身長と体重に関する正規性の検定

2015SS097 平田翔馬
指導教員：小藤俊幸

1　はじめに

正規分布が重要とされているのは，自然現象の中に正規分布に従う分布をしている現象が様々存在しているからである．さらに，最小二乗法をはじめとする，多くの統計的手法において，正規分布が仮定されていることからも，データの正規性を確認しておくことは重要である．

統計的方法における正規性の確認は，大きく分けて 2 つの方法で行うことができる．一つ目はグラフを利用し，視覚的に確認する方法である．二つ目はデータの分布を評価する，適合度の検定おこなう方法である．

今回は，統計処理ソフト R を使用して，文部科学省の平成 30 年度の学校保健統計調査から身長と体重の年齢別分布を分析する．そして，一般的に身長は正規分布し，体重は正規分布しないとされていることを確認していく．

2　グラフによる視覚的な確認

視覚的に正規性を確認する方法として代表的なものに，ヒストグラムと Q-Q プロットがある．例として，17 歳の女子の図を用いる．

2.1　ヒストグラムによる確認

ヒストグラムとは，度数分布表を柱状のグラフに表現したものである．Y 軸には度数もしくは相対度数をとり，X 軸には階級値をとる．

図 1，2 はそれぞれ，17 歳女子の身長と体重のヒストグラムである．曲線は正規分布の密度曲線を表し，破線はデータの平均値 μ を示している．身長のヒストグラムは，この密度曲線に沿っていて，正規分布に従っていると考えられる．それに対し，体重のヒストグラムは密度曲線よりも右に偏っていて，正規分布に従わないと考えられる．

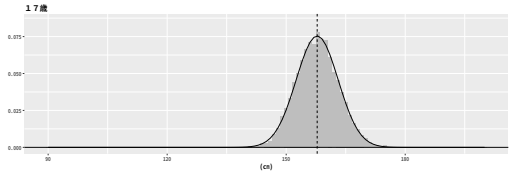

図 1　平成 30 年 17 歳女子の身長のヒストグラム

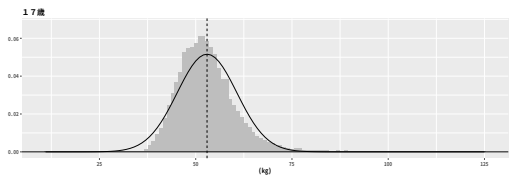

図 2　平成 30 年 17 歳女子の体重のヒストグラム

2.2　Q-Q プロットによる確認

Q-Q プロットとは，二つの分布を互いに対してプロットすることで比較する方法である．今回は，Y 軸にデータの分位数をとり，X 軸に正規分布の分位数をとり Q-Q プロットを作成する．

図 3 と 4 に，17 歳女子の身長と体重の Q-Q プロットを示す．実線は $y = \sigma x + \mu$ の直線であり，データが正規分布に従う場合を表している．身長は，ほとんどの点が直線にのっているので，正規分布に従っていると考えられる．体重は，直線とは大きく離れているため，正規分布に従わないと考えられる．

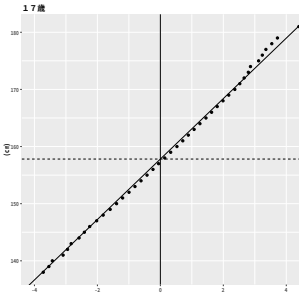

図 3　平成 30 年 17 歳女子の身長の Q-Q プロット

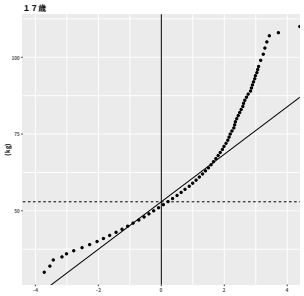

図 4　平成 30 年 17 歳女子の体重の Q-Q プロット

3　分布の適合度検定

適合度検定とは，統計学における仮説検定のうち，対象とする確率分布のもとでの期待値に対する，観測データの当てはまりやすさを検定するものである．

正規分布に対して適合度検定をする際は，帰無仮説 H_0 を「観測データは正規分布に従う」，対立仮説 H_1 を「観測データは正規分布に従わない」として，帰無仮説が棄却される有意水準を上側 5 %で片側検定を行う．

3.1　カイ二乗検定

カイ二乗検定は，あるデータが目的とする分布に従うかどうかを調べるための，適合度検定の一つである．この検定で用いる検定統計量，カイ二乗値 χ^2 は次のように定義される．

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1) \tag{1}$$

ここでの O_i と E_i はそれぞれ，観測度数と期待度数であり，k はグループの個数である．

例として，17 歳男子の身長と体重のカイ二乗値はそれぞ

れ 12.3，4.8×10^5 である．身長の上側 5 %点は 135.4 となるので，「身長は正規分布に従わないとはいえない」，体重の上側 5 %点は 139.9 となるので，「体重は正規分布に従わない」という検定結果となる．次に，16 歳男子の身長のカイ二乗値は 227.6 となるので，「身長は正規分布に従わない」となる．ここで，16 歳男子の身長の Q-Q プロット，図 5 を確認すると，ほとんどの点が直線に沿っており，正規分布に従っているように見える．しかし，直線から離れた左下にある外れ値が影響して，カイ二乗値が大きくなってしまう．図 6 は，この外れ値をデータから除いて描いた Q-Q プロットである．さらに，この外れ値を除いて計算したカイ二乗値は 24.3 となり，「身長の分布は正規分布に従わないとはいえない」という検定結果になる．

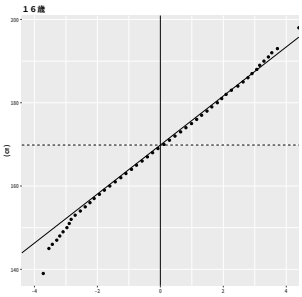

図 5　平成 30 年 16 歳男子の身長の Q-Q プロット

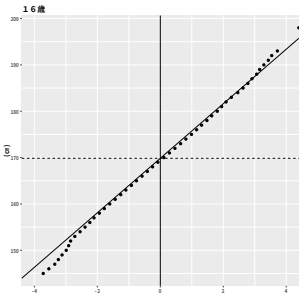

図 6　平成 30 年 16 歳男子の身長から外れ値を除いた Q-Q プロット

3.2　ジャック-ベラ検定

ジャック-ベラ検定は，データが正規分布に従う歪度と尖度を有しているかを調べる適合度検定である．検定統計量 JB は，歪度 Sk と尖度 Ku を用いて次のように定義される．

$$JB = \frac{n}{6}(Sk^2 + \frac{1}{4}Ku^2) \sim \chi^2(2) \qquad (2)$$

17 歳女子の身長と体重それぞれの検定統計量 JB の値は，1.68 と 960.8 である．自由度 2 のカイ二乗分布での上側 5 %点は 5.99 なので，検定結果は「身長は正規分布に従わないとはいえない」となり，「体重は正規分布に従わない」となる．

3.3　R を用いた適合度検定

統計処理ソフト R に実装されている，シャピロ-ウィルク検定とコルモゴロフ-スミルノフ検定を行う．

R の持つ無作為抽出を行う関数を使い，度数分布表からデータの再抽出をして適合度検定を行う．検定統計量と p 値には 1000 回の検定の平均を用い，再抽出するデータの個数を 100 個から 500 個まで 100 個ずつ増やして検定を行う．

女子の身長について，どちらの検定もサンプル数が 400 個までは，ほとんどの場合で p 値は 5 %より大きく，帰無仮説は棄却されないため，検定の結果は「身長のは正規分布に従わないとはいえない」となる．しかし，サンプル数が 500 個になると，シャピロ-ウィルク検定では p 値が 5 %を下回る年齢が存在するのに対して，コルモゴロフ-スミルノフ検定ではすべての年齢で p 値が 5 %よりも大きい．

女子の体重については，サンプル数が 100 個の場合，シャピロ-ウィルク検定ではほとんどの場合で帰無仮説が棄却されるのに対して，コルモゴロフ-スミルノフ検定はすべての年齢で帰無仮説は棄却されずに体重が正規分布に従うという結果になる．200 個の場合でも、シャピロ-ウィルク検定ではすべての年齢で帰無仮説は棄却されるが，コルモゴロフ-スミルノフ検定では帰無仮説を棄却できないものが存在する．

以上のことから，シャピロ-ウィルク検定はサンプル数が少ない時に，コルモゴロフ-スミルノフ検定はサンプル数が多い時に有効な検定の手法だと考えられる．

4　おわりに

データの正規性を確認する方法として，二つの図的表現と四つの適合度検定について検証した．ヒストグラムや Q-Q プロットを用いると，視覚的に分かりやすく正規性の有無が判断できる．カイ二乗検定とジャック-ベラ検定のように，度数分布表からでも検定統計量が計算できるものは，様々なデータに対して容易に検定を行うことができる．一方で，要約されたデータに外れ値が存在する場合，検定統計量がその影響を大きく受ける．R 上での検定には，シャピロ-ウィルク検定とコルモゴロフ-スミルノフ検定を用いた．これにより，サンプル数が検定の結果に影響することを確認できた．

検定の方法によって，結果に違いがみられたものの，一般に知られている「身長は正規分布に従う」，「体重は正規分布に従わない」という説は，ほとんどの場合で正しいことを実証することができた．

参考文献

[1] James O. Adefisoye: *Testing Normality: An Assessment of the Performances of Several Univariate Tests of Normality*. LAP LAMBERT Academic Publishing, 2016.

[2] 刈屋武昭・勝浦正樹: 『統計学 第 2 版 〈プログレッシブ経済学シリーズ〉』. 東洋経済新報社, 2008.

[3] 柴田義貞: 『正規分布 特性と応用 UP 応用数学選書 3』. 財団法人 東京大学出版, 1981.

[4] 白旗慎吾: 『統計解析入門』. 共立出版株式会社, 1992.

[5] 中川重和『統計学 One Point 16 正規性の検定』共立出版株式会社, 2019.

[6] 松下貢『統計分布を知れば世界がわかる』中央公論新社, 2019.