

1. 分布が一山でないとか、非対称である場合など正規性が成り立たない場合、ただし正規性が成り立つ場合でも、サンプルサイズがある程度大きくなれば、検定に比べて少し検出力(有意になりやすさ)が低下するものの、それほど性質が悪いわけではない。

### 2. 順序カテゴリカルデータ

データが、-, 土, +, ++等で示される場合がある。このようなデータについては平均を求めることはできないし、また-と土の間隔、土と+の間隔も等しいわけではない。しかしながら、土は一より、+は土より程度がよいという順序関係はわかるので、順位和検定は行える。実際には、-, 土, +, ++に1, 2, 3, 4の数値を与えてNPARIWAYプロシジャで解析すればよい。

### 3. 検出限界以下のデータが存在する場合

分析データでは“検出限界以下”という形でのデータが得られる場合がある。このようなデータについては、たとえば検出限界が1 ppmであれば、実際の値が0.1 ppmであっても0.7 ppmであっても1 ppm以下として同じ扱いを受ける。このようなケースを“打切り”が存在するという。打切りデータが存在する場合でも1 ppm以下は2 ppmや3 ppmよりは小さいという順序関係はわかるので、順位和検定は行える。

#### NPARIWAY プロシジャ

書式 PROC NPARIWAY オプション;

CLASS 変数;

[VAR 変数のリスト;]

[BY 変数のリスト;]

文例 proc npariway data=school;

class sex;

var rika syakai;

機能 対応のない2群(または多群)のノンパラメトリック検定を行う。

## 6-4 対応のある2群の検定 (UNIVARIATE プロシジャ)

対応のある検定とは、たとえば次のような例で用いられる。

ある薬物を8匹の犬に投与して、血液中の成分Aを測定したところ、次の結果を得た。投与前後でAの濃度に差があるかを調べたい。同一の個体について、投与前と投与後の値が測定されている。投与前のデータ8件は、投与後のデータ8件と犬の個体番号によって対応付けられる。

犬の番号	1	2	3	4	5	6	7	8	平均
投与前	3.51	3.07	3.29	3.03	3.38	3.30	3.15	3.25	3.24750
投与後	3.39	3.39	3.20	3.11	3.17	3.09	3.17	3.09	3.20125
差	0.12	-0.32	0.09	-0.08	0.21	0.21	-0.02	0.16	0.04625

図表 6.4 UNIVARIATE プロシジャによる対応のある検定

Univariate Procedure			
Variable=DIF			
Moments			
N	8	Sum Wgts	8
Mean	0.04625	Sum	0.37
Std Dev	0.180629	Variance	0.032627
Skewness	-1.31523	Kurtosis	1.511099
USS	0.2455	CSS	0.228388
CV	390.5489	Std Mean	0.063862
T:Mean=0	0.724218	Pr> T	0.4924
Num ^= 0	8	Num > 0	5
M(Sign)	1	Pr>= M	0.7266
Sgn Rank	7	Pr>= S	0.3594
Quantiles(Def=5)			
100% Max	0.21	99%	0.21
75% Q3	0.185	95%	0.21
50% Med	0.105	90%	0.21
25% Q1	-0.05	10%	-0.32
0% Min	-0.32	5%	-0.32
		1%	-0.32
Range	0.53		
Q3-Q1	0.235		
Mode	0.21		
Univariate Procedure			
Variable=DIF			
Extremes			
Lowest	Obs	Highest	Obs
-0.32(	2)	0.09(	3)
-0.08(	4)	0.12(	1)
-0.02(	7)	0.16(	8)
0.09(	3)	0.21(	5)
0.12(	1)	0.21(	6)

対応のある  $t$  検定はデータから個体ごとに投与前と投与後の差を求めて、その差が0であるかを検定する。このため差の平均をそのバラツキの大きさと比較する。

SAS では 1 変量解析用の UNIVARIATE プロシジャによって対応のある検定を行うことができる。次にプログラム例を示す。

```
data pair;
  input x y @@;
  dif=x-y;
cards;
```

```

3. 51 3. 39 3. 07 3. 39 3. 29 3. 20 3. 03 3. 11
3. 38 3. 17 3. 30 3. 09 3. 15 3. 17 3. 25 3. 09
;
proc univariate data=pair;
    var dif;

```

UNIVARIATE プロシジャでは、ある変量の平均あるいはメディアンが0であるかを検定するのに3種類の検定が用意されている。

T:Mean=0 平均が0かを調べる Student の  $t$  検定(対応のある  $t$  検定)

M (Sign) 符号検定(条件付き2項検定)

Sgn Rank ウィルコクソンの符号付き順位和検定

それぞれの検定の  $p$  値は 0.4924, 0.7266, 0.3594 でいずれも 5%水準で有意ではない。このことから薬物投与によって、成分 A の濃度は変化しないことがわかる。

#### 対応のある検定の注意

(1) 対応のあるデータに対応のない  $t$  検定を適用すると

対応のあるデータに対して、対応を無視して形式的に対応のない  $t$  検定を適用することも可能であるがこれはまずい方法である。対応のない  $t$  検定では、投与前と投与後のそれぞれの平均を求めて、その差が0であるかを検定することになる。対応のある  $t$  検定では、これとは逆に先に個体ごとに投与前と投与後の差を求めてから平均を計算する。 $t$  検定を検定する分子自体は、対応のある  $t$  検定(差の平均)でも対応のない  $t$  検定(平均の差)でも等しいが、対応のない  $t$  検定では、ある個体は成分 A の濃度が生まれつき高く、ある個体は低いといった個体差が誤差に含まれるため、個体差が大きい場合には対応のある  $t$  検定に比べて有意になりにくい。対応のある  $t$  検定では、先に投与前と投与後の差をとることによって、A の濃度の個体差が誤差から除かれるので、差の検出感度が上がる。手続きとしては次のようなことになる。

対応のある  $t$  検定：個体ごとに投与前と投与後の差をとってから平均を求める。

対応のない  $t$  検定：投与前と投与後の平均を求めて差をとる。

(2) 3種の対応のある検定の使い分け

対応のある  $t$  検定では差の分布に正規分布を前提としている。ウィルコクソンの符号付き順位和検定や符号検定では、正規分布でなくても対称分布であれば、すなわち前と後の変数の分布が同じであれば構わない。外れ値に対する頑健性は

M(Sign) → Sgn Rank → T:Mean=0

の順であり、分布が正規分布のときの検出力はちょうどこの逆順である。

(3) Version 5 と Version 6 との違い

UNIVARIATE プロシジャでは対応のある  $t$  検定に関する機能が Version 6 から大きく変化しているので注意が必要である。

1. 符号検定が追加された。
2. 符号付き順位和検定の計算のしかたが変更された。

Release 6.07 から 0 でないデータの数が 20 以下のときには、可能なすべての並べかえのパターンを数え上げて、得られたデータより極端なパターンが得られる確率を計算する正確な検定を行うようになった。20 を越えるときは  $t$  分布による近似を行って  $p$  値を計算するが、この近似は通常の教科書ではあまり使われない方法を用いている。詳しくはマニュアルを参照されたい。

### 3. $p$ 値のデータセットへの出力

3 種類の検定の検定統計量と  $p$  値が OUTPUT ステートメントによってデータセットに落とせるようになった。対応のある検定を行う目的のみの場合には UNIVARIATE プロシジャの出力は冗長である。必要なら NOPRINT オプションを指定して、PRINT プロシジャで検定の結果のみ出力し直した方がよい。統計量と  $p$  値のキーワードを示す。

	統計量	$p$ 値
対応のある $t$ 検定	T	PROBT
符号検定(条件付き 2 項検定)	MSIGN	PROBM
符号付き順位和検定	SIGNRANK	PROBS

先の例で対応のある  $t$  検定で対応のある  $t$  検定の結果をデータセット O\_UNI に落としたいときは次のように指定すればよい。t = と probt = によって変数名を指定する。

```
proc univariate data=pair noprint;
  var dif;
  output out=o_uni t=t probt=p;
```

## 6-5 カイ2乗検定とフィッシャーの正確検定(FREQ プロシジャ)

### 2×2の分割表

分割表(contingency table)はあるクラス(分類)に含まれる対象の数を表にしたものである。分割表の中でもっともよく使われるのが2×2の分割表である。分割表はクロス集計表とも呼ばれる。これはデータを行・列2方向に配置した要因 A と B によって分類し、度数を示したものである。

たとえば次のような形式のデータを扱う。具体例として、40 人の患者をランダムに 20 人ずつ対照群と薬物投与群に分け、一定期間経過後の疾患の治癒を調べた結果を 2×2 の分割表にまとめたとしよう。

2×2の分割表

一般型				具体例			
要因 A/要因 B	B1	B2	計	処置/治癒	なし	あり	計
A1	$f_{11}$	$f_{12}$	$f_{1\cdot}$	対 照	13	7	20
A2	$f_{21}$	$f_{22}$	$f_{2\cdot}$	薬物投与	8	12	20
計	$f_{\cdot 1}$	$f_{\cdot 2}$	$f_{\cdot \cdot}$	計	21	19	40

このような分割表について、要因 A と要因 B の関連に興味が置かれる場合がしばしばある。具体例では対照群では治癒しなかった患者の数が多く、薬物投与群では治癒した患者の数が多くなっ