14　疑わしいデータの取扱い

同じ測定を繰り返したとき，1 つのデータだけが他のものに比べて飛び離れた値をとることがある。実験条件とは別の要因でそのような飛び離れた値が得られたならば，そのデータは採用せず棄却することになる。しかし，何かの基準に基づいて棄却すべきかどうかの判断を下さないと，実験データに実験者の恣意が入り込むことになる。ここでは，そのような場合によく用いられる棄却検定のいくつかの例を紹介し，棄却検定に実際に使用できるような Excel ワークシートを作成する。

この節では主に次の書物を参考にした。

- 藤森　利美,『分析技術者のための統計的方法』，丸善，東京，1986

14.1　問題設定

同じ測定を繰り返して n 個のデータを得た。それを小さい順（昇順）に並べたところ，x_1 から x_{n-1} までは，おおむね似た値になったが，x_n だけは少々離れた値（異常値）になった。このとき，x_n をデータから除外する（棄却する）ことは，統計的に妥当か？

データの最大値ではなく最小値が異常であるという問題設定もあり得る。その場合は，データを小さい順でなく大きい順（降順）に並べたとすれば，以下の説明があてはまる。

ここでは，異常値が 1 つのみである場合に限って話を進める。また，特に断らない限り $3 \le n \le 10$ の場合を取り扱うことにする。それ以外の場合については，上記の参考書を参照して欲しい。

14.2　4d ルール

簡単なので $4 \le n$ の場合に広く用いられているが，統計学的な根拠に乏しい。

1. 異常値以外の平均 $\langle x \rangle'$ をとる。

$$\langle x \rangle' = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i \tag{14.1}$$

2. 異常値以外について偏差の絶対値 $|x_i - \langle x \rangle'|$ の平均をとり，それを d とする。

$$d = \frac{1}{n-1} \sum_{i=1}^{n-1} \left| x_i - \langle x \rangle' \right| \tag{14.2}$$

3. 異常値 x_n とそれ以外の平均の $\langle x \rangle'$ の差の絶対値を求め，それを d' とする。

$$d' = \left| x_n - \langle x \rangle' \right| \tag{14.3}$$

4. $d' > 4d$ ならば異常値 x_n は棄却する。そうでなければ棄却しない。

最後の手順で 4d の代わりに 2.5d が用いられることもある。その場合は「2.5d ルール」という。4d の場合も，2.5d の場合も，第 1 種の誤り（有効なデータが切り捨てられてしまう 誤り）が生じやすい。

Excel 2007

1. Web ページから 4d ルール説明用の CSV ファイルをダウンロードして Excel で開く。

	A	B	C	D	E	F	G	H
1	No	データ	偏差絶対値	異常値が先頭に来るようにデータをソートしてください				
2	1			データ数は4以上10以下です				
3	2							
4	3			疑わしい値以外の平均				
5	4							
6	5			疑わしい値以外の残差の平均(d)				
7	6							
8	7			dの4倍		dの2.5倍		
9	8							
10	9			疑わしい値とそれ以外の平均との差(d')				
11	10							
12				4dルールの判定		2.5dルールの判定		
13								

2. 課題提出用のファイル名で，Excel 形式で保存する。
3. B2:B11 の範囲にデータを入力する。空白があってもかまわない。とりあえず，適当に 1 付近の数値を 5 個と 2 付近の数値を 1 つ入れて，残りは空白にする。
4. 異常値が先頭に来るようにソートする。
   (i) B 列を選択し，メニュー [データ] → [並べ替え]。
   (ii)「並べ替えの前に」で「現在選択されている範囲を並べ替え」を ON → [並べ替え]。
   (iii)「並べ替え」で「最優先されるキー」は「データ」，そして「降順」を ON → [OK]（もしも異常値が最小値なら「降順」の代わりに「昇順」を ON）。
5. D5 には異常値以外 B3:B11 の平均を計算する。あとで自由に別のデータを入れたときにも使えるように，空白部分まで範囲に含めておく。
6. B2 が空白ならば C2 を空白に，そうでなければ C2 に異常値以外の平均と B2 との差の絶対値を計算する。
   C2 に「=IF(B2="","",ABS(B2-D5))」
7. C2 をコピーして C3:C11 に貼り付ける。もし，前の操作で「IF」を使わずに，「=ABS(B2-D5)」としたらここで何がおきるか，確かめてみること。
8. D7 には C3:C11 の平均を計算する。
9. D9 には D7 の 4 倍，F9 には D7 の 2.5 倍を計算する。
10. D11 は C2 と同じ。
11. D13 には，D11 が D9 より大きければ「**棄却**」と表示され，そうでなければ「**棄却不可**」と表示されるようにする。
    D13 は「=IF(D11>D9,"**棄却**","**棄却不可**")」
12. これを応用して，F13 には，D11 が F9 より大きければ「**棄却**」と表示され，そうでなければ「**棄却不可**」と表示されるようにする。

14.3 Q テスト

Q テストは，Dean と Dixon によって考えられた方法で，統計学的にも根拠がある。

1. 全データの範囲 R を求める。

$$R = |x_n - x_1| \tag{14.4}$$

2. 異常値とそれに最も近いデータの差の絶対値 R' を求める。

$$R' = |x_n - x_{n-1}| \tag{14.5}$$

3. R'/R が，データ数 n ごとに予め与えられている Q 値より大きければ異常値 x_n は棄却する。そうでなければ棄却しない。
   Dean と Dixon の原論文 (R. B. Dean and W. J. Dixon, *Anal. Chem.*, **23**, 636 (1951)) には，90% の信頼限界における Q 値の表が与えられている。

n	$Q_{0.90}$
3	0.90
4	0.76
5	0.64
6	0.56
7	0.51
8	0.47
9	0.44
10	0.41

1. Web ページから Q テスト説明用の CSV ファイルをダウンロードして Excel で開く。

	A	B	C	D	E	F	G
1	No	データ	異常値が先頭に来るようにデータをソートしてください				
2	1		データ数は3以上10以下です				
3	2					Q値の表	
4	3		データ数(n)			3	
5	4					4	
6	5		データの範囲(R)			5	
7	6					6	
8	7		異常値と近接する値の差(R')			7	
9	8					8	
10	9		R'/R			9	
11	10					10	
12			Q値				
13							
14			判定				
15							

2. 「ホーム」→「セル」→「書式」→「ワークシートの移動またはコピー」で，「4*d* テスト」と同じファイルにワークシートを移動する。
3. B2:B11 の範囲にデータを入力する。空白があってもかまわない。とりあえず，適当に 1 付近の数値を 5 個と 2 付近の数値を 1 つ入れて，残りは空白にする。
4. G4:G11 の範囲に Q 値の表を入力する。
5. C5 にデータ数「=COUNT(B2:B11)」。
6. C7 にデータの範囲「=MAX(B2:B11)−MIN(B2:B11)」。
7. C9 に異常値と近接する値との差「=ABS(B2−B3)」。
8. C11 に R'/R「=C9/C7」。
9. C13 に Q 値「=VLOOKUP(C5,F4:G11,2)」。

   [Excel 関数の説明] 「VLOOKUP」は表の中から値を取り出すときに使う。この関数には 3 つの引数がある。この例で説明すると，C5 という値（検索値）を F4:G11 の範囲の左端の列（この場合 F 列の F4:F11）で検索し，その行のなかで 2 列目（3 番目の引数「2」，この場合 G 列を指す）に書かれている数値を返す。

   [Excel 関数の説明] 「VLOOKUP」では列（縦方向 vertical ）を検索したが，行（横方向 horizontal ）を検索する場合「HLOOKUP」を用いる。
10. C15 に判定「=IF(C11>C13,"**棄却**","**棄却不可**")」。

14.4 Dixon の方法

Q テストを発表した後に Dixon が考えた方法で，しばしば用いられる。

1. データ数 n によって次の中から式を選び，r 値を求める。

$$r = \frac{|x_n - x_{n-1}|}{|x_n - x_1|} \quad \text{for} \quad 3 \le n \le 7 \tag{14.6}$$

$$r = \frac{|x_n - x_{n-1}|}{|x_n - x_2|} \quad \text{for} \quad 8 \le n \le 10 \tag{14.7}$$

$$r = \frac{|x_n - x_{n-2}|}{|x_n - x_2|} \quad \text{for} \quad 11 \le n \le 13 \tag{14.8}$$

$$r = \frac{|x_n - x_{n-2}|}{|x_n - x_3|} \quad \text{for} \quad 14 \le n \le 25 \tag{14.9}$$

2. ある有意水準（例えば 0.10）で表に与えられた棄却限界よりも r 値が大きければ，その有意水準（危険率）で異常値を棄却できる。そうでなければ棄却できない。

棄却限界の表

n	有意水準			n	有意水準		
	0.10	0.05	0.01		0.10	0.05	0.01
3	0.886	0.941	0.988	10	0.409	0.477	0.597
4	0.679	0.765	0.889	11	0.517	0.576	0.679
5	0.557	0.642	0.780	12	0.490	0.546	0.642
6	0.482	0.560	0.698	13	0.467	0.521	0.615
7	0.434	0.507	0.637	14	0.492	0.546	0.641
8	0.479	0.554	0.683	15	0.472	0.525	0.616
9	0.441	0.512	0.635				

Excel 2007

ここでは，$3 \le n \le 13$ のみに対応したワークシートを作成する。

1. Web ページから Q テスト説明用の CSV ファイルをダウンロードして Excel で開く。

	A	B	C	D	E	F	G	H	I
1	No	データ	異常値が先頭に来るようにデータをソートしてください						
2	1		データ数は3以上13以下です						
3	2					棄却限界の表			
4	3		有意水準	表の列			0.1	0.05	0.01
5	4					3			
6	5		データ数(n)	表の行		4			
7	6					5			
8	7		分子の後半			6			
9	8					7			
10	9		分母の後半			8			
11	10					9			
12	11		r値			10			
13	12					11			
14	13		棄却限界			12			
15						13			
16			判定			14			
17						15			

2. 「ホーム」→「セル」→「書式」→「ワークシートの移動またはコピー」で，「4*d* テスト」，「Q テスト」と同じファイルにワークシートを移動する。
3. B2:B14 の範囲にデータを入力する。空白があってもかまわない。とりあえず，適当に 1 付近の数値を 5 個と 2 付近の数値を 1 つ入れて，残りは空白にする。
4. G5:I15 の範囲に上の表の該当する部分を入力する。
5. C5 に求めたい有意水準（例えば「0.1」）を入力する。
6. D5 には C5 に入力した有意水準が，表の何列目に書かれているのかを表示する。
   D5 に「=MATCH(C5,G4:I4,0)」
   [Excel 関数の説明] 「MATCH」の働きをここの例で説明する。C5 という値（検索値）を G4:I4 の範囲で検索し，そのなかの何番目のセルにあったかを返す。最後の引数「0」は，検索値が見つからなかったときにエラーを返すことを指定している。
7. C7 にデータ数「=COUNT(B2:B14)」。
8. D7 には C7 に表示されたデータ数が，表の何行目に書かれているのかを表示する。
   D7 に「=MATCH(C7,F5:F15,0)」
   [Excel 関数の説明] 「MATCH」を，今度は縦の範囲で使用している
9. C9 に r の計算の分子の後半を書く。「=IF(C7<=10,B3,B4)」
   $n \le 10$ なら x_{n-1} なので B3，$11 \le n \le 13$ なら x_{n-2} なので B4。
10. C11 に r の計算の分母の後半を書く。「=IF(C7<=7,VLOOKUP(C7,A2:B14,2),VLOOKUP(C7-1,A2:B14,2))」
    $n \le 7$ なら x_1，$8 \le n \le 13$ なら x_2 なので，n の値が変わっても対応できるように，このようにした。
11. C13 に r を計算「=ABS((B2-C9)/(B2-C11))」。
12. C15 に有意水準とデータ数に対応する規格限界の値を表示する「=INDEX(G5:I15,D7,D5)」。

[Excel 関数の説明] 「INDEX」の働きをここの例で説明する。G5:I15 という範囲のなかで D7 行目 D5 列目のセルの値を返す。

13. C17 に判定結果を書く「=IF(C13>C15,"**棄却**","**棄却不可**")」

14.5 Grubbs の方法

手計算では Dixon の方法よりも計算手順が煩雑だが，より推奨されている方法に Grubbs の方法がある。Excel を使った場合，こちらの方がより簡単である。

1. 全データの平均 $\langle x \rangle$，不偏分散 V，標準偏差 σ を計算する。

$$\langle x \rangle = \frac{1}{n} \sum_{n}^{j=1} x_j \tag{14.10}$$

$$\begin{aligned} V &= \frac{1}{n-1} \sum_{n}^{j=1} \left(x_j - \langle x \rangle \right)^2 \\ &= \sigma^2 \end{aligned} \tag{14.11}$$

2. 次の式で定義される T を計算する。

$$T = \frac{|x_n - \langle x \rangle|}{\sigma} \tag{14.12}$$

3. T の値が右の表にあげた棄却限界よりも大きければ，その有意水準（危険率）で異常値 が棄却できる。なお，この表は異常値が最大値になるか最小値になるかわからない場合の表である。もしも，異常値が常に大きめに出る（あるいは小さめに出る）ことが予め予想されるなら，有意水準の値を半分にして使用する。

棄却限界の表

n	有意水準				n	有意水準			
	0.10	0.05	0.02	0.01		0.10	0.05	0.02	0.01
3	1.153	1.155	1.155	1.155	10	2.176	2.290	2.410	2.482
4	1.463	1.481	1.492	1.496	11	2.234	2.355	2.485	2.564
5	1.672	1.715	1.749	1.764	12	2.285	2.412	2.550	2.636
6	1.822	1.887	1.944	1.973	13	2.331	2.462	2.607	2.699
7	1.938	2.020	2.097	2.139	14	2.371	2.507	2.659	2.755
8	2.032	2.126	2.221	2.387	15	2.409	2.549	2.705	2.806
9	2.110	2.215	2.323	2.387					