13 スチューデントの t

前回，実験値に基づいて母集団の平均を推定する方法を学んだ。しかし，そこで用いた方法は，母集団の分散が予めわかっている場合にのみ適用できる方法であった。実際には，母集団の分散が予めわかっている場合は極めてまれである。実用的には，母集団の分散が未知の場合についての平均の推定法を知る必要がある。ここでは，そのような方法として，「t 推定」，「t 検定」と呼ばれるものを学ぶ。

この節で参考にした書物は前章と同じである。

13.1 t 分布 t distribution

ここでも前回と同様の乱数を題材にする。次の式を用いて 2 つの一様乱数 U_1, U_2 から 1 つの乱数 X を発生させる。

$$X = 5 + \sqrt{-\log U_1}\cos(2\pi U_2) \tag{13.1}$$

理想的には平均 $\mu = 5$，分散 $\sigma^2 = 1/(2\ln 10) = 0.217147\cdots$ の乱数になる。

Excle

1. Web ページから t-分布説明用のファイルをダウンロードし，Excel で開く。

	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		n=3	3			n=6	6						
2	行番号	平均	不偏分散	t		平均	不偏分散	t					乱数
3	1												4.412075
4	2												

2. ワークシート名を「t-bunpu」から「t-分布」に変更する。
3. 課題提出用のファイル名で，Excel 形式で保存する。
4. S3 に「=5+SQRT(−LOG(RAND()))∗ COS(2∗PI()∗RAND())」と入力
5. S3 をクリップボードにコピーし，S3:AD1002 の範囲に貼り付ける。

X_i の中から n 個を標本として抽出し $x_1, x_2, \cdots, x_n$ とする。次のような量 t を定義し，抽出操作を繰り返して t の分布を調べる。

$$t = \frac{\langle x\rangle - \mu}{\sqrt{V/n}} \tag{13.2}$$

ただし，μ は母集団 X の平均，$\langle x\rangle$ は n 個の標本に関する平均，V は n 個の標本に関する不偏分散である。

$$\langle x\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{13.3}$$

$$V = \frac{1}{\phi}\sum_{i=1}^{n} (x_i - \langle x\rangle)^2 \tag{13.4}$$

また，ϕ は自由度で，ここでは次のような量になる。

$$\phi = n - 1 \tag{13.5}$$

サンプル数が $n = 3, 6$ の 2 種類の場合について t の分布をもとめ，正規分布と比較する。

Excel 2007

1. 引き続き先のシートを用いる。

2. G 列に番号を打つ。
3. H3 に 3 つのサンプルの平均「=AVERAGE(S3:U3)」，I3 に 3 つのサンプルの不偏分散「=VAR(S3:U3)」，J3 に 3 つのサンプルで求めた t の値。「=(H3-$B$5)/SQRT(I3/$I$1)」。
4. L3 に 6 つのサンプルの平均「=AVERAGE(S3:X3)」，M3 に 6 つのサンプルの不偏分散「=VAR(S3:X3)」，N3 に 6 つのサンプルで求めた t の値。「=(L3-$B$5)/SQRT(M3/$M$1)」。
5. H3:N3 をコピーし，H4:N1002 に貼り付け。
6. 度数分布を求めるために, 右のように入力する。ダウンロードしたファイルではすでに入力されている。

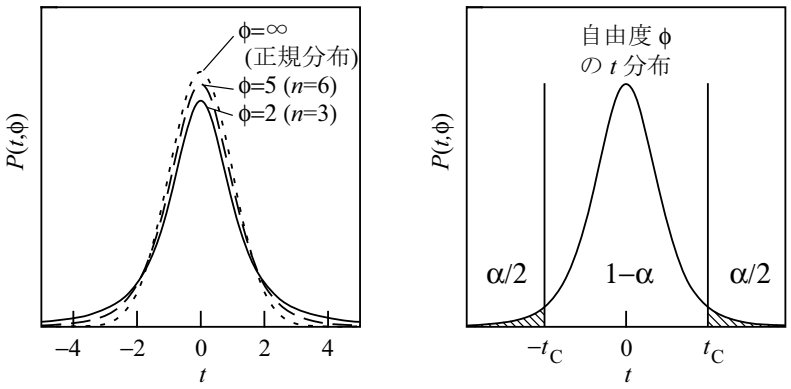
   (i) C9:C30 の範囲に
   「=FREQUENCY(J3:J1002,A9:A29)/COUNT(J3:J1002)/0.5」，Ctrl+Shift+Enter。

   (ii) D9:D30 の範囲に
   「=FREQUENCY(N3:N1002,A9:A29)/COUNT(N3:N1002)/0.5」，Ctrl+Shift+Enter。

	A	B	C	D	E
1					
2					
3	理想値				
4	母集団				
5	平均	5	分散	0.217147	
6					
7	t分布				
8	範囲	値	phi=2	phi=5	正規分布
9	−5	−5.25			
10	−4.5	−4.75			
11	−4	−4.25			
12	−3.5	−3.75			
13	−3	−3.25			
14	−2.5	−2.75			
15	−2	−2.25			
16	−1.5	−1.75			
17	−1	−1.25			
18	−0.5	−0.75			
19	0	−0.25			
20	0.5	0.25			
21	1	0.75			
22	1.5	1.25			
23	2	1.75			
24	2.5	2.25			
25	3	2.75			
26	3.5	3.25			
27	4	3.75			
28	4.5	4.25			
29	5	4.75			
30		5.25			

7. 正規分布を求める。
   (i) E9 に「=NORMDIST(B9,0,1,0)」，E9 をコピーし，E10:E30 に貼り付け。
8. グラフを描く。
   (i) B8:E30 の四角い範囲を選択。
   (ii) メニュー [挿入] → [グラフ] →「散布図」で，形式は右下折れ線のみを選択→ [完了]。
   (iii) X 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「-5」に，「最大値」を「5」，「Y/数値軸との交点」を「-5」にして [OK]。
   (iv) Y 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「0」に，「最大値」を「0.5」にして [OK]。

何度か再計算させてみて（空のセルを選択し Delete を押す），分布の様子をみる。t の分布は，正規分布に比べて頭が低く，すそが高い。そして n が小さいほど，その傾向が顕著である。このような分布は Student の t 分布と呼ばれ，t の値の他に自由度 $\phi(= n - 1)$ に依存する。Student というのは，この分布を発見したイギリスの化学者 W. S. Gossert のペンネームである。Excel には Student の t 分布そのものを計算する関数は用意されていない（おそらく，実用上あまり必要ないから）。いまの例について t 分布のグラフを書くと図のようになる。自由度 ϕ が無限大の場合に t 分布は正規分布と一致する。



13.2 t 分布による区間推定

前回，母集団の分散が既知の場合について，n 個のサンプルの平均 $\langle x \rangle$ から基準値 z を求め，z が平均 0，分散 1 の正規分布にしたがうことを利用して，母集団の平均を推定した。z の代わりに t を，正規分布の代わりに t 分布を用いれば，同様の方法で，母集団の分散が未知の場合について母集団の平均を推定できる。

n 個のサンプリングを繰り返して行って標本平均 $\langle x \rangle$ と標本の不偏分散 V を何度も求めたとき，$t = (\langle x \rangle - \mu) / \sqrt{V/n}$ がある正の値 t_C 以下になる確率が $[1 - (\alpha/2)]$ であるとする（例えば $\alpha = 0.05$ で $1 - (\alpha/2) = 0.975$ ）。t_C の値は，自由度 $\phi = n - 1$ の t 分布の性質から求めることができる（具体的には $n = 6, \phi - 5$ で $1 - (\alpha/2) = 0.975$ なら $t_C = 2.571$ ）。このとき，t が t_C 以上になる確率は $\alpha/2$ である。さらに，t が $-t_C$ 以下になる確率も $\alpha/2$ である。つまり，$-t_C \le t \le t_C$ となる確率（信頼率）は $1 - 2 \times (\alpha/2) = 1 - \alpha$ である（いまの例なら $1 - \alpha = 0.95$ で，パーセントで言えば 95%）。これを次のように表す。

$$P(-t_C \le t \le t_C) = 1 - \alpha \tag{13.6}$$

次に，カッコの中のみを変形する

$$-t_C \le t \le t_C \tag{13.7}$$

$$-t_C \le \frac{\langle x \rangle - \mu}{\sqrt{V/n}} \le t_C \tag{13.8}$$

$$\langle x \rangle - t_C \sqrt{\frac{V}{n}} \le \mu \le \langle x \rangle + t_C \sqrt{\frac{V}{n}} \tag{13.9}$$

確率の形で書くと次のようになる。

$$P\left(\langle x \rangle - t_C \sqrt{V/n} \le \mu \le \langle x \rangle + t_C \sqrt{V/n}\right) = 1 - \alpha \tag{13.10}$$

この式の意味するところを確認しておく。母集団の分散が未知である場合について，サンプル数 n の測定を行って平均値 $\langle x \rangle$ と不偏分散 V を得た。そして，信頼区間 $\langle x \rangle - t_C \sqrt{V/n} \le \mu \le \langle x \rangle + t_C \sqrt{V/n}$ を算出した。このサンプリングと信頼区間の算出を何度も繰り返して行ったとき，計算された信頼区間のうち $(1 - \alpha)$ の割合のものは母集団の平均 μ を含んでいる。ただし，t_C の値は，自由度 $\phi = n - 1$ の t 分布の性質から求める。

13.3 t 分布による区間推定の具体例

前回，母集団の分散が既知である場合に用いた例について，母集団の分散が未知であるとみなして，t 分布を用いた区間推定の具体例を示す。

Excel 2007

1. 前回のワークシートを開く。

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	データ											
3	4.969781	5.316936	5.163603	5.023216	5.366835	4.197772	5.446009	5.079964	4.384615	5.272127	4.309788	4.883102
4												
5	平均値	不偏分散	サンプル数	自由度								
6	4.951145	0.184208	12	11								
7												
8	z推定					t推定						
9	母集団分散	標本平均分散										
10	0.217147	0.018096										
11	alpha(a)	信頼率	1−(a/2)	Zc		alpha(a)	信頼率	tc				
12	0.05	0.95	0.975	1.959964								
13												
14	下限	平均	上限			下限	平均	上限				
15	4.687491	4.951145	5.2148									
16												
17	z検定					t検定						
18	無帰仮説					無帰仮説						
19	平均値	5				平均値	5					
20	下限	上限	判定			下限	上限	判定				
21	4.736346	5.263654	○									

2. 準備

   (i) 右の部分を用いる。

   (ii) F12 に α の値を入力「0.05」，G12 に信頼率を計算「=1−F12」。

	F	G	H
8	t推定		
9			
10			
11	alpha(a)	信頼率	tc
12			
13			
14	下限	平均	上限
15			
16			
17	t検定		
18	無帰仮説		
19	平均値	5	
20	下限	上限	判定
21			

3. 信頼区間の計算

   (i) H12 に「=TINV(F12,D6)」。

   [Excel 関数の説明] 「TINV」は，t 分布で，自由度 ϕ と信頼率 $(1-\alpha)$ が与えられたときに，t が $-t_C$ から t_C までの値をとる確率 $P(-t_c \le t \le t_C)$ の値が $(1-\alpha)$ となるような t_C の値を返す関数。引数は 2 つで，はじめは α の値，2 番目は自由度 ϕ。

   (ii) F15 に区間の下限を計算「=A6−H12∗SQRT(B6/C6)」。

   (iii) G15 に区間の中心を表示「=A6」。

   (iv) H15 に区間の上限を計算「=A6+H12∗SQRT(B6/C6)」。

4. このシートでデータを消せば少ないデータの場合も計算できるし，α の数字をかえれば異なる信頼率でも計算できる。

13.4 t 分布による検定

t 分布を利用した検定も同様に考えられる。

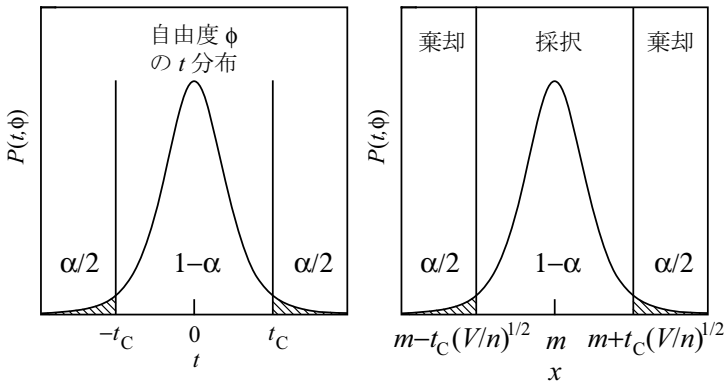
ここでは，母集団の分散は未知であるとする。先ほどの区間推定の場合と同じように t を定義する。また，ある信頼率の下で，同様に t_C を定義する。無帰仮説を「母集団の平均値が $\mu = m$ である」とたてる。この仮定が正しければ，t はその信頼率の確率で次の区間に含まれるはずである。

$$-t_C \le t \le t_C \tag{13.11}$$

$$-t_C \le \frac{\langle x \rangle - m}{\sqrt{V/n}} \le t_C \tag{13.12}$$

$$m - t_C\sqrt{\frac{V}{n}} \le \langle x \rangle \le m + t_C\sqrt{\frac{V}{n}} \tag{13.13}$$

したがって，得られた平均値 $\langle x \rangle$ がこの範囲に含まれていなければ，その信頼率で，$\mu = m$ という仮説が否定されることになる。



区間推定で用いたデータで検定の具体例を示す。

Excel 2007

1. シートの次の部分を用いる。

	F	G	H
17	t検定		
18	無帰仮説		
19	平均値	5	
20	下限	上限	判定
21			

2. G19 に無帰仮説の平均値「5」を入力。
3. F21 に区間の下限を計算「=G19−H12∗SQRT(B6/C6)」。
4. G21 に区間の上限を計算「=G19+H12∗SQRT(B6/C6)」。
5. H21 に判定を○×で表示する。「=IF(A6<F21,"×",IF(A6>G21,"×","○"))」
6. 再計算を繰り返せば，時々判定が×になる。

13.5 χ^2 分布 ~ 補足 ~

これまで，サンプリングによる少数のデータから母集団の平均を予測する方法を考えてきた。測定の目的によっては，母集団の分散を予測したい場合がある。母集団の分散に対する不偏推定量は標本の不偏分散であることはすでに述べた。しかし，区間推定や検定を行うためには，不偏推定量の他に，標本の不偏分散の分布に関する知識が必要である。そのときに重要になるのが χ^2 分布である（「カイじじょうブンプ」とよむ）。

サンプルの偏差二乗和を S，母集団の分散を σ^2 としたとき，χ^2 を次のように定義する。

$$\chi^2 = \frac{S}{\sigma^2} \tag{13.14}$$

$$S = \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 \tag{13.15}$$

先に発生した 12000 の乱数を用いて，サンプル数が $n = 3, 6$ の場合について χ^2 の分布をもとめる。

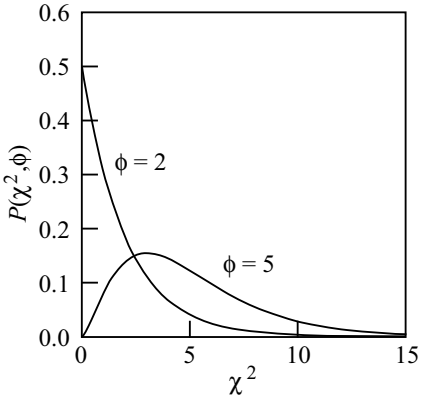
Excel 2007

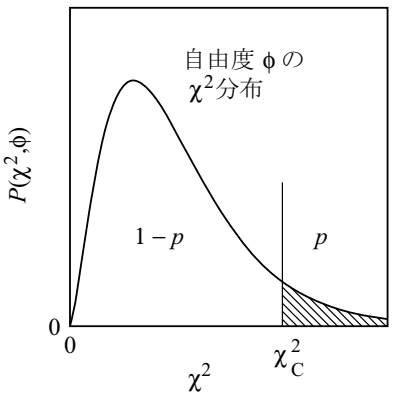
1. この章の始めで使ったワークシートを用いる。Web でダウンロードしたファイルには含まれていないので注意すること。
2. D5 に母集団の分散の理論値を計算「=1/(2∗LN(10))」。
3. K3 に 3 つのサンプルで χ^2 を計算「=DEVSQ(S3:U3)/$D$5」，K3 をコピーして K4:K1002 に貼り付け。
4. O3 に 6 つのサンプルで χ^2 を計算「=DEVSQ(S3:X3)/$D$5」，O3 をコピーして O4:O1002 に貼り付け。
5. 度数分布を求めるため，右のように入力する。ダウンロードしたファイルでは，すでに入力されている。
6. C34:C65 に 3 つのサンプルの χ^2 の度数分布「=FREQUENCY(K3:K1002,A34:A64)/COUNT(K3:K1002)/0.5」，Ctrl+Shift+Enter。
7. D34:D65 に 6 つのサンプルの χ^2 の度数分布「=FREQUENCY(O3:O1002,A34:A64)/COUNT(O3:O1002)/0.5」，Ctrl+Shift+Enter。
8. B33:D65 の四角い範囲を選択。
9. メニュー [挿入] → [グラフ] →「散布図」で，形式は右下折れ線のみを選択→ [完了]。

	A	B	C	D
32	chi^2分布			
33	範囲	値	phi=2	phi=5
34	0	0		
35	0.5	0.25		
36	1	0.75		
37	1.5	1.25		
38	2	1.75		
39	2.5	2.25		
40	3	2.75		
41	3.5	3.25		
42	4	3.75		
43	4.5	4.25		
44	5	4.75		
45	5.5	5.25		
46	6	5.75		
47	6.5	6.25		
48	7	6.75		
49	7.5	7.25		
50	8	7.75		
51	8.5	8.25		
52	9	8.75		
53	9.5	9.25		
54	10	9.75		
55	10.5	10.25		
56	11	10.75		
57	11.5	11.25		
58	12	11.75		
59	12.5	12.25		
60	13	12.75		
61	13.5	13.25		
62	14	13.75		
63	14.5	14.25		
64	15	14.75		
65		15		

この分布を χ^2 分布という。この分布は自由度 ϕ によって変化する。平均は ϕ，分散は 2ϕ である（実際に確認してみること）。

この χ^2 分布を用いて，母集団の分散に関する推定や検定を行うことは，これまでの内容を理解していれば，比較的簡単な応用問題である。ただし，分布が左右対称ではないことには十分に注意を払う必要がある。





Excel には，χ^2 と ϕ の値から上側確率 p を求める関数「CHIDIST」と，p と ϕ の値からそれに相当する χ^2 を求める関数「CHIINV」がある。「=CHIDIST(χ^2,ϕ)」，「=CHIINV(p,ϕ)」のように使用する。