20.9, 61.1, 57.2, 51.0, 46.6, 41.2, 21.0, 56.3, 49.5, 49.3, 22.4, 23.5

解答

表 11-3 正規確率紙の方法

番号	データ	累積比率	x 値
1	20.9	0.076923	-1.42608
2	21.0	0.153846	-1.02008
3	22.4	0.230769	-0.73632
4	23.5	0.307692	-0.50240
5	41.2	0.384615	-0.29338
6	46.6	0.461538	-0.09656
7	49.3	0.538462	0.09656
8	49.5	0.615385	0.29338
9	51.0	0.692308	0.50240
10	56.3	0.769231	0.73632
11	57.2	0.846154	1.02008
12	61.1	0.923077	1.42608

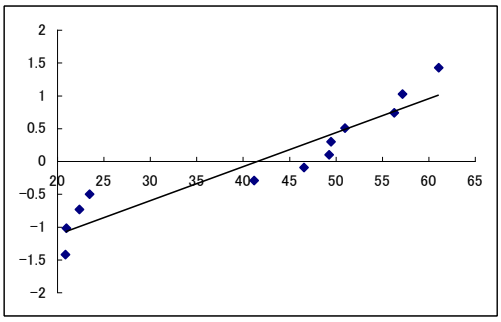

図 11-4 正規確率紙の方法

直線状に並んでいると言えないので、正規分布とは言えない。($p < 0.0392$)

11.3 母平均と指定値との比較（正規性あり）

前節で述べた方法は視覚的に正規性を調べる方法ですから､その結果になかなか自信が持てないと思います。そこでよく使われるのが名前だけ紹介した数値的方法です。統計ソフト等でこの方法を用いると、データの分布が正規分布と異なることは容易に示せます。しかしその逆は「このデータでは正規分布と異なるといえない」というだけで、積極的に正規性を支持するものではありません。ただ通常この「正規分布と異なるといえない」ということで不本意ながら正規性が示されたと解釈することが多いようです。この場合、正規性に少しでも怪しいところがあれば、次の節で述べる Wilcoxon の符号付き順位和検定と併用すればよいと思います。これは分布型を問わない検定方法ですので、正規分布でもそうでなくても利用できます。ここでは、正規性を認めて検定手法の説明をしましょう。

例

ある会社 20 社のある商品の従業員 1 人当り売上高のデータを集めたら､平均 241（万円）、不偏分散から求めた標準偏差 14（万円）であった。これらの会社の売上高は 226（万円）に比べて差があるといえるか。正規分布を仮定し、有意水準 5%で判定せよ。

理論　母平均の t 検定

正規分布する標本について、標本の母平均 μ_1 と指定値 μ とを比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で判定する。但し、データ数 n、標本平均 $\bar{x}$、不偏分散 u^2 とする。

帰無仮説 H_0： $\mu_1 = \mu$　　（平均に差がない）

対立仮説 H_1： $\mu_1 \neq \mu$　　（平均に差がある，両側検定）

H_0 のもとで　　$t=\dfrac{\sqrt{n}(\bar{x}-\mu)}{u}\sim t_{n-1}$ 分布　　(11.1)

$p=tdist(t,n-1,2)$ として、 $p<\alpha$ のとき、H_0 を棄却して H_1 を採択する。

解答

$n=20,\ \bar{x}=241,\ \mu=226,\ u=14$ として、統計量 t を求めると以下のようになります。

$$t=\frac{\sqrt{20}(241-226)}{14}=4.791574$$

自由度は $20-1=19$ より、検定確率値は $tdist()$ 関数を用いて以下のようになります。

$$p=tdist(4.791574,8,2)=0.000127\cong 0.0001$$

$p<0.05$ より、１人当りの売上高に差があるといえると判定されます。

解説

t 分布の座標値から確率を求める Excel 関数は、座標値 t, 確率 p, 自由度 d として、以下のように与えられています。確率値を求める場合、最後のパラメータで両側確率か、片側確率かを指定します。

$p=tdist(t,d,2)$　　両側検定

$p/2=tdist(t,d,1)$　　片側検定

$t=tinv(p,d)$　　両側検定

量的なデータの指定値との比較の問題は、データが正規分布する場合とそうでない場合とで取り扱い方が違うことは以前説明しました。ではなぜ正規分布する場合だけ特別に t 検定を利用するのでしょうか。一般的な方法があれば、どちらの場合もそれを用いればよいはずです。その理由は正規分布するデータでは、一般的な方法に比べて、t 検定がより差を見出し易いからです。但し、t 検定は正規分布からずれると全く意味のないものになってしまいますので、十分注意して使用する必要があります。

さて、データの範囲が広く小さい方に多く集まっているような場合、データの対数を取ると正規分布に近い分布を得ることがあります。図 11.1a はデータをそのまま利用したヒストグラムですが、図 11.1b は自然対数（底が e の対数）を取ったデータを用いたヒストグラムです。前者は正規分布から相当外れていますが、後者は正規分布に近い形をしています。この場合一般の検定を利用することも考えられますが、対数を取って正規分布にして検定を行った方が、良い結果が得られます。このように対数を取ったデータが正規分布するような分布を対数正規分布といいます。

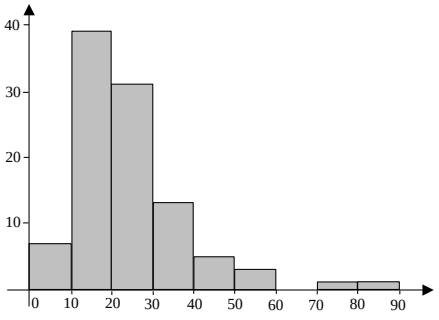

図 11.1a 度数分布

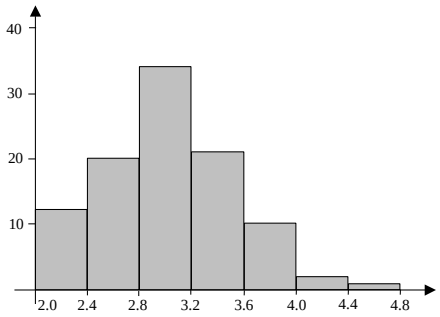

図 11.1b 自然対数を取った度数分布

数学的解説

ここでは統計量 t が t 分布に従うことを簡単に示しておきましょう。確率変数 X_i が独立で $X_i \sim N(\mu,\sigma^2)$ 分布とすると、平均は

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(\mu, \sigma^2/n) \text{ 分布}$$

となることを 7.3.4 節の問題で示しましたが、これを用いると $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 分布となることが分かります。ここでは指定値 μ の値は分かっていますが、σ の値は分かりません。そこでこれを不偏分散 u^2 から求めた標準偏差 u で代替します。不偏分散には 8.4 節で述べたように、以下の関係があることが知られています。

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

そこで 8.3 節で述べた t 分布の定義から、以下のような関係が分かります。

$$\frac{\sqrt{n}(\overline{X}-\mu)}{u} = \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\dfrac{\sqrt{n-1}\,u}{\sigma}\Big/\sqrt{n-1}} \sim t_{n-1} \text{ 分布}$$

この式では σ の値が u に置き換わり、左辺はすべてデータから求まります。

問題

以下のデータの正規性が認められているとき、平均は 5.5 と比べて差があるといえるか。有意水準 5%で判定せよ。

8.4, 4.6, 5.2, 6.3, 7.2, 5.8, 6.0, 5.4, 4.9, 6.9

解答

$n = 10$，$\mu = 5.5$，$\bar{x} = 6.07$，$u = 1.16719$

$$t = \frac{\sqrt{10}(6.07-5.5)}{1.16719} = 1.544305$$

$$p = tdist(1.544305, 9) = 0.156912 \cong 0.157$$

$p > 0.05$ より、差があるとは言えない。

11.4 母集団の中央値と指定値との比較（正規性なし）

データに正規性が見られないとき、上記の t 検定は使えず、データの分布によらない検定手法を利用します。このような検定を総称してノンパラメトリック検定といいます。これに対して前節の t 検定のように、正規性を利用する検定をパラメトリック検定と呼びます。データの正規性に少しでも不安がある場合､我々は両方の手法を併用することをお勧めします。正規性がある場合、ノンパラメトリック検定は使えないのではなく､パラメトリック検定の方がより明確に差が出るというだけです。しかし、逆に正規性が認められない場合、パラメトリック検定の結果は何の意味も持ちません。

ノンパラメトリック検定では､何らかの形でデータに順位を付け､その順位和を用いて検定を行う場合が多く見られます。この教科書で登場するものとしては、この節と 13.2 節で説明する Wilcoxon の符号付き順位和検定､12.4 節で学ぶ Wilcoxon の順位和検定､及び 15.2 節の Spearman の順位相関係数等が、代表的なノンパラメトリックな手法です。これらの理論についてはかなり深い数学的背景があり､この教科書の範囲外ですので利用法のみをまとめて解説します。

ここで、章の見出しに中央値の検定と書いていますが、これはデータ数の多いとき、後に述べる順位和が正規分布に従うことを使った平均値の検定になります｡順位和の分布の平均値は中央値に相当しますので中央値の検定としています。

例

ある会社のある商品の 1 人当り売上高（万円）は以下の通りである。これらの会社の売上高は 226（万円）に比べて差があるといえるか。有意水準 5%で判定せよ。

206, 235, 155, 172, 180, 199, 151, 172, 291, 182, 260

理論　Wilcoxon の符号付き順位和検定

標本データ x_i の中央値 m' と指定値 m を比較し、差があるかどうか有意水準 $\alpha\times100\%$ で判定する。

帰無仮説 H_0： $m'=m$　　中央値に差がない

対立仮説 H_1： $m'\neq m$（両側検定）　　中央値に差がある

新しい変数 $z_i=X_i-m$ を考える。$|z_i|$ の小さい順に 0 を除いて順位 r_i を付け、$z_i=0$ の場合を除いて z_i の正負で 2 群に分ける。但し、同数値の場合は、順位平均を取る。例えば、5 位が 2 つの場合は、両方 (5+6)/2=5.5 とする。

各群のデータ数を r, s（$n=r+s$）、順位和を R_r, R_s とし、小さい方の順位和を R とする。

データ数が少ない（$n\le50$）とき

補遺 3 の数表を参照し､両側確率を α として $R\le R_1$ のとき､H_0 を棄却して H_1 を採択する。

データ数が多い（$n>50$）とき

H_0 のもとで　　$z=\dfrac{|R-n(n+1)/4|-1/2}{\sqrt{n(n+1)(2n+1)/24}}$ ～ $N(0,1)$ 分布（正の部分）　　(11.2)

$p=2\cdot(1-normsdist(z))$ として、 $p<\alpha$ のとき、H_0 を棄却して H_1 を採択する。

解答

まず以下のような表を作ります。順位は Excel の rank(数値,範囲,1)関数を用いて昇順に付けます。このとき同順位は同じ数字ですから、これを平均順位に置き換えて訂正順位とします。同じ順位を見つけるにはメニュー［データー並べ替え］を用いると便利です。

データ	差	|差|	順位	訂正順位
206	-20	20	2	2
235	9	9	1	1
155	-71	71	10	10
172	-54	54	7	7.5
180	-46	46	6	6
199	-27	27	3	3
151	-75	75	11	11
172	-54	54	7	7.5
291	65	65	9	9
182	-44	44	5	5
260	34	34	4	4

この表から、データと指定値との差が正のものと負のものに分けて順位和を求めます。訂正順位で四角で囲んだものは差が正になるものです。この順位合計を求める際にも並べ替えを用いると簡単です。結果は差が正になる群が 14、負になる群が 52 となります。２つの順位和から小さい方を選んで $R=14$ とします。

補遺３の数表から $n=11$ で $\alpha=0.05$ の値 $R_1=10$ を求めて、以下のような結論になります。

$R>R_1$　より、中央値に差があるとはいえない。

解説

Wilcoxon の符号付き順位和検定には、2 種類の方法があります。１つはここで述べたように、データの値から指定値を引いて絶対値をとり順位を付ける方法、もう１つは２つの対応するデータ間で引き算を行なって絶対値をとり順位を付ける方法です。これら２つの方法とも、求めた差の正負で群を分けて順位和を求める方法は同じです。後者については対応のあるデータに対する Wilcoxon の符号付き順位和検定として、13.3 節で学びます。

手順をまとめておくと、まず個々のデータから指定値を引き、その絶対値をとります。絶対値の小さい順に順位を付け、データから指定値を引いた差の値が正のものと負のものとで２つの群に分け、それぞれの順位の合計を取ります。例えば、4 位が２つあるような同順位のものについては、２つを(4+5)/2=4.5 位とします。また、データから中央値を引いた差が 0 のデータは除外します。２つの群のうち、順位合計の小さいものを選び、その値によって検定しますが、データ数が少ない場合は表によって、データ数が多い場合は検定量 z を求め、それが標準正規分布に従うことを利用して検定を行ないます。

もう少し分り易く言い換えると、この検定方法は指定値に近いデータから順に順位を付け、指定値より小さい側と大きい側で順位合計を取るものです。分布が指定値より大きい側に偏っていればいるほど、大きい側の順位合計は大きくなります。このとき指定値に近いところから順位を付けていますので、この傾向はより顕著に効いてきます。これがこの検定のうまいところです。

11.5 母平均推定のためのデータ数の決定

最後に少し本筋から離れて、調査等をするときにデータ数をどのように決めるのかという問題について考えてみようと思います。10.4 節で母比率と指定比率の比較の場合について述べましたが、ここでは正規分布するデータで指定値との比較の問題に絞ってその考え方を学びます。

例

母集団の標準偏差が 5cm であるとき、標本平均 169cm として指定値 170cm と異なることを有意水準 5%で示すためには、いくらのデータ数が必要か。

理論

指定値が μ、母分散 σ^2 の場合、有意水準 $\alpha\times100\%$ で、標本平均 $\bar{x}$ から推測される母平均が指定値と等しくないことを判定するために必要なデータ数を求める。但し、検定は両側検定とする。

$$Z=\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\sim N(0,1)\text{ 分布}\quad\text{を用いて、}$$

$$n>\frac{Z(\alpha/2)^2\sigma^2}{(\bar{x}-\mu)^2}=\frac{normsinv(1-\alpha/2)^2\sigma^2}{(\bar{x}-\mu)^2}\qquad(11.3)$$

注） $Z(\alpha/2)$ は標準正規分布上側確率 $\alpha/2$ の座標値である。

Excel でこれは $normsinv(1-\alpha/2)$ と表示される。

解答

$$n>\frac{normsinv(0.975)^2\times5^2}{1^2}=96.03619\quad\text{より、標本は 97 以上必要である。}$$

解説

11.3 節では、データ数と指定値、標本平均、不偏分散から観測値の出現確率を求めましたが、今回は指定値、標本平均、母分散の値と観測値の出現確率（有意水準の値）からデータ数を求めています。検定の場合とデータ数の決定の場合とで分散が、不偏分散と母分散で異なっていますが、母分散が分からない場合は、近似的に不偏分散で代用しても大きな問題はありません。

検定で有意差が出るためには、統計値が $Z(\alpha/2)$ より大きいことが条件です。

$$Z(\alpha/2)<\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\quad\text{を変えて}\quad\sqrt{n}>\frac{Z(\alpha/2)\sigma}{\bar{x}-\mu}$$

両辺の 2 乗をとって、以下の式を得ます。

$$n > \frac{Z(\alpha/2)^2\sigma^2}{(\bar{x}-\mu)^2}$$

ここで、不偏分散ではなく母分散を使った理由は、$\frac{\sqrt{n}(\bar{x}-\mu)}{u} \sim t_{n-1}$ 分布の関係から (11.3) に相当する式が

$$n > \frac{tinv(\alpha, n-1)u^2}{(\bar{x}-\mu)^2}$$

となり、座標値を求める際に、自由度としてこれから求めようとしているデータ数を使わなければならなくなるからです。