12　推定と検定

実験値の精度や誤差を理解する上で重要な推定と検定の考え方を，前回と同様に乱数を用いたシミュレーションによって学ぶ。ここで説明するのは，母集団の分散が予めわかっている場合に用いられる方法で，「z 推定」，「z 検定」と呼ばれている方法である。

この節では主に次の書物を参考にした。

- 永田　靖，『入門統計解析法』，日科技連，東京，1992
- 菅　民郎，『Excel で学ぶ統計解析入門』，オーム社，東京，1999

12.1　不偏推定量 unbiased estimator

実験室における測定では，ホンの数個のデータから母集団の性質を推定する必要がある。自分が 3 回しか測定していないからといって，母集団の数が 3 だと思ってはいけない。ホントは何度でも繰り返せる測定を 3 回だけやってみたと思うべきである。では，母集団の性質を推定するのに適した量は何か？サンプリング操作を何度も繰り返したとき，推定値の平均が真の値になるようなものを，不偏推定量という。

ここでは，Excel で発生させた乱数を題材に，母集団の平均と分散に対する不偏推定値は何かを考える。

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。

$$X = 5 + \sqrt{-\log U_1}\cos(2\pi U_2) \tag{12.1}$$

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。ln は底が e の自然対数，log は底が 10 の常用対数としているので注意すること。ここから，サンプル数 $n = 3, 6, 11$ の 3 種類の標本 x_i を繰り返しサンプリングする。

Excel 2007

1. 乱数の準備
   (i) Web ページから不偏推定量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			n=6			n=11					
2	行番号	平均	分散	不偏分散	平均	分散	不偏分散	平均	分散	不偏分散			乱数
3	1												5.756375
4	2												

   (ii) ワークシート名を「huhen」から「不偏推定量」に変更する。
   (iii) 課題提出用のファイル名で，Excel 形式で保存する。
   (iv) S3 に「=5+SQRT(−LOG(RAND()))∗COS(2∗PI()∗RAND())」と入力。
   (v) 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。
2. サンプリング
   (i) G3:G1002 には前回と同様の操作で 1〜1000 まで番号をつける。
   (ii) H3 に「=AVERAGE(S3:U3)」(3 つのサンプルの平均)。
   (iii) I3 に「=VARP(S3:U3)」(3 つのサンプルの標本分散)。
   (iv) J3 に「=VAR(S3:U3)」(3 つのサンプルの不偏分散)。
   (v) 同様に 6 個のサンプルについて K3 に「=AVERAGE(S3:X3)」，L3 に「=VARP(S3:X3)」，M3 に「=VAR(S3:X3)」。
   (vi) 同様に 11 個のサンプルについて N3 に「=AVERAGE(S3:AC3)」，O3 に「=VARP(S3:AC3)」，P3 に「=VAR(S3:AC3)」。
   (vii) H3:P3 をコピーし H4:P1002 に貼り付け。

用語を確認しておく。この授業では「標本分散」は次の量である。

$$V = \frac{1}{n}\sum_{n}^{i=1}(x_i - \langle x\rangle)^2 \tag{12.2}$$

そして，「不偏分散」は次の量である。

$$V = \frac{1}{n-1}\sum_{n}^{i=1}(x_i - \langle x\rangle)^2 \tag{12.3}$$

ただし，不偏分散のことを単に「分散」と言うことも多い。n が非常に大きいとき（例えば 1000），標本分散と不偏分散の違いは問題にならない。しかし，通常の測定操作では n が小さい（例えば $n = 3, 6, 11$）ので，両者の違いが問題になる。

標本分散や不偏分散を求めるには平均 $\langle x\rangle$ が決まっていなければならない。n 個のデータは n 個すべてが様々な値をとりうるとき，「自由度」が n であるという。平均 $\langle x\rangle$ が決まっていて固定されている状態では，n 個のデータは n 個すべてが自由に変化しうる変数ではない。$(n-1)$ 個のデータが決まれば，それらと平均 $\langle x\rangle$ の値から，残り 1 個のデータは必然的に決定されてしまう。つまり，偏差 $(x_i - \langle x\rangle)$ を計算する際には $(n-1)$ 個のデータのみが自由に値をとりうるので，サンプル数は n でも自由度は $(n-1)$ になる。ここで言う「標本分散」は，偏差二乗和をデータ数 n で割った値であり，「不偏分散」は偏差二乗和を自由度 $\phi = n - 1$ で割った値である。

では，不偏分散は何故「不偏」分散なのか？先ほど求めた $n = 3, 6, 11$ のサンプリングを 1000 回繰り返したデータを具体例として考えてみる。前回考えた「平均の分散」と，今回の「分散の平均」を取り違えないこと。

Excle

先ほどのワークシートで次のような表を作る。ダウンロードしたファイルでは，すでに入力されている。

	A	B	C	D	E
1					
2					
3		母集団			
4	n	12000	3	6	11
5	自由度				
6	平均				
7	分散				
8	不偏分散				
9					
10	理想値	母集団			
11	平均				
12	分散				

1. C5 に「=C4−1」と入力。C5 をコピーして，D5, E5 に貼付け
2. B6 に「=AVERAGE(S3:AD1002)」（母集団の平均）。
3. B7 に「=VARP(S3:AD1002)」（母集団の標本分散）。
4. B8 に「=VAR(S3:AD1002)」（母集団の不偏分散）。
5. C6 に「=AVERAGE(H3:H1002)」（3 つの平均の平均）。
6. C7 に「=AVERAGE(I3:I1002)」（3 つの標本分散の平均）。
7. C8 に「=AVERAGE(J3:J1002)」（3 つの不偏分散の平均）。
8. D6 に「=AVERAGE(K3:K1002)」（6 つの平均の平均）。
9. D7 に「=AVERAGE(L3:L1002)」（6 つの標本分散の平均）。
10. D8 に「=AVERAGE(M3:M1002)」（6 つの不偏分散の平均）。
11. E6 に「=AVERAGE(N3:N1002)」（11 の平均の平均）。
12. E7 に「=AVERAGE(O3:O1002)」（11 の標本分散の平均）。
13. E8 に「=AVERAGE(P3:P1002)」（11 の不偏分散の平均）。
14. B11 に「=5」（母集団の平均の理想値）。B12 に「=1/(2∗LN(10))」（母集団の分散の理想値）。

$n = 3, 6, 11$ のいずれの場合も，平均の平均は 5 付近の値をとり，母集団の平均とほぼ一致する。つまり，少

数サンプルの平均は母集団の平均に対する不偏推定量になっている。

平均とは異なり，標本分散の平均はサンプル数によって明らかに異なる値をとる。つまり，少数サンプルの標本分散は母集団の分散に対する不偏推定量になっていない。しかし，不偏分散の平均についてみれば，$n = 3, 5, 11$ のいずれの場合も，母集団の分散（および不偏分散）とほぼ等しい値をとる。つまり，少数サンプルの不偏分散は母集団の分散に対する不偏推定量になっている。

12.2 点推定 point estimation

前節の結果から明らかなように，母集団の平均を推定したければ少数サンプルの平均をとればよい。また，母集団の分散を推定したければ，少数サンプルの不偏分散をとればよい。

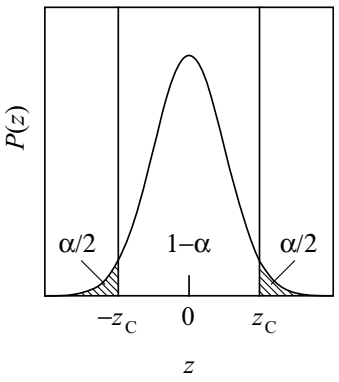
このように，平均や分散といった統計量をひとつの値として推定する操作を「点推定」という。

12.3 区間推定 interval estimation

次に，母集団の平均の存在範囲を区間として推定する「区間推定」について考える。ここでは，母集団の分散が予め σ^2 であるとわかっていると仮定する。n 回の測定を行い，その標本平均 $\langle x \rangle$ を得たとする。この測定は無限に繰り返すことのできるサンプリング操作のうちのただ 1 つの例である。このようなサンプリングを繰り返したときの平均値の分布については前回学んだ。それは，母集団の平均を μ，分散を σ^2 としたときに，サンプル数 n の平均 $\langle x \rangle$ から次のような量 z を求めれば，z は平均 0，分散 1 の正規分布にしたがうというものである。

$$z = \frac{\langle x \rangle - \mu}{\sqrt{\sigma^2/n}} \tag{12.4}$$

n 個のサンプリングを繰り返して行ったとき，z がある正の値 z_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 $1 - (\alpha/2) = 0.975$ ）。z_C の値は，標準正規分布の性質から求めることができる（具体的には $1 - (\alpha/2) = 0.975$ なら $z_C = 1.960$ ）。このとき，z が z_C 以上になる確率は $\alpha/2$ である。さらに，z が $-z_C$ 以下になる確率も $\alpha/2$ である。つまり，$-z_C \le z \le z_C$ 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。



$$P(-z_C \le z \le z_C) = 1 - \alpha \tag{12.5}$$

次に，カッコの中のみを変形する

$$-z_C \le z \le z_C \tag{12.6}$$

$$-z_C \le \frac{\langle x \rangle - \mu}{\sqrt{\sigma^2/n}} \le z_C \tag{12.7}$$

$$\langle x \rangle - z_C \sqrt{\frac{\sigma^2}{n}} \le \mu \le \langle x \rangle + z_C \sqrt{\frac{\sigma^2}{n}} \tag{12.8}$$

確率の形で書くと次のようになる。

$$P\left(\langle x \rangle - z_C \sqrt{\sigma^2/n} \le \mu \le \langle x \rangle + z_C \sqrt{\sigma^2/n}\right) = 1 - \alpha \tag{12.9}$$

この式の意味するところをは何か。

母集団の分散が σ^2 であることがわかっている場合について，サンプル数 n の測定を行って平均値 $\langle x \rangle$ を得た。そして，信頼区間 $\langle x \rangle - z_C \sqrt{\sigma^2/n} \le \mu \le \langle x \rangle + z_C \sqrt{\sigma^2/n}$ 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，z_C の値は，標準正規分布の性質から求める。

このとき，信頼区間の境界 $\langle x\rangle - z_C\sqrt{\sigma^2/n}$ および $\langle x\rangle + z_C\sqrt{\sigma^2/n}$ のことを，その信頼率における「信頼限界」confidence limit という。信頼率には 0.95 が用いられることが多いが，もちろん，目的によって，0.90 でも 0.99 でも適当な値を用いればよい。

12.4 区間推定の具体例

今回の乱数を測定データとみなして具体例を計算してみる。母集団の分散は $\sigma^2 = 1/(2\ln 10) = 0.217147\cdots$ であると予めわかっているとする。

Excel 2007

1. Web ページから推定・検定説明用ファイルをダウンロードし，Excel で開く。

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	データ											
3	5.160588	4.80774	5.169312	5.081859	5.254435	4.635186	5.24188	5.101463	4.618777	5.379715	5.050192	5.41736
4												
5	平均値	不偏分散	サンプル数	自由度								
6												
7												
8	z推定					t推定						
9	母集団分散	標本平均分散										
10												
11	alpha(a)	信頼率	1-(a/2)	Zc		alpha(a)	信頼率	tc				
12												
13												
14	下限	平均	上限			下限	平均	上限				
15												
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												

2. 「ホーム」→「セル」→「書式」→「ワークシートの移動またはコピー」で，「不偏推定量」と同じファイルにワークシートを移動する。
3. ワークシート名を「kenntei」から「推定・検定」に変更する。
4. 準備
   (i) A3:L3 に先ほどと同じ式に基づく乱数が表示されている。
   (ii) A6 に 12 個の乱数の平均を計算する。「=AVERAGE(A3:L3)」
   (iii) B6 に 12 個の乱数の不偏分散を計算する。「=VAR(A3:L3)」
   (iv) C6 に 12 個の乱数のデータ数を表示する。「=COUNT(A3:L3)」
   (v) D6 に 12 個の乱数の自由度を表示する。「=C6−1」
   (vi) A10 に乱数の母集団分散（理論値）を計算する。「=1/(2∗LN(10))」
   (vii) B10 に乱数の標本平均の分散（理論値）を計算する。「=A10/C6」
5. 信頼区間の計算
   (i) A12 に α の値を入力「0.05」。
   (ii) B12 に信頼率を求める「=1−A12」。C12 に $1-(\alpha/2)$ を計算。「=1−A12/2」
   (iii) D12 に z_C を計算。「=NORMSINV(C12)」
   [Excel 関数の説明]「NORMSINV」は，ある値（ここでは 0.975）が与えられたときに，標準正規分布で $-\infty$ から z_C までの累積確率 $P(z \le z_C)$ の値が p となるような z_C の値を返す関数。引数は累積確率の値のひとつだけ。
   (iv) A15 に区間の下限を計算。「=A6−D12∗SQRT(B10)」
   (v) B15 に区間の中心を計算。「=A6」
   (vi) C15 に区間の上限を計算。「=A6+D12∗SQRT(B10)」
6. 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。

12.5 検定 test

次に，n 回の測定を行って平均値 $\langle x \rangle$ を得たときに，「母集団の平均値がある値 $\mu = m$ をとる」という仮説が統計的に妥当であるといえるかどうかという問題を考える。このような作業は「検定」とよばれる。ここでも，母集団の分散は予め σ^2 であるとわかっていると仮定する。

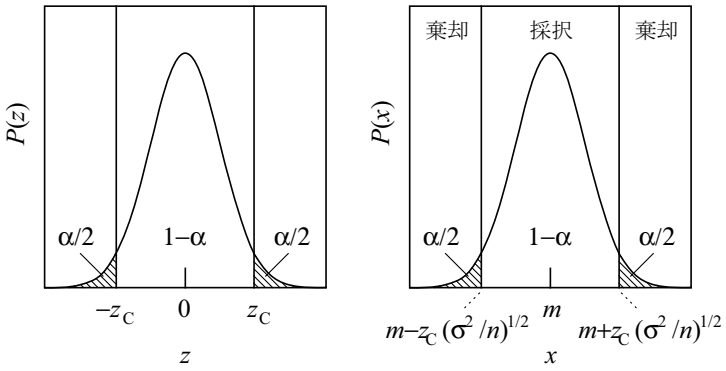
先ほどの区間推定の場合と同じように z を定義する。また，ある信頼率の下で，同様に z_C を定義する。母集団の平均値が $\mu = m$ であるという仮定がある信頼率の下で正しければ，z はその信頼率の確率で次の区間に含まれるはずである。

$$-z_C \le z \le z_C \tag{12.10}$$

$$-z_C \le \frac{\langle x \rangle - m}{\sqrt{\sigma^2/n}} \le z_C \tag{12.11}$$

$$m - z_C \sqrt{\frac{\sigma^2}{n}} \le \langle x \rangle \le m + z_C \sqrt{\frac{\sigma^2}{n}} \tag{12.12}$$

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。はじめに立てた仮説を「無帰仮説」null hypothesis，それを否定する仮説を「対立仮説」alternative hypothesis という。また，α のことを有意水準 level of significance という。



12.6 検定の具体例

区間推定で用いたデータで検定の具体例を示す。

Excel 2007

1. 先ほどのワークシートの次の部分を利用する。

	A	B	C
17	z検定		
18	無帰仮説		
19	平均値	5	
20	下限	上限	判定
21			

2. B19 に無帰仮説でたてた平均値「5」を入力。
3. A21 に区間の下限を計算。「=B19−D12∗SQRT(B10)」
4. B21 に区間の上限を計算。「=B19+D12∗SQRT(B10)」
5. C21 に判定結果を○×で表示する。「=IF(A6<A21,"×",IF(A6>B21,"×","○"))」
6. 再計算を繰り返せば，時々判定が×になる

[Excel 関数の説明] 「IF」にはコンマで区切られた 3 つの引数がある。1 番目は条件，2 番目は条件が成り立つ場合の処理（数式等），3 番目は条件が成り立たないときの処理。ここでは文字列を表示させるのが処理の中身で，その場合，文字列をダブルクォーテーションマーク「 " ··· " 」でくくる。ここでは，

「IF」を入れ子にして用いており，はじめの条件が満たされない場合，次の条件でさらに 2 つの場合に分けている。じっくり見て考えてください。

当然のことながら，このような検定では，本当は無帰仮説が成り立っているのにそれを否定してしまうような誤りと，本当は無帰仮説が成り立っていないのにそれを肯定してしまうような誤りとが生じる。はじめのものは第一種の過誤，あとのものは第二種の過誤と一般に呼ばれている。