(15. Juli 2004)

このテキストは，

http://math.cs.kitami-it.ac.jp/~fuchino/chubu/statistics-04s-misc.pdf

に置いてあるものと同じものです．また，講義の web ページ

http://math.cs.kitami-it.ac.jp/~fuchino/chubu/statistics-04s.html

にもリンクしてあります．

1　母比率の推定

例題 1 ある食品の中から，100 個を無作為に抽出して，その重量を測定したところ，10 個が規格外れであった．この食品の不良率を信頼係数 95% で推定せよ．

多数の個体からなる全体の統計量 A を知るために，全体のごく一部についてのデータをとって，この測定値から全体での値を結論する方法を**推定** という．推定したい統計量を持つ全体のことを**母集団**とよび，データを測定するために抽出した母集団の一部を**標本**とよぶ．標本が n 個の個体からなるとき，この標本の**サイズ**は n であるという．また標本をできるだけデータの偏りのないようにとることを，この標本を**無作為抽出**するという．推定では母集団の統計量 A を直接測定するわけではないので，その結論は，「この統計量 A が区間 $[a, b]$ に含まれる値になる確率は $\mu\%$ である.」というような形式になることが多い，“μ %” としては 95% または 99% (つまり 0.95 または 0.99) の値がとられることが多い，このとき,「統計量の A の**信頼係数** $\mu\%$ での**推定区間**は $[a, b]$ である」という言い方をする．

以下では，ある母集団で，ある性質（特性）A を持つ個体の母集団全体に対する比率が p であるとき，この p を信頼係数 $1-\alpha$ で推定することを考える．上で述べたように，α としては 0.05, 0.01 などの値を採用することが多い．大きさが n の標本をこの母集団から抽出して，標本の k 番目の要素が A を持つなら 1 そうでないなら 0 をとるような確率変数を X_k とすれば，$E(X_k) = 0\cdot(1-p)+1\cdot p = p, V(X_k) = (0-p)^2(1-p)+(1-p)^2p = p - p^2 = p(1-p)$ となる．　（理想的な標本では） X_k, $k = 1, 2, \ldots, n$ は互いに独立となると考えられる．したがって，標本の要素のうち特性 A を持つものの数を与える確率変数 $X = X_1 + X_2 + \cdots + X_n$ を考えると，中心極限定理により，n が十分に大きいときには，X/n の分布は $N(p, p(1-p)/n)$ で近似できる．そこで，以下では，X/n が $N(p, p(1-p)/n)$ に従うとして，議論することにすると，

$$Z = \frac{X/n - p}{\sqrt{p(1-p)/n}}$$

は標準正規分布 $N(0,1)$ に従うから，$P(|Z|<\lambda)<1-\alpha$ となる λ に対し[1]，

$$\begin{aligned}|Z|<\lambda &\Leftrightarrow -\lambda \ < \ Z \ < \ \lambda \\ &\Leftrightarrow -\lambda\sqrt{p(1-p)/n} \ < \ X/n-p \ < \ \lambda\sqrt{p(1-p)/n} \\ &\Leftrightarrow p-\lambda\sqrt{p(1-p)/n} \ < \ X/n \ < \ p+\lambda\sqrt{p(1-p)/n}\end{aligned}$$

となる．X/n の期待値は p だから，X/n を p で置き換え，標本中の A を持つ個体の個数の実測値を m として，p を $p^*=m/n$ で置き換えると，

$$p^*-\lambda\sqrt{p^*(1-p^*)/n} \ < \ p \ < \ p^*+\lambda\sqrt{p^*(1-p^*)/n}$$

となる．したがって，区間

$$\left[p^*-\lambda\sqrt{p^*(1-p^*)/n},\ p^*+\lambda\sqrt{p^*(1-p^*)/n}\ \right]$$

が信頼係数 $1-\alpha$ での p の推定区間と考えられる．

1.1　例題 1 の解説

以上の考察を例題 1 に適用する： この問題では

$$p^*=\frac{1\not{0}}{10\not{0}}=\frac{1}{10}$$

となるから，信頼係数 95% $(1-0.05)$ での不良率の推定区間は，

$$\left[\ \frac{1}{10}-1.96\times\sqrt{\frac{1}{10}\times\frac{9}{10}/100},\ \frac{1}{10}-1.96\times\sqrt{\frac{1}{10}\times\frac{9}{10}/100}\ \right]=[0.0412,\ 0.1588]$$

となる．

2　母比率の検定

例題 2 （教科書 p.88 の例 5.10） ある選挙区で無作為に抽出した有権者 300 人について調査したところ，165 人が A 候補者を支持していた．このとき A 候補者は過半数支持を受けていると言えるか？ 有意水準 5% で検定せよ．

仮説検定は，ある仮定 H_0 のもとに，確率を計算し，その結果測定された実測値がきわめて低い確率 $p=1-\alpha$ を持つ事象を与える値の領域 W に属すことを示し，このことから，この仮定 H_0 がほぼ誤りと言えることを結論し，この仮定 H_0 の否定 H_1 がほぼ成り立っていると推測する論法である．(あるいは，実測値がこのような領域 W に属さないことを見て，H_0 は必ずしも否定できないことを結論する．）推定でと同じように α としては，0.05 または 0.01 などの値が用いられることが多い．
H_0 のことを **帰無仮説** （きむかせつ）とよび H_1 を **対立仮説** とよぶ．また W を **破却領域** （はきゃくりょういき）といい，α を **有意水準**という．実測値から得られた値が

[1] Z が標準正規分布に従うことから，$\alpha=0.05$ なら，λ はおよそ 1.96 となり，$\alpha=0.01$ なら， λ はおよそ 2.58 である．

W に属すことが示せたときには，帰無仮説 H_0 は有意水準 α で破却（はきゃく）される，と言い，そうでなかったときには，帰無仮説 H_0 は有意水準 α で採択（さいたく）されるという．
母集団で，ある特性 A を持つ個体の割合 p に関する検定を考える．帰無仮説 H_0 として，$p = p_0$ である，という仮定をたてて考えてみる．今，サイズが n の標本の中で A を持つものの数を与える確率変数 X を考えると，前節と同様に，

$$Z = \frac{X/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

は標準正規分布 $N(0,1)$ に従うと考えてよいから，有意水準 α の破却領域は，$P(|Z| > \lambda) = \alpha$ となるような λ に対し，$|Z| > \lambda$ を与えるような X の値の領域となる．
このことから，x_0 を実際の標本での X の測定値とするとき，

$$z_0 = \frac{x_0/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

として，$|z_0| > \lambda$ なら H_0 は有意水準 α で破却され，$|z_0| \leq \lambda$ なら，H_0 は有意水準 α で採択される，とすればよいことがわかる．また，$z_0 > \lambda$ のときには，対立仮説 H_0 として，$p > p_0$ を採択することができ，$z_0 < \lambda$ のときには $p < p_0$ を採択できる．

2.1 例題 2 の解説

上の例題 2 では，“p_0 が $1/2$ ある”という仮定を H_0 として，標本の実測値からの Z の計算結果を z_0 とすると，

$$z_0 = \frac{165/300 - 1/2}{\sqrt{1/2 \cdot (1-1/2) \cdot 1/300}} = 1.7320508075688787\cdots$$

となるから，$|z_0| < 1.96$ となり，H_0 つまり，この候補者が過半数の有権者によって支持されているという仮定は有意水準 5%で採択される．

上と同じ方法によって，例題 2 とは若干タイプの異なる次のような検定を行なうこともできる：

例題 3 2000 年の国勢調査によると，全国の就業者の 29.1% はサービス業に従事している．同年 S 大学の卒業生で就業した者のうち 200 人を無作為抽出してアンケートをとったところ，サービス業に就職したのはこのうちの 36.5% だった．これは全国平均より高率と言えるか？ 有意水準 0.05 で検定せよ．

2.2 例題 3 の解説

この例題では，就業者全体のうち “サービス業に従事している” という特性を持つ個体の割合は確定している．ここで問われているのは，この値が就業者全体の部分集団である “S 大学の卒業生” という母集団でも採択できるかどうかである．
したがって，“この母集団でサービス業に就職した者の割合が 29.1% 以下である” という仮定が H_0 となる．上と同じ記号を用いることにして，

$$z_0 = \frac{0.365 - 0.291}{\sqrt{0.291(1-0.291)/200}} = 2.30397151534653\cdots > 1.96$$

となるから，H_0 が破却され，S 大生でサービス業に就職した者の割合は全国平均より高いことが，有意水準 0.05 で言えることがわかった．ただし，この場合，有意水準 0.01 で検定すると，$|z_0| < 2.58$ だから， H_0 は破却できない．このことからも分るようにサービス業に就職した者の割合が全国平均より高率とも言いきれない可能性も，捨てきれるわけではない．