統計特講III

ー基礎統計 classic

S1-39(08)基礎統計科編

2009

夏期

大 1・2

S13909

まえがき

◇統計という学問は、人に教わったことをそのまま覚えるという事では、けっして身に付きません。統計を自分のものにするためには、自分で必死こいて考えなければなりません。そのさい考えるということは、何をすることでしょうか。それは第 1 に状況を生き生きと思い浮かべることです。何がどうなってどんな風に推測されるのか、等を頭のなかで実験　―思考実験―　してみることです。そして第 2 には、その状況と統計的推測を支配している法則が何なのかをしっかりと見定めることです。そして第 3 に、その一般的な法則を具体的なケースに即して書き下さなければなりません。第 4 に必要な計算をやり切って、その結果が統計的に納得のゆくものであるかどうかを吟味することです。

◇『基本事項』『例題』『演習問題』の 3 部構成になっています。
『基本事項』は比較的丁寧に書き下されていますので、統計的に何を示しているかを考えてながら、手を動かしつつ読んで下さい。目で追ってはいけません。手で追ってください。
『例題』は基本事項の意味や考え方、法則・公式の使い方の練習になるようなものを選びました。

◇教科書の第何章に対応するかも書かれているので、教科書を横において勉強していただければ、統計の理解はより深まります。

注：このシケプリは廣松先生(退官)と倉田先生(木 5)の過去問をベースに作られています。安藤先生(水 1)と藤縄先生(月 2)と小林先生(金 5)のテストはものすごい論述を書かせたりするかもしれません。その辺は、各自対応していただきたいと思います。

※安藤先生の基礎統計の過去問は見当たらない。レポートが出ているらしい。
藤澤先生は基礎統計を担当するのは今年が初めて。
小林先生は自身の web ページで過去問を 1 年分公開している。見た感じ、そこまでずれてなかった印象。各自ご確認を。

もくじ

第一講 統計学の基礎知識

教科書:第一章~第二章

§1.1　統計学の教養(読み流してよい)

・記述統計と推測統計

記述統計では，母集団全てを調べて得た資料を要約して特徴的な各種数値を算出し，それより分析対象の性質を記述する．推測統計では，標本のみを観察し，これから確率論を用いて母集団全体の特徴を推測する．

・全数調査(センサス)と標本調査

全数調査とは母集団に属する全ての単位について調べること．国勢調査がこの例．一方，標本調査とは母集団から標本を抽出してそれらを観察すること．但し，標本が‘正しく’選ばれていなければ、正しい考察を得られない．

・量的データと質的データ

量的データとは、調査結果が数値で表現しうるデータのこと。一方質的データとは、『男・女』のように、数値で表現できず、どのカテゴリーに属しているか、どんな状態にあるか、をだけがわかるデータのこと。

・時系列データと横断面(クロス・セクション)データ

時系列データとは同一の対象に対して，異なる時点での観測値からなるデータ．横断面データとは異なる対象の同時点の観測値からなるデータ．

・原データと統計資料

原データとは、実験や調査をして得られた『生のデータ』である。それに対し統計資料とは、原データに対し、何らかの統計処理を施した後のデータのことである。

・統計資料のうち、第一義統計(調査統計)と第二義統計(業務統計)

統計資料のうち、『あらかじめ統計資料を作成する目的で調査を行って、その結果として得られたデータを集計したもの、またはそのデータそのもの』を第一義統計という。一方『統計資料の作成を目的とせずに得られたデータ』を集計し得られた資料を第二義統計という。

・一次統計と二次統計

一次統計とは、統計資料のことである。一方二次統計とは、統計資料をさらに加工して得られた統計資料のことである。

※時系列データを扱う際の注意----------

(1) 期間の取り方の代表例

年 annual/半年/四半期 quarterly/月 monthly/旬/日 daily/時/分/秒[1]

(2) 暦年 Calendar Year と会計年度 Fiscal Year

・暦年(CY)　…1/1~12/31
・会計年度(FY)…4/1~3/31

CY で統計を整理しているのか、FY で整理しているのかをはっきりさせなければならない。

注) 日本以外の国の統計資料と日本のを比較する際、会計年度での時系列データは慎重に取り扱う。国によって『年度の定義』が異なるからである。

日本　…4/1~3/31　　2008 年 10 月 23 日は
アメリカ…10/1~9/30　　日本では 2008 年度、米国では 2009 年度

例題 1. 統計のウソ

「統計のウソ」という言葉について論ぜよ。　(廣松)

解答

統計調査をおこなって得られたデータをそのまま鵜呑みにするのは、危険である。そもそも、データの定義が、統計を取る側の人間にとって有利となるように定義づけられている可能性もある。また、標本に基づく統計では、その選び方によって同じ母集団でも統計数値は大きく変わりうる。標本抽出に作為をいれれば統計で人をだますのはカンタンであるし、データを発表する側が、自分らにとって都合のいいデータだけを発表することでも人を欺くことができる.

[コメント]

データの定義、統計手法の選択、結果の表現のうち１つ以上にトリックがある場合が多い。

作為抽出の面白い例として，世論調査の一手法があげられる。『お昼ごろに電話帳からランダムにサンプリングして電話をかけ回答を求める』という RDD 方式という手法があるが、お昼頃といったら電話に出るのは基本的に主婦であるため、偏った資料になってしまう恐れがある。このように、無作為抽出だと思いながらも実際は作為抽出となることもある。

≪終わり≫

[1] 金融工学など

§1.2　1 次元のデータ

1 次元のデータ

…異なる対象に対して、ある 1 種類の数値について調査したデータ

$$(x_1, x_2, \cdots, x_i, \cdots, x_n)$$

ここでnは『データの大きさ・サイズ』という。

1.2.1　度数分布とヒストグラム

度数分布表

第i階級	階級値v_i	度数f_i	相対度数	累積度数	累積相対度数
第 1 階級(0 点以上 10 点未満)	5	12	0.032	12	0.032
第 2 階級(10 点以上 20 点未満)	15	10	0.027	22	0.059
第 3 階級(20 点以上 30 点未満)	25	19	0.051	41	0.110
第 4 階級(30 点以上 40 点未満)	35	42	0.113	83	0.223
第 5 階級(40 点以上 50 点未満)	45	72	0.193	155	0.416
第 6 階級(50 点以上 60 点未満)	55	82	0.220	237	0.635
第 7 階級(60 点以上 70 点未満)	65	54	0.145	291	0.780
第 8 階級(70 点以上 80 点未満)	75	38	0.102	329	0.882
第 9 階級(80 点以上 90 点未満)	85	25	0.067	354	0.949
第 10 階級(90 点以上 100 点未満)	95	19	0.051	373	1.000
合計		373	1.000		

上のように、いくつかの階級を設定する。

・階級値v_i…その階級の代表値。一般的に、階級の両端の値の平均。

・度数f_i…それぞれの階級に観測値がいくつあるか。$\sum f_i = n$

・第i階級の相対度数…$f_i \Big/ \sum_i f_i = \dfrac{f_i}{n}$

・第i階級の累積度数…$\sum_{j \leqq i} f_j$

・第i階級の累積相対度数…$\sum_{j \leqq i} \left(\dfrac{f_j}{n} \right)$

ヒストグラムとは、度数分布表を棒グラフにしたものである。

※階級数・階級幅の決定について----------

階級をどのようにとるかについて特に決まったルールはないが、できればヒストグラムがきれいな形になるように階級数・階級幅を設定したい。そのような階級数kを与えてくれる経験則がある。

$$k \fallingdotseq 1 + log_2 n$$

これを『スタージェスの公式』という(繰り返すが経験則である)。

階級幅については、一般的にはどの階級も等しい階級幅であることが望まれるので、上で求めたkを用いて、観測値が取りうる範囲をk等分すればよい。その際、上限と下限には区切りのよい値にする。統計の種類(所得・敷地面積etc)によっては、上限・下限が設定できないものや、階級幅が一定でないものもあるが、『基礎統計』の範囲内では無視して良い。(参:教科書 P.19~26)

ヒストグラムの形

・単峰型…分布の山が、一つである分布
・右に歪んだ分布…分布の山が左に寄っていて、右側に長い裾を引いている分布
・双峰型…ヒストグラムの山が二つある分布

> 双峰型の場合、観測対象の集団には『性質が異なる 2 集団が混在している』ことが多い。正しい結果と考察を得るには、度数分布表を作る前にこの 2 集団を分別しておく必要がある。この“分別”を『層別』という。

ローレンツ曲線については P.26~27 を参照。

1.2.2　平均・メディアン・モード

原データ$(x_1, x_2, \cdots, x_i, \cdots, x_n)$がある場合(簡単のため$x_1 < x_2 < \cdots < x_n$とする)

(1) 平均(mean)$\mu\left(\bar{x}\text{などとも表記する}\right)$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(2) メディアン・中央値(median)Me

(Me 以下のデータ数) = (Me 以上のデータ数)

をみたすような数がMeとなる

データのサイズnが奇数$(2m + 1)$のときは小さい方から$(m + 1)$番目の数x_{m+1}

偶数$(2m)$のときは小さい方からm番目と$(m + 1)$番目の数の平均$\frac{x_m + x_{m+1}}{2}$

とする

(3) モード・最頻値(mode) Mo

一番多く登場する数(モードは度数分布表がある場合に考えるのが普通)

度数分布表がある場合

第i階級	階級値v_i	度数f_i

(データのサイズをnとする)

(1')平均

$$\mu = \frac{1}{n}\sum_i f_i v_i$$

(2')メディアンMe…初めて相対累積度数が 0.5 を超える階級の階級値

(3')モードMo…度数が最大の階級の階級値

※モードは層別ができていない集団においては適切な値を与えない。

※四分位点$\mathrm{Q}_1, \mathrm{Q}_3$

…メディアンMeの考え方を応用したもの。Meは観測値を小さい順に数えて 50%のところにある数値であった。この『50%』を『25%』『75%』に置き換えた点がそれぞれ$\mathrm{Q}_1, \mathrm{Q}_3$である。

例題 2. 平均の最小二乗性

$$f(c) = \sum_{i=1}^{n}(x_i - c)^2$$

を最小とするcを求めよ。 (倉田)

解答

$$f' = \sum_{i=1}^{n}(-2)(x_i - c) = -2\sum_{i=1}^{n} x_i + 2nc$$

$$f' = 0 \quad \Leftrightarrow \quad c = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$f'' = 2n > 0$ より$c = \bar{x}$で最小となる

≪終わり≫

1.2.3　散らばりを表す尺度

データサイズをnとする

(1)平均偏差[2]

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\mu|$$

…『各観測値が平均からどれだけ離れているか』の平均

(2) 分散σ^2

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}({x_i}^2-2\mu x_i+\mu^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\mu^2$$

$$= \left(\text{二乗和の平均}\right) - (\text{平均の二乗})$$

度数分布で与えられている時は、

$$\sigma^2 := \frac{1}{\mathrm{n}}\sum_{i=1}^{n}f_i(v_i-\mu)^2 = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}f_i{v_i}^2-\mu^2$$

(3)標準偏差$\sigma = \sqrt{\text{分散}\sigma^2}$

標準偏差は、原データと次元が同じ

※四分位偏差Q

$$\mathrm{Q} = (\mathrm{Q_3}-\mathrm{Q_1}) \div 2$$

両端 4 分の 1 を切り落としているので、外れ値(極端な値)の影響を受けにくい

[2] $\sum(x_i-\mu) = 0$

※箱ヒゲ図----------

分布の散らばりや歪み具合を視覚的に表すのに効果的な『箱ヒゲ図』というものがある。
統計資料ではよく利用されるので一応紹介しておく。

例題 3. 箱ヒゲ図

以下のデータはある会社の社員 36 人の血糖値のデータである。これについて，箱ヒゲ図を描け。

109,89,93,120,75,103,98,106,136,112,97,80
107,78,95,80162,116,93,88,102,90,96,100
48,70,99,115,93,82,112,57,94,88,105,102 (廣松)

解答

36÷4=9 である．大小順に並べ替えると

48,57,70,75,78,80,80,82,88,88,89,90,
93,93,93,94,95,96,97,98,99,100,102,102
103,105,106,107,109,112,112,115,116,120,136,162

第一四分位点 $Q_1 = 88$(下から 9 番目) [3]　　第三四分位点 $Q_3 = 107$(上から 9 番目)

メディアン $Me = (96 + 97) \div 2 = 96.5$　　四分位偏差 $Q = (Q_3 - Q_1) \div 2 = 9.5$

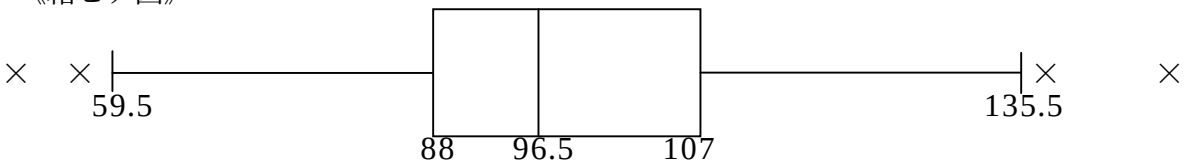
$Q_1 - 3Q = 88 - 9.5 \times 3 = 59.5$(この値を下ヒンジという)

$Q_3 + 3Q = 107 + 9.5 \times 3 = 135.5$(この値を上ヒンジという)

[箱ヒゲ図の描き方]

①第一四分位点Q_1から第三四分位点Q_3までを一辺とする長方形を描く。
②メディアンMeの位置に線を入れる。
③第一四分位点Q_1から下ヒンジまでヒゲを伸ばす
④第三四分位点Q_3から上ヒンジまでヒゲを伸ばす
⑤下ヒンジ以下 or 上ヒンジ以上の値を外れ値として、その位置に×マークをつける。
⑥完成。

《箱ヒゲ図》



≪終わり≫

---

[3] 四分位点の取り方は人それぞれなので、下(上)から何番目の値かを明記しておく。

1.2.4　変動係数と変数変換

・変動係数

$$\text{C.V.} = \frac{\sigma}{\mu}$$

分布の中心が著しく異なる場合は、分散/標準偏差で、散らばり具合を比較することは適切でない。したがって、平均の違いも考慮にいれたこの変動係数(無次元)を用いる。

・観測データの(一次)変換

$$(x_1, x_2, \cdots, x_i, \cdots, x_n)$$

に対し、

$$z_i = ax_i + b$$

という変換を行うとき、z_i の平均 $\bar{z}$,分散$S_z{}^2$,標準偏差S_zを、 x_i の平均 $\bar{x}$,分散 $S_x{}^2$標準偏差S_xで表すと、

$$\bar{z} = a\bar{x} + b$$
$$S_z{}^2 = a^2 S_x{}^2$$
$$S_z = |a| S_x$$

となる。(定義式から計算するだけ)

特に、

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

と変換すると

$$\bar{z} = 0 \qquad S_z = 1$$

となる。この変換を標準化という。
さらに

$$\mathrm{T_i} = 10z_i + 50$$

とすると、これは偏差値を表す。平均 50、標準偏差 10 である。

第二講 相関と回帰

教科書:第三章

一般には多次元で扱うのだが、基礎統計の講義内では 2 次元を扱う。
いま以下のようなデータサイズnの 2 次元データが得られた。

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

xとyの関係を調べるときには、2 つの視点が存在する。

①相関…xとy の間に区別を設け、対等に見る。xとy の相互関係を扱う。
②回帰…x がy を決める様子を扱う(一方通行)。またはその逆。

§2.1　相関

2.1.1　相関とは

サイズnの 2 次元データ$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$を$xy$ 平面にプロットした図を散布図又は相関図という。以下の 4 つのような図を考える。

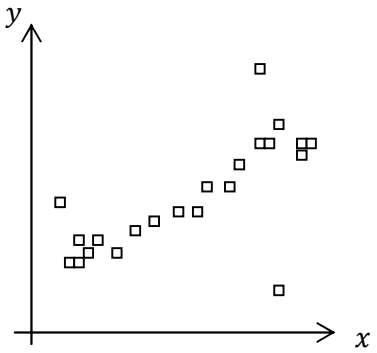


図 1

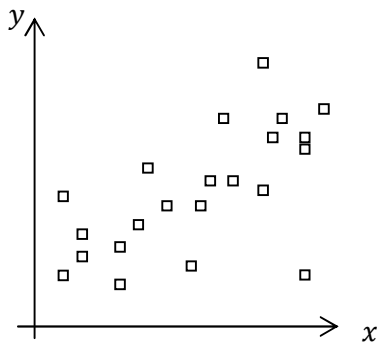


図 2

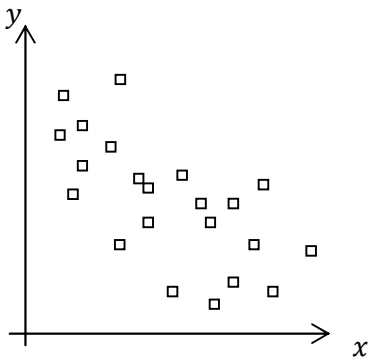


図 3

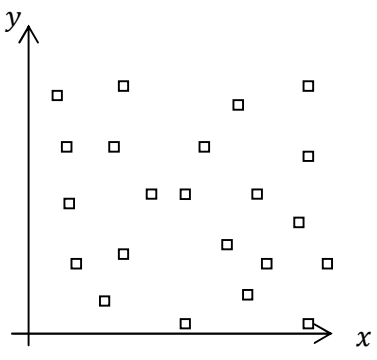


図 4

図 1,2 のようにxとyとの関係が傾き正の直線関係に近いとき、

『xとyとの間には正の相関関係がある』

という。直線関係への一致の度合いは、「強い・弱い」で表す。すなわち、図 1 の方が図 2 のより直線関係が大きく出ているので、図 1 の状態は

『xとyとの間には強い正の相関関係がある』

と表す。傾き負の直線関係に近い図 4 は

『xとyとの間には負の相関関係がある』

という。図 3 にはどこを見ても直線関係は見受けられないので、この場合は

『xとyとの間には相関関係はない。xとy は無相関である。』

という。相関という言葉に対する注意として

相関関係≠因果関係　　無相関≠無関係

という 2 点を忘れないように注意。

2.1.2　相関係数

相関の度合(直線への当てはまり具合)を数値で表現するために以下の二つを定義する。

$$\text{共分散}C_{xy} = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$$

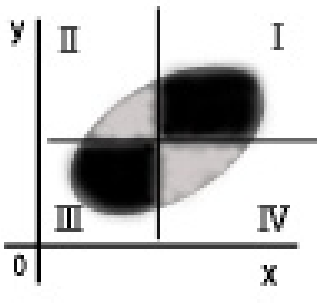
$$= \frac{1}{n}\left(\sum x_i y_i\right) - \bar{x}\bar{y}$$

$$\text{相関係数}r_{xy} := \frac{C_{xy}}{S_x S_y}$$

$$= \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}}$$

$$= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\{\sum(x_i-\bar{x})^2\}\{\sum(y_i-\bar{y})^2\}}}$$

・共分散C_{xy}



領域Ⅰ：$x > \bar{x}, y > \bar{y}$ 領域Ⅱ：$x < \bar{x}, y > \bar{y}$
領域Ⅲ：$x < \bar{x}, y < \bar{y}$ 領域Ⅳ：$x > \bar{x}, y < \bar{y}$

領域Ⅰ・Ⅲでは$(x_i-\bar{x})(y_i-\bar{y}) > 0$
領域Ⅱ・Ⅳでは$(x_i-\bar{x})(y_i-\bar{y}) < 0$

したがって

点が領域Ⅰ・Ⅲに多い(正の相関が強い)ほど共分散は大きい

点が領域Ⅱ・Ⅳに多い(負の相関が強い)ほど共分散は小さい

※共分散は観測値とは異なる次元を持つ。

・相関係数$r_{xy} = \dfrac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}} = \dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\{\sum(x_i-\bar{x})^2\}\{\sum(y_i-\bar{y})^2\}}}$ (無次元量)

これの特徴を考えてみよう

例題 4. 相関係数

相関係数の取り得る値の範囲を求めよ。

解答

シュワルツの不等式

$$\left\{\sum_{i=1}^{n}(A_i)^2\right\}\left\{\sum_{i=1}^{n}(B_i)^2\right\} \geqq \left\{\sum_{i=1}^{n}(A_iB_i)\right\}^2$$

において、

$$A_i = x_i - \bar{x} \qquad B_i = y_i - \bar{y}$$

とすれば

$$\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} \geqq \left\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right\}^2$$

$$\therefore \frac{\{\sum(x_i-\bar{x})(y_i-\bar{y})\}^2}{\{\sum(x_i-\bar{x})^2\}\{\sum(y_i-\bar{y})^2\}} \leqq 1$$

$$i.e. \qquad {r_{xy}}^2 \leqq 1$$

$$\therefore -1 \leqq r_{xy} \leqq 1$$

※P.49 のような方法もある。

■ $-1 \leqq r_{xy} \leqq 1$ の等号成立は$\dfrac{y_i-\bar{y}}{x_i-\bar{x}} = (i$ によらない一定値$)$なので

$$y_i - \bar{y} = k(x_i - \bar{x})$$

とおける。このとき、

$$y\text{の分散}{S_y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$= \frac{k^2}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= k^2{S_x}^2$$

$$\therefore k = \pm\frac{S_y}{S_x}$$

$r_{xy} = 1$ のとき$k = \frac{S_y}{S_x}$で、$r_{xy} = -1$ のとき$k = -\frac{S_y}{S_x}$となる。

すなわち、相関係数が±1 のときはデータが

$$y_i - \bar{y} = \pm\frac{S_y}{S_x}(x_i - \bar{x})$$

をみたしてなければいけない。

相関係数が 1 のことを、正の完全相関といい、相関係数が－1 のことを、負の完全相関という。

≪終わり≫

注意

相関係数は非常に便利な数値であるが、数値だけを見て判断を下すと誤解を生みやすい。
散布図も描いて判断するべきである。ただし、散布図を描く際には必ずデータの層別をすること。

例題 5. 相関関係と因果関係

2 変数x, yのデータにおいて,xの観測値に対してyの観測値が一義的に定まるならば,xとyの相関係数はr＝± 1であるといえるか.もしいえないとするならば,どういう場合かを実例をあげて示せ.(廣松)

解答

言えない．相関係数が±1 ということは，xとyの間に 1 次の関係があるということである．xに対しyが 1 義的に決まる例として，$y = \sqrt{1 - x^2}$を考える．点(x, y)はxy平面上で半円を描き，一次の関係はない．したがって相関係数は±1 ではない．　≪終わり≫

その他の相関係数に関しては P.52~58 を参照せよ。

§2.2　回帰分析

以下ではxがyを決める様子を調べる。y をx の一次式で近似することを考える。すなわち

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

y_i: 被説明変数　x_i: 説明変数

α, β: 回帰係数　ε_i: 誤差項

とする。

2.2.1　最小二乗法

近似が最適なものになるためには、誤差項の二乗和

$$\sum(\varepsilon_i)^2 = \sum(y_i - \alpha - \beta x_i)^2$$

が最小であることが望ましい。これをL(α, β)とすると、

$$\frac{\partial \mathrm{L}}{\partial \alpha} = \sum(-2)(y_i - \alpha - \beta x_i)$$

$$\frac{\partial \mathrm{L}}{\partial \beta} = \sum(-2x_i)(y_i - \alpha - \beta x_i)$$

であり、$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \beta} = 0$がL(α, β)が最小となるための必要条件である。(この 2 式を正規方程式という)

$$i.e. \begin{cases} \sum y_i - n\alpha - \beta \sum x_i = 0 \\ \sum x_i y_i - \alpha \sum x_i - \beta \sum {x_i}^2 = 0 \end{cases}$$

これを解いて

$$\begin{cases} \beta = \dfrac{\sum x_i\, y_i - n\overline{x} \cdot \overline{y}}{(\sum {x_i}^2) - n\overline{x}^2} = \dfrac{C_{xy}}{{S_x}^2} = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} (= \beta_0 \text{とする}) \\ \alpha = \overline{y} - \beta\overline{x} (= \alpha_0 \text{とする}) \end{cases}$$

よって求めたい回帰直線は

$$y = \alpha_0 + \beta_0 x$$

となる。β_0を回帰係数という。

2.2.2　相関係数の意味

データを直線で近似しようという目的からもわかるが、回帰と相関にはつながりがある。

$$r_{xy} := \frac{C_{xy}}{S_x S_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum(x_i - \bar{x})^2\}\{\sum(y_i - \bar{y})^2\}}}$$

$$\beta_0 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{C_{xy}}{{S_x}^2}$$

より

$$\beta_0 = r_{xy}\frac{S_y}{S_x}$$

である。このことから相関係数は、x, y を直線で近似したときの当てはまりの良さを表しているということが漸く確認された。

2.2.3　決定係数

$$\hat{y_i} = \alpha_0 + \beta_0 x_i$$

とおくと

$$\varepsilon_i = y_i - \hat{y_i}$$

とかけて、これを回帰残差と呼ぶ。するとε_i は$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \beta} = 0$より

$$\sum \varepsilon_i = 0 \qquad \cdots (2-1)$$

$$\sum \varepsilon_i x_i = 0 \qquad \cdots (2-2)$$

をみたす。式(B)は$\{\varepsilon_i\}$と$\{x_i\}$がベクトルとして直交しているという性質を表す重要な式である。
ここで

$$L(\alpha_0, \beta_0) = \sum (\varepsilon_i)^2 = \sum (y_i - \hat{y_i})^2$$

が持つ意味を考える。似た形の

$$\sum (y_i - \bar{y})^2$$

を考えてみよう。これは、原データy_i の平均$\bar{y}$ からの散らばりの総和(変動という)を表す。
そして、x_i から説明した$\hat{y_i}$ の平均も$\bar{y}$ であるので、

$$\sum (\hat{y_i} - \bar{y})^2$$

は$\hat{y_i}$ の変動を表す。
この 2 つの変動は値として近い方が良いので、その差を考えてみる。

$$\sum (y_i - \bar{y})^2 - \sum (\hat{y_i} - \bar{y})^2$$

$$= \sum ({y_i}^2 - 2 y_i \bar{y} - {\hat{y_i}}^2 + 2 \hat{y_i} \bar{y})$$

$$= \sum (y_i - \hat{y_i})(y_i - 2\bar{y} + \hat{y_i})$$

$$= \sum (y_i - \hat{y_i})\{(y_i - \hat{y_i}) + 2(\hat{y_i} - \bar{y})\}$$

$$= \sum (y_i - \hat{y_i})^2 + 2 \sum (y_i - \hat{y_i})(\hat{y_i} - \bar{y})$$

$$= \sum {\varepsilon_i}^2 + 2 \sum \varepsilon_i (\hat{y_i} - \bar{y})$$

ここで

$$\sum \varepsilon_i(\hat{y_i} - \bar{y}) = \sum \varepsilon_i(\alpha_0 + \beta_0 x_i - \bar{y})$$

$$= \beta_0 \sum \varepsilon_i x_i + (\alpha_0 - \bar{y}) \sum \varepsilon_i$$

$$= 0 \quad (\because \text{式}(A)(B))$$

したがって

$$\sum (y_i - \bar{y})^2 - \sum (\hat{y_i} - \bar{y})^2$$

$$= \sum {\varepsilon_i}^2 + 2 \sum \varepsilon_i(\hat{y_i} - \bar{y})$$

$$= \sum {\varepsilon_i}^2$$

すなわち

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y_i} - \bar{y})^2 + \sum {\varepsilon_i}^2$$

となる。この式は、『元々の変動』は『回帰直線で説明し得る変動』と『残差の二乗和(=説明できない変動)』に分解される、ということを示している。

このことを踏まえて、決定係数を定義する。

$$\text{決定係数} \eta^2 = \frac{\sum (\hat{y_i} - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum {\varepsilon_i}^2}{\sum (y_i - \bar{y})^2}$$

$$\left(= \frac{\text{回帰直線で説明し得る変動}}{\text{元々の変動}} = 1 - \frac{\text{残差の二乗和}(\text{説明できない変動})}{\text{元々の変動}}\right)$$

η^2が 0 に近いほど、$\hat{y_i}$ はy_i から外れていて、η^2が 1 に近いほど、$\hat{y_i}$ はy_i に一致している。

例題 6. 決定係数と相関係数

決定係数と相関係数の関係を調べよ。 (倉田)

解答

$$\eta^2 = \frac{\sum (\hat{y_i} - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

$$= \frac{\sum \{(\alpha_0 + \beta_0 x_i) - (\alpha_0 + \beta_0 \bar{x})\}^2}{\sum (y_i - \bar{y})^2}$$

$$= {\beta_0}^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

$$= {\beta_0}^2 \frac{{S_x}^2}{{S_y}^2}$$

$$= {r_{xy}}^2 \qquad \left(\because \beta_0 = r_{xy}\frac{S_y}{S_x}\right)$$

よって決定係数は相関係数の二乗と一致する。

≪終わり≫

※多次元データにおける決定係数と相関係数----------

$(p+1)$次元データ(サイズn)

$$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i) \quad (i = 1,2,\cdots,n)$$

において、$x_{i1}, x_{i2}, \cdots, x_{ip}$ の一次式で y_i を説明する様子を分析するのを重回帰分析という。
2 次元データのときと同じように

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

とおいて、最小二乗法で$\alpha, \beta_1, \beta_2, \cdots, \beta_p$を求める。その$\alpha, \beta_1, \beta_2, \cdots, \beta_p$ に対して決定係数を

$$\eta^2 = 1 - \frac{\sum {\varepsilon_i}^2}{\sum (y_i - \bar{y})^2}$$

と定める。多次元データにおいては、(重)相関係数 R は決定係数の正の平方根として定義されている。すなわち

$$\mathrm{R} = \sqrt{\eta^2}$$

である。
例題 6 で見た、『相関係数の二乗が決定係数に一致する』という不思議な性質は、多次元においては『定義』だったのである。

≪終わり≫

演習問題 1　回帰分析(倉田 03)

16 組の父子の身長を計測したところ、$(x_1, y_1), (x_2, y_2), \cdots, (x_{16}, y_{16})$ なるデータが得られたものとする。xは父の身長、yは子の身長とする。単位は cm とする。このデータから以下の数値が得られた。

$$\text{父の平均} = \bar{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = 166.3 \qquad \text{父の分散} = \frac{1}{16}\sum_{i=1}^{16}(x_i - \bar{x})^2 = 31.33$$

$$\text{子の平均} = \bar{y} = \frac{1}{16}\sum_{i=1}^{16} y_i = 173.1 \qquad \text{子の分散} = \frac{1}{16}\sum_{i=1}^{16}(y_i - \bar{y})^2 = 27.71$$

$$\text{父と子の共分散} = \frac{1}{16}\sum_{i=1}^{16}(x_i - \bar{x})(y_i - \bar{y}) = 24.13$$

(1) 子の身長yの父の身長xへの回帰直線$y = a + bx$を計算せよ。
(2) 回帰係数bの解釈を述べよ。
(3) 決定係数を計算せよ。

演習問題 2　回帰分析(倉田 04)

50 組の父子の身長を計測したところ、$(x_1, y_1), (x_2, y_2), \cdots, (x_{50}, y_{50})$ なるデータが得られたものとする。xは父の身長、yは子の身長とする。単位は cm とする。このデータから以下の数値が得られた。

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = 167.2 \qquad {S_x}^2 = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})^2 = 38.7$$

$$\bar{y} = \frac{1}{50}\sum_{i=1}^{50} y_i = 172.4 \qquad {S_y}^2 = \frac{1}{50}\sum_{i=1}^{50}(y_i - \bar{y})^2 = 31.8$$

$$C_{xy} = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})(y_i - \bar{y}) = 12.0$$

(1) 子の身長yの父の身長xへの回帰直線$y = a + bx$を計算せよ。
(2) 回帰係数bの解釈を述べよ。
(3) 決定係数を計算せよ。

~Tea Break Ⅰ~　　確率

教科書:第四章

1　基本事項覚え書き

[P.69]

・標本点…サイコロでいう 1,2,3,4,5,6　　ωで表す

・標本空間／全事象…サイコロでいう{1,2,3,4,5,6}　　Ωで表す.

・事象…標本空間の部分集合

・空集合φ

[P.70]

・根元事象…1 つの標本点だけからなる,これ以上分解できない事象

・複合事象…2 つ以上の根元事象に分解できる事象

[P.73]

・ベン図

[P.74]

・排反事象

・和事象…A∪B

・積事象…A∩B

※ある標本空間をΨとする．Ψが∪と∩に閉じているとき(任意のA, B $\in \Psi$に対して，A∪B $\in \Psi$，A∩B$\in \Psi$であるとき)Ψは加法族であるという

・$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

・$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

[P.75]

・補事象…高校では$\overline{A}$　大学からはA^c

・ド・モルガンの法則

[P.80]

・加法定理…$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

[P.81~82]

・条件付き確率…B が起こった場合に A が起こる確率

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

これより　乗法定理：$P(A \cap B) = P(B) \cdot P(A|B)$　　を得る.

[P.82]

・事象 A,B が独立($\Leftrightarrow P(A|B) = P(A)$) $\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

2 確率の定義[P.75~78] (読み流してよい)

2.1 ラプラスの定義

試行の根元事象が N個あって、それらは同程度に確からしいとする[4]．この中に,事象Aにとって都合のよいような根元事象(＝Aが起こるような根元事象)がR個あれば,事象Aの確率は

$$P(A) = \frac{R}{N}$$

で定義される．
この定義の最大の利点は,確率が場合の数の数え上げに帰着することである．

『同程度に確からしい』と考えられない場合はラプラスの定義は適用できない。そこで確率の頻度説が登場する．

2.2 頻度による確率の定義

一般に事象Aを生み得る実験をn回おこない,そのうちAがn_A回出るとする．

$$n \to \infty \text{のとき} \quad \frac{n_A}{n} \to \alpha$$

となるならば,$P(A) = \alpha$と定義する．

しかしながら,この定義も完全なものではない.「実験をn回おこなう」のは現実での話であるが，nを無限大に飛ばすのは理論上の話であって現実の話ではない．頻度による定義はこのようなギャップを含んでいるため，理論的に完全でない．このような困難を克服したのが確率の功利主義的定義である．

2.3 功利主義的定義

(1)任意の事象 A に対して $0 \leqq P(A) \leqq 1$
(2)$P(\varphi) = 0$(又は $P(\Omega) = 1$)
(3)任意のA_iとA_jが排反のとき($\Leftrightarrow A_i \cap A_j = \varphi$)

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

以上(1)(2)(3)をみたす P のことを確率と定義する．

[4] 『同程度に確からしく』の仮定の正しさを疑うかもしれないが、『同程度に確からしくない』とする十分な理由がない限りは『同程度に確からしい』と考える．(理由不十分の原則)

3 ベイズの定理[P.84]

ある試行において,得られた結果をA,原因を$H_1, H_2, \cdots, H_n$とする.『Aが起こったとき,原因がH_iである』確率$P(H_i|A)$に関して、次のような公式がある(証明は P.84 を見よ)

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}$$

ここで $P(H_i)$はH_iの事前確率、$P(H_i|A)$はH_iの事後確率という．Aが起こった前に考えたものか，後に考えたものかに従っている.

演習問題 3　条件付き確率(倉田 06)

(1)ある都市において全体の 0.05%の住民があるウイルスに感染していることが知られている。そのウイルスについて新しい診断法が開発されたとする。その診断法は感染者の 99.8%を正確に検出するが、非感染者の 0.3%を誤って(感染者であると)判断する。この年の住民 1 名を検査したところ、ウイルスが検出された。この人が実際に感染者である確率はいくらか。
(2)上問で導かれた結果についてコメントせよ。

演習問題 4　事象の独立(倉田 05)

2 つの事象 A、B が独立であることの定義を述べよ。(1,2 行)

演習問題 5　条件付き確率(廣松)

3つの機械A,B,Cで全製品の50%,30%,20%を製造している工場がある.これらの機械の不良品率は3%,4%,5%であるとする.
(1)1つの製品を無作為に選んだとき,それが不良品である確率を求めよ.
(2)その不良品が機械 A から製造されたものである確率を求めよ。

演習問題 6　条件付き確率(廣松)

2 つのツボがある.第一のツボには白玉が 3 個,赤玉 1 個入っており,第二のツボには白玉が 1 個,赤玉が 3 個入っている.今いずれかのツボから玉を 1 個取り出したところ,白玉であった.玉を取り出したのはどちらのツボか?

演習問題 7　条件付き確率(廣松)

あるメーカーの電球の不良品検査において，20 個を検査して，そのうち不良品が 1 個までならば，合格であるとする．このとき，真の不良品率が 5%であるのに不合格となる確率はいくらか．また真の不良品率が 10%のとき，合格となる確率はいくらか.

演習問題 8　条件付き確率(廣松)

10 本のうち 2 本が当たりであるようなクジがある. 甲乙丙 3 人がこの順に 1 本ずつクジを引くとき，甲乙丙の間に損得はあるか．(結論だけでなく，根拠も示すこと)ただし，引いたクジは戻さないとする.

演習問題 9　条件付き確率(廣松)

ツボ A は 5 個の赤玉と 3 個の白玉を含んでおり, ツボ B は 2 個の赤玉と 6 個の白玉を含んでいるとする.
(1)ツボから 1 個ずつ玉を取り出したとき,2 個の玉が同色である確率はいくらか.
(2)各ツボから 2 個ずつ非復元で取り出したとき,4 個の玉が同色である確率はいくらか.

第三講 確率分布総論[5]

教科書:第五章

§3.1　確率変数・確率分布

確率変数とは、出方が確率によって支配されている量である。すなわち確率変数には、確率が対応している。
確率変数が大文字(X, Y, ⋯)のときは変数としてみており、小文字$(x, y, \cdots)$のときは実現値として扱うのが普通である。

3.3.1　離散型確率変数・確率分布

可算集合$\{x_1, x_2, \cdots\}$の中の値をとる(とびとびの値をとる)確率変数は離散型といわれる。
ひとつの値に対し、それぞれ確率が対応している。

$$x_k \longrightarrow p_k \qquad (k = 1,2,\cdots,)$$

確率変数x_kが一つに定まれば、確率p_kが決まる。すなわちpはx の関数であるので

$$(\mathrm{P}(\mathrm{X} = x_k) =)p_k = f(x_k)$$

とかける。この$f(x)$を確率質量関数[6]という。$f(x)$は確率の規格化に対応して

$$f(x_k) \geqq 0 \quad (k = 1,2,\cdots), \sum_k^{\infty} f(x_k) = 1$$

をみたす。

3.1.2　連続型確率変数・確率分布

ある区間(開でも閉でもかまわない)の中の値を取る(数直線にベタ付けされた)確率変数を連続型確率変数という。

連続型にも、確率分布を定義したい

離散量の和は$\sum$で与えられるのに対し、連続量の和は$\int \quad dx$ で与えられる。
このことを用いて、確率の規格化条件に着目する。

---

[5] まだ具体的な確率分布の説明をしていないので、この章では一般的に通じる性質しか説明できない。それゆえ、やや分かりづらい部分もあるかと思われる。第四講の各論を読み終えたら、ぜひ読みなおしていただきたい。

[6] 単に確率関数ともいうが、連続型の理解のためにはこの方が良い。確率を『重み』『質量』ととらえることは少なくない。

$$(\text{離散型}) \qquad \sum_{k}^{\infty} f(x_k) = 1$$

から類推すると、

$$(\text{連続型}) \qquad \int_{-\infty}^{\infty} ?(x)\,dx = 1 \qquad \cdots(3-1)$$

となるような$?(x)$が連続型での確率分布となりそうである。

式$(3-1)$の右辺の”1”は確率であり、いまはこれを『質量』と捉えていた。よって、dxを長さと捉えれば、$?(x)$は『(線)密度』ということになる。

実際、この$?(x)$を確率密度関数といい、(まぎらわしいが)これにも記号$f(x)$を用いる[7]。

この密度関数を用いて、連続型確率変数Xが$a \leqq X \leqq b$をみたす確率は

$$P(a \leqq X \leqq b) = \int_{a}^{b} f(x)dx$$

と定義する。(面積が確率を表す)

この定義から連続型では一点$X = a$となる確率は

$$P(X = a) = P(a \leqq X \leqq a) = \int_{a}^{a} f(x)dx = 0$$

となることがわかる[8]。

$f(x)$は確率の規格化条件として

$$f(x) \geqq 0 \quad , \quad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

をみたす。2 式目の必要条件として

$$\lim_{x\to\pm\infty} f(x) = \lim_{x\to-\infty} f(x) = 0$$

を得る。

・累積確率分布…確率変数Xがある値以下となる確率を与える。

$$(\text{離散型}) \qquad \Phi(x_k) = P(X \leqq x_k) = \sum_{i=1}^{k} f(x_i)$$

$$(\text{連続型}) \qquad \Phi(x) = P(X \leqq x) = \int_{-\infty}^{x} f(t)dt$$

---

[7] 『確率分布』『$f(x)$』という同じ表現を用いていても、その中身が離散型か連続型かをはっきりと区別して考える必要がある。

[8] 感覚的には受け入れにくいであろうが、無限を考えると、このような現実とのギャップが生じるのだ、程度でよい。
(密度不均一な金属棒に対し、『端からちょうど3cm のところの質量はどれくらいですか?』というようなものである)

§3.2　確率分布総論

3.2.1　確率分布の基本的指標[9]

①平均μ = E[X]

$$(\text{離散型}) \qquad E[X] := \sum_x xf(x)$$

$$(\text{連続型}) \qquad E[X] := \int_{-\infty}^{\infty} xf(x)dx$$

Xの関数$\Psi(X)$に対しては

$$(\text{離散型}) \qquad E[(\Psi(X)] := \sum_x \Psi(x)f(x)$$

$$(\text{連続型}) \qquad E[\Psi(X)] := \int_{-\infty}^{\infty} \Psi(x)f(x)dx$$

とする。

性質(a, bは定数し、X, Yは確率変数とする)

$$E[b] = b,\ E[aX + b] = aE[X] + b,\ E[X + Y] = E[X] + E[Y]$$

②分散$\sigma^2 = V[X] := E[(X-\mu)^2] \equiv E[X^2] - \mu^2$

性質(a, bは定数し、Xは確率変数とする)

$$V[b] = 0 \quad , \quad V[aX + b] = a^2\, V[X]$$

③歪度$\alpha_3 := \dfrac{E[(X-\mu)^3]}{\sigma^3} \equiv \dfrac{1}{\sigma^3}(E[X^3] - 3\mu E[X^2] + 2\mu^3)$

確率分布関数の歪み具合を表す。

$\alpha_3 > 0 \Leftrightarrow$ 右の裾が長い

$\alpha_3 < 0 \Leftrightarrow$ 左の裾が長い

④尖度$\alpha_4 := \dfrac{E[(X-\mu)^4]}{\sigma^4} \equiv \dfrac{1}{\sigma^4}(E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4)$

確率分布関数の中心μまわりの尖り具合を表す。

ただし、普通は正規分布の尖度 3を基準するので,$\alpha_4 - 3$を考える。

$\alpha_4 - 3 > 0 \Leftrightarrow$ 正規分布より尖っている

$\alpha_4 - 3 < 0 \Leftrightarrow$ 正規分布より丸く鈍い

[9] ③④そのものは流し読みするだけで、覚えようとしなくてよい。

3.2.2　モーメント・モーメント母関数

・モーメント

3.2.1 で見たように、確率分布関数の形状を表現する指標は

$$E[X^r]$$

を用いて表される。この$E[X^r]$を原点まわりのr次モーメントといい、μ_rで表す。また$E[(X-\mu)^r]$をμまわりのr次モーメントといい、μ_r'で表す。(μとはμ_1のこと)

このように定義したモーメントを全ての次数rについて知ることができれば、その確率分布の形状(⇒性質)を一意に定めることができる。

・モーメント母関数

$$M_X(t) := E[e^{tX}]$$

と定義すると、

$$M_X{}^{(r)}(0) = \mu_r$$

となる。この$M_X(t)$をモーメント母関数という。

モーメント母関数のメリットは、比較的簡単な計算で各次数のモーメントが計算できる点である。

証明

マクローリン展開より

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

$$\therefore e^{tx} = \sum_{i=0}^{\infty} \frac{(tx)^i}{i!}$$

(離散型)

$$\begin{aligned} M_X(t) &:= E[e^{tX}] \\ &= \sum_x e^{tx} f(x) \\ &= \sum_x \left[\left\{ \sum_{i=0}^{\infty} \frac{(tx)^i}{i!} \right\} f(x) \right] \\ &= \sum_{i=0}^{\infty} \left[\frac{t^i}{i!} \sum_x \{x^i f(x)\} \right] \\ &= \sum_{i=0}^{\infty} \left(\frac{t^i}{i!} \mu_i \right) \end{aligned}$$

(連続型)

$$\begin{aligned} M_X(t) &:= E[e^{tX}] \\ &= \int_{-\infty}^{\infty} e^{tx} f(x) dx \\ &= \int_{-\infty}^{\infty} \sum_{i=0}^{\infty} \frac{(tx)^i}{i!} f(x) dx \\ &= \sum_{i=0}^{\infty} \left[\frac{t^i}{i!} \int_{-\infty}^{\infty} x^i f(x) dx \right] \\ &= \sum_{i=0}^{\infty} \left(\frac{t^i}{i!} \mu_i \right) \end{aligned}$$

$$\therefore M_X{}^{(r)}(t) = \left\{ \sum_{i=0}^{r-1} \left(\frac{t^i}{i!} \mu_i \right) + \frac{t^r}{r!} \mu_r + \sum_{i=r+1}^{\infty} \left(\frac{t^i}{i!} \mu_i \right) \right\}^{(r)}$$

$$= \mu_r + \sum_{i=r+1}^{\infty} \left\{ \frac{t^{i-r}}{(i-r)!} \mu_i \right\}$$

$$\therefore M_X{}^{(r)}(0) = \mu_r$$ ■

§3.3　チェビシェフの不等式

いかなる確率変数X(平均$\mu = E[X]$、分散$\sigma^2 = V[X]$)に対しても、以下の不等式が常に成立する。

$$P(|X-\mu| \geqq k\sigma) \leqq \frac{1}{k^2} \qquad (k \in \boldsymbol{R})$$

証明

(離散型)

$$\begin{aligned}
&\sigma^2 \\
&= E[(X-\mu)^2] \\
&= \sum_x \{(x-\mu)^2 f(x)\} \\
&= \sum_{|x-\mu| \geqq k\sigma} \{(x-\mu)^2 f(x)\} + \sum_{|x-\mu| \leqq k\sigma} \{(x-\mu)^2 f(x)\} \\
&\geqq \sum_{|x-\mu| \geqq k\sigma} \{(x-\mu)^2 f(x)\} \\
&\geqq \sum_{|x-\mu| \geqq k\sigma} \{(k\sigma)^2 f(x)\} \\
&= (k^2\sigma^2) \sum_{|x-\mu| \geqq k\sigma} f(x) \\
&= (k^2\sigma^2) P(|X-\mu| \geqq k\sigma)
\end{aligned}$$

(連続型)

$$\begin{aligned}
&\sigma^2 \\
&= E[(X-\mu)^2] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&= \int_{|x-\mu| \geqq k\sigma} (x-\mu)^2 f(x)dx + \int_{|x-\mu| \leqq k\sigma} (x-\mu)^2 f(x)dx \\
&\geqq \int_{|x-\mu| \geqq k\sigma} (x-\mu)^2 f(x)dx \\
&\geqq \int_{|x-\mu| \geqq k\sigma} (k\sigma)^2 f(x)dx \\
&= (k^2\sigma^2) \int_{|x-\mu| \geqq k\sigma} f(x)dx \\
&= (k^2\sigma^2) P(|X-\mu| \geqq k\sigma)
\end{aligned}$$

$$\therefore \sigma^2 \geqq (k^2\sigma^2) P(|X-\mu| \geqq k\sigma)$$

両辺を$k^2\sigma^2$で割って

$$P(|X-\mu| \geqq k\sigma) \leqq \frac{1}{k^2}$$

■

この式は、いかなる確率変数に対しても成立する点で、確率分布が不明のときに有効である。

演習問題 10　確率密度関数(倉田 05)

連続型確率変数Xの確率密度関数$f(x)$が次式のように与えられている。

$$f(x) = \begin{cases} 6x(1-x) & (0 < x < 1) \\ 0 & (\text{それ以外}) \end{cases}$$

このとき、$E[X]$、$E[X^2]$、$V[X]$を求めよ。

第四講 確率分布各論

教科書:第六章

§4.1　離散型確率分布

4.1.1　二項分布

二種類の可能な結果S(成功),F(失敗)を生じ得る試行(S, Fを生じる確率はそれぞれ$p, 1-p$とする)を独立にn回繰り返す。この試行(n回全体)をベルヌーイ試行という。Sが生じた回数がx回となる確率は

$$f(x) = {}_nC_x \cdot p^x \cdot (1-p)^{n-x}$$

で表される。このような確率分布を二項分布$Bi(n,p)$と呼び、特に$Bi(1,p)$はベルヌーイ分布と呼ばれる。

・確率変数 X が二項分布$Bi(n,p)$に従うとき、

$$E[X] = np\,, V[X] = np(1-p)$$

4.1.2　ポアソン分布

ベルヌーイ試行において、試行回数nが極めて大きく、成功確率pが極めて小さいときに、成功回数(Sが生じる回数)がx回となる確率は、$np = \lambda$として

$$f(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$

で表される。このような確率分布をポアソン分布$Po(\lambda)$という。

・確率変数Xがポアソン分布$Po(\lambda)$に従うとき、

$$E[X] = \lambda\,, V[X] = \lambda$$

■ポアソンの小数の法則

$np = \lambda\,, n \to \infty (\Rightarrow p \to 0)$のとき

$${}_nC_x \cdot p^x \cdot (1-p)^{n-x} \to e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$

証明

$${}_nC_x = \frac{n!}{(n-x)!\,x!} = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-x+1) \cdot \frac{1}{x!}$$

$$p^x = \left(\frac{\lambda}{n}\right)^x = \frac{\lambda^x}{n^x}$$

$$(1-p)^{n-x} = \left(1-\frac{\lambda}{n}\right)^{n-x} = \left(1-\frac{\lambda}{n}\right)^{-x} \cdot \left\{\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right\}^{-\lambda}$$

$\therefore {}_nC_x \cdot p^x \cdot (1-p)^{n-x}$

$$= \left\{ n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-x+1) \cdot \frac{1}{x!} \right\} \cdot \left(\frac{\lambda^x}{n^x} \right) \cdot \left[\left(1 - \frac{\lambda}{n} \right)^{-x} \cdot \left\{ \left(1 - \frac{\lambda}{n} \right)^{-\frac{n}{\lambda}} \right\}^{-\lambda} \right]$$

$$= \left\{ 1 \cdot \left(1 - \frac{1}{n} \right) \cdot \left(1 - \frac{2}{n} \right) \cdot \cdots \cdot \left(1 - \frac{x-1}{n} \right) \right\} \cdot \left(\frac{\lambda^x}{x!} \right) \cdot \left(1 - \frac{\lambda}{n} \right)^{-x} \cdot \left\{ \left(1 - \frac{\lambda}{n} \right)^{-\frac{n}{\lambda}} \right\}^{-\lambda}$$

$$\to \{ 1 \cdot 1 \cdot 1 \cdot \cdots \cdot 1 \} \cdot \left(\frac{\lambda^x}{x!} \right) \cdot 1 \cdot e^{-\lambda}$$

$$= e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$ ■

実際に計算するときは、λに標本データの平均を採用する。

例題 7. ポアソン分布

ある電話交換台に 1 時間あたり平均して 120 回の通話(コール)があるとする。さらに、この交換台の処理能力は 1 分間に 3 コール以内であるとする。いまある 1 分間をとったとき、この交換台では処理しきれないだけのコールがある確率は、どの程度か。計算は e のままでもよい。　　(廣松)

解答

平均 1 時間に 120 回のコールがあることより、平均 1 分間に 2 回のコールがある。したがって，1 分間にあるコールの回数Xは$\lambda = 2$のポアソン分布に従う。つまり、確率分布関数は

$$f(x) = \frac{e^{-2} \cdot 2^x}{x!}$$

となる．もとめる確率は余事象の考え方にしたがって,

$$1 - f(0) - f(1) - f(2) - f(3) = \cdots = 1 - \frac{19}{3} e^{-2} \quad \cdots (\text{答})$$

4.1.3　幾何分布

ベルヌーイ試行と同様に、成功S・失敗Fのみを生じる試行を考える。(それぞれの確率はp、$1-p$) このとき、x回目で初めてSとなる確率は、

$$f(x) = (1-p)^{x-1} \cdot p$$

で表される。これを幾何分布$Ge(p)$という。

一回の試行に一定時間かかるとすれば、xは待ち時間を表すことになり、幾何分布$Ge(p)$は成功までの待ち時間分布となる。

・確率変数Xが幾何分布$Ge(p)$に従うとき

$$E[X] = \frac{1}{p}, V[X] = \frac{1-p}{p^2}$$

・累積確率関数

$$F(x) = P(X \leqq x) = \sum_{t=1}^{x} f(t) = \sum_{t=1}^{x} p(1-p)^{t-1} = 1-(1-p)^x$$

・無記憶性(a, b正とする)

離散型確率変数Xが幾何分布$Ge(p)$に従う $\Leftrightarrow$ $P(X = a+b|X > b) = P(X = a)$

証明(⇒のみ)

$$\begin{aligned} P(X = a+b|X > b) &= \frac{P(X = a+b \text{かつ} X > b)}{P(X > b)} \\ &= \frac{P(X = a+b)}{P(X > b)} \\ &= \frac{P(X = a+b)}{1-P(X \leqq b)} \\ &= \frac{f(a+b)}{1-F(b)} \\ &= \frac{(1-p)^{a+b-1} \cdot p}{1-\{1-(1-p)^b\}} \\ &= (1-p)^{a-1} \cdot p \\ &= f(a) \\ &= P(X = a) \end{aligned}$$ ■

$X = b$の時点から見て、時間aだけ経過した$X = a+b$で事象Sが起こる確率$P(X = a+b|X > b)$は、
$X = 0$の時点から見て、時間aだけ経過した$X = a$で事象Sが起こる確率$P(X = a)$と等しい。
つまり、待ち時間がaである確率はどの時点から考えたかということに依存しない、事象 S が生じる確率は時間によらず一定であることを主張している。
これは、『事象Sがランダムに生じる』という条件に対する、定義式である。

4.1.4　一様分布

$$f(x) = \frac{1}{N} \quad (x = 1,2,\cdots,N)$$

で表される確率分布を一様分布という。

・確率変数Xが一様分布に従うとき

$$E[X] = \frac{N-1}{2}, V[X] = \frac{(N-1)^2}{12}$$

§4.2　連続型確率分布

4.2.1　指数分布

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geqq 0) \\ 0 & (x < 0) \end{cases}$$

と表される確率分布を指数分布$Ex(\lambda)$という。指数分布は連続的な待ち時間分布である[10]。

・確率変数Xが指数分布$Ex(\lambda)$に従うとき

$$E[X] = \frac{1}{\lambda}, V[X] = \frac{1}{\lambda^2}$$

・累積確率分布

$$F(x) = P(X \leqq x) = \int_{-\infty}^{x} f(x)\,dx = \begin{cases} 1 - e^{-\lambda x} & (x \geqq 0) \\ 0 & (x < 0) \end{cases}$$

・無記憶性(a, b正とする)

連続型確率変数Xが指数分布$Ex(\lambda)$に従う $\Leftrightarrow$ $P(X > a + b | X > b) = P(X > a)$

証明(⇒のみ)

$$\begin{aligned} P(X > a + b | X > b) &= \frac{P(X > a + b \text{かつ} X > b)}{P(X > b)} \\ &= \frac{P(X > a + b)}{P(X > b)} \\ &= \frac{1 - P(X \leqq a + b)}{1 - P(X \leqq b)} \\ &= \frac{1 - F(a + b)}{1 - F(b)} \\ &= \frac{1 - \{1 - e^{-\lambda(a+b)}\}}{1 - \{1 - e^{-\lambda b}\}} \\ &= e^{-\lambda a} \\ &= 1 - F(a) \\ &= 1 - P(X \leqq a) \\ &= P(X > a) \end{aligned}$$ ■

解釈は幾何分布の無記憶性と同様であり、やはり『事象がランダムに生じる』という条件に対する、定義式である。

---

[10] 離散的な待ち時間分布を表す幾何分布の累積確率分布

$$F(x) = 1 - (1 - p)^x$$

において、$(1 - p) = e^{-\lambda}$とすれば、

$$F(x) = 1 - e^{-\lambda x}$$

となり、指数分布の累積確率分布と一致する。このことから指数分布も待ち時間分布の性質をもつことがわかる、形式的ではあるが。興味のある人は次ページの『信頼性工学』の項目を参照せよ。(あくまでも参考である。基礎統計の範囲を大幅に逸脱している)

例題 8. 指数分布

ある病院では、外来患者の待ち時間はほぼ平均 30 分の指数分布に従うとする。この病院を訪れた外来患者が 20 分以上待たされる確率を求めよ。 (廣松)

解答

待ち時間Xが平均 30 分の指数分布に従う．すなわち，$\frac{1}{\lambda} = 30 \Leftrightarrow \lambda = \frac{1}{30}$で

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{30} e^{-\frac{1}{30}x}$$

求める確率は

$$\int_{20}^{\infty} f(x)dx = \cdots = e^{-\frac{2}{3}} \qquad \cdots 答$$

※信頼性工学----------

システムの故障確率などを考察する工学分野：信頼性工学と呼ばれるものを、少々紹介する。

◆t：システム起動からの経過時刻(連続量)

◆残存確率$R(t)$：システム全体のうち、時刻tでまだ故障が発生していないシステムの割合[11]

◆故障確率$F(t)$：システム全体のうち、時刻tでは既に故障してしまっているシステムの割合

このとき、

$$R(t) + F(t) = 1$$

◆故障時間tの密度関数$f(t)$：時刻tから単位時間後に故障するシステムの、全体に対する割合

このとき、

$$F(t) = 1 - R(t) = \int_0^t f(t)dt \quad , \quad \frac{dF(t)}{dt} = -\frac{dR(t)}{dt} = f(t)$$

◆瞬間故障率$\lambda(t)$：時刻tまで残存していたシステムが、時刻tから単位時間に故障する確率

$$\lambda(t) \equiv \frac{f(t)}{R(t)}$$ [12]

このとき、

$$\lambda(t) \equiv \frac{f(t)}{R(t)} = -\frac{1}{R(t)}\frac{dR(t)}{dt} = -\left(\frac{d}{dt} lnR(t)\right)$$

となる。初期条件

$$R(0) = 1$$

を考慮して解くと、

[11] すなわち、時刻tまでに故障が 1 件も発生しない確率を表す。その意味で『信頼度関数』とも呼ばれる。

[12] 条件付き確率としてみると捉えやすい。

$$R(t) = e^{-\int_0^t \lambda(t)dt}$$

となる。

$$\int_0^t \lambda(t)dt = H(t)$$

を累積ハザード関数という。

例　瞬間故障率$\lambda(t)$がシステム作動からの経過時間tに比例する場合(比例定数a)

$$\lambda(t) = at$$

より、

$$R(t) = e^{-\int_0^t (at)dt} = e^{-\frac{a}{2}t^2}$$

$$\therefore f(t) = -\frac{dR(t)}{dt} = ate^{-\frac{a}{2}t^2}$$

ポアソン分布・指数分布の導出

同時に作動開始した無数のシステムがある場合、任意の時刻tまでに故障するシステムの数を考える。このとき、以下の 3 つの条件[13]を仮定する。

(1)瞬間故障率$\lambda(t)$は時間によらず一定(λ)
　　⇒微小時間Δtにシステムが故障する確率は$\lambda\Delta t$に等しい
(2)2 個以上のシステムが同時に故障することはない
(3)システムの故障は互いに独立である

時刻tまでの故障数がiである確率を$P_i(t)$とする[14]。

①時刻$t + \Delta t$で故障数が 0 である確率は、時刻tまで 1 件も故障が発生せず、引き続く時間Δtの間も故障が発生しない確率である。

$$\therefore P_0(t + \Delta t) = P_0(t) \cdot (1 - \lambda\Delta t)$$

②時刻$t + \Delta t$での故障数が$i \geqq 1$である確率は、以下の 2 つの排反事象の和である。

・時刻tまでの故障数がiであり、引き続く時間Δtの間は故障が発生しない
・時刻tまでの故障数が$i - 1$であり、引き続く時間Δtの間に 1 件の故障が発生する

$$\therefore P_i(t + \Delta t) = P_i(t) \cdot (1 - \lambda\Delta t) + P_{i-1}(t) \cdot \lambda\Delta t \qquad (i \geqq 1)$$

13　ポアソン分布・指数分布の時に仮定した条件と比較してみよ。

14　tは連続量、iは 0 以上の整数値を取る離散量である。

①②において、時間間隔Δtを無限小に取ると、以下の微分方程式 2 本を得る。

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \qquad \cdots(4-1)$$

$$\frac{dP_i(t)}{dt} = -\lambda\{P_i(t) - P_{i-1}(t)\} \qquad (i \geqq 1) \qquad \cdots(4-2)$$

$t = 0$ではシステムの故障はないとしたときの初期条件

$$P_0(0) = 1\,, P_i(0) = 0 \quad (i \geqq 1)$$

も考慮して$(4-1), (4-2)$を解くと[15]

$$P_i(t) = \frac{(\lambda t)^i}{i!} \cdot e^{-\lambda t} \qquad \cdots(4-3)$$

となる。

・式$(4-3)$において$t = t_0$と固定してみる。$\lambda t_0 = N_0$とおけば

$$P_i(t_0) = \frac{(N_0)^i}{i!} \cdot e^{-N_0} = Po(N_0)$$

すなわち、ある時刻t_0までに発生する故障件数iはポアソン分布$Po(N_0)$に従うことがわかる。

・式$(4-3)$において$i = 0$としてみる。$P_0(t)$は時刻tまでに発生するシステム故障件数が 0 である確率を示すが、これは定義により残存確率$R(t)$である。

$$\therefore R(t) = P_0(t) = \frac{(\lambda t)^0}{0!} \cdot e^{-\lambda t} = e^{-\lambda t}$$

これより、故障時間tの密度関数$f(t)$は

$$f(t) = -\frac{dR(t)}{dt} = \lambda e^{-\lambda t} = Ex(\lambda)$$

すなわち、ランダムなシステム故障の待ち時間分布は指数分布$Ex(\lambda)$であることがわかる。

例題 9. ポアソン分布と指数分布

ある窓口への客の到着は、平均 2 分に 1 人である。このとき,以下の確率を答えよ。

(1)5 分間に 1 人も客が到着しない確率

(2)5 分間に 2 人以上の客が到着する確率

(3)客の到着間隔が 3 分以上になる確率 (廣松･改題)

解答

(1)【ポアソン分布$Po(\lambda_1)$を利用】

『5 分間に』なので、確率変数Xを「5 分間の来客数」とすると、

『平均 2 分に 1 人』⇔『平均 5 分に 2.5 人』

[15] 微分方程式の解法の一つ、『定数変化法』を繰り返し用いる。

ゆえ、確率変数Xは$Po(\lambda_1 = 2.5)$にしたがい、確率質量関数は

$$f(x) = \frac{e^{-2.5} \cdot (2.5)^x}{x!}$$

となる。5 分間に 1 人も客が来ない確率は$x = 0$を代入して

$$f(0) = \frac{e^{-2.5} \cdot (2.5)^0}{0!} = e^{-2.5} \cdots (答)$$

【指数分布$Ex(\lambda_2)$を利用】

来客の待ち時間Yの平均は 2 分ゆえ、Yがしたがう分布を$Ex(\lambda_2)$とすると、

$$\frac{1}{\lambda_2} = 2 \Leftrightarrow \lambda_2 = \frac{1}{2}$$

よって、確率密度関数は

$$g(y) = \frac{1}{2} e^{-\frac{1}{2}y}$$

となる。いま、来客の到着時間間隔が 5 分以上になるので、求める確率は

$$\int_5^\infty g(y)dy = \left[-e^{-\frac{1}{2}y}\right]_5^\infty = e^{-2.5} \cdots (答)$$

(2)

『5 分間に 2 人以上』=『全事象』－『5 分間に 0 人』－『5 分間に 1 人』より、求める確率は

$$1 - f(0) - f(1) = 1 - 3.5e^{-2.5} \cdots (答)$$

(3)【ポアソン分布$Po(\lambda_3)$を利用】

3 分間の来客数が 0 人であることが必要十分条件。よって、3 分間の来客数をZとすると、その平均は 1.5 ゆえ、 Zは$Po(\lambda_3 = 1.5)$にしたがい、確率質量関数は

$$h(z) = \frac{e^{-1.5} \cdot (1.5)^z}{z!}$$

したがって、求める確率

$$h(0) = e^{-1.5} \qquad \cdots (答)$$

【指数分布$Ex(\lambda_2)$を利用】

求める確率は、

$$\int_3^\infty g(y)dy = \left[-e^{-\frac{1}{2}y}\right]_3^\infty = e^{-1.5} \cdots (答)$$

4.2.2　一様分布

$$f(x) = \begin{cases} \dfrac{1}{b-a} & (a \leqq x \leqq b) \\ 0 & (x < a, b < x) \end{cases}$$

で表される確率分布を一様分布という。

・確率変数Xが一様分布に従うとき

$$E[X] = \frac{a+b}{2}, V[X] = \frac{(b-a)^2}{12}$$

例題 10. 一様分布

毎時 0 分と 30 分に発車する汽車がある。そのことを知らずに、汽車に乗るために駅に着たときの待ち時間をT分とする。以下の問題に答えよ。

(1)Tのとる値の範囲を求めよ。

(2)Tがx分以下である確率$P(T \leqq x)$を求めよ。

(3)Tの期待値と分散を求めよ。　(廣松)

解答

(1)$0 \leqq T \leqq 30$

(2) Tは一様分布(連続型)にしたがうので、確率密度関数を

$$f(x) = p$$

とすると

$$\int_0^{30} f(x)dx = 1 \Leftrightarrow 30p = 1 \Leftrightarrow p = \frac{1}{30}$$

$$\therefore P(T \leqq x) = \int_0^x f(x)dx = \frac{x}{30} \quad \cdots (答)$$

(3)Tの期待値μ…

$$\mu = \int_0^{30} xf(x)dx = \int_0^{30} \frac{x}{30}dx = 15 \quad \cdots (答)$$

Tの分散σ^2…

$$\sigma^2 = \int_0^{30} (x-\mu)^2 f(x)dx = 75 \quad \cdots (答)$$

※駅が客を待つ時間は、指数分布にしたがう。この場合、一様な確率でやってくるのは電車ではなく乗客である。

4.2.3　正規分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

で表わされる確率分布を正規分布$N(\mu, \sigma^2)$という。特に、$N(0,1)$を標準正規分布という。正規分布は自然界や人間社会の数多くの現象に当てはまる点で重要である。

・確率変数Xが正規分布$N(\mu, \sigma^2)$に従うとき

①$E[X] = \mu\,, V[X] = \sigma^2$[16]

②$Y = aX + b$は正規分布$N(a\mu + b, a^2\sigma^2)$に従う[17]

③$Z = \frac{x-\mu}{\sigma}$は標準正規分布$N(0,1)$に従う

例題 11. 正規分布

ある試験の得点は正規分布$N(68,100)$にしたがっているとする。このとき、以下の問いに答えよ。

(1)60点以上70点以下の得点を取る人の確率を求めよ。

(2)得点上位 30%を優にするとして、およそ何点以上が優になるかを求めよ.　　(廣松)

解答

得点Xは$N(68,100)$に従う。よって

$$Z = \frac{X-68}{10}$$

は$N(0,1)$に従う。

(1)

$$60 \leqq X \leqq 70 \quad \Leftrightarrow \quad -0.8 \leqq Z \leqq 0.2$$

ゆえ

$$P(60 \leqq X \leqq 70) = P\left(-0.8 \leqq Z \leqq 0.2\right) = 1 - 0.42074 - 0.21186 = 0.3674 \quad \cdots (\text{答})$$

(2)

$$P(Z \geqq 0.52) = 0.30153 \qquad P(Z \geqq 0.53) = 0.29806$$

より、$P(Z \geqq Z_0) = 0.300$となるZ_0は$0.52 < Z_0 < 0.53$をみたす。

よって優をとるのに必要な点は

$$0.52 < Z_0 = \frac{X_0 - 68}{10} < 0.53$$

$$73.2 < X_0 < 73.3$$

より、74 点である。…(答)

---

[16]期待値μと分散σ^2は、密度関数の中に含まれていることに注意せよ。

[17] $E[Y] = a\mu + b\,, V[Y] = a^2\sigma^2$は簡単。$Y$が正規分布に従うことは、モーメント母関数を計算することで示される。

演習問題 11　確率分布・条件付き確率(倉田 03)

(1)表の出る確率が 0.6 であるようなコインを 5 回投げるような試行を考える。表の出る回数をXとするとき、$X = 3$となる確率$P(X = 3)$を求めよ。
(2)表の出る確率が 0.0002 であるようなコインを 10000 回投げるような試行を考える。表の出る回数をXとするとき、$X = 3$となる確率$P(X = 3)$を求めよ。
(3)表の出る確率が 0.6 であるようなコインを表が出るまで投げ続けるような試行を考える。X回目に始めて表が出るとするとき、$X = 3$となる確率$P(X = 3)$を求めよ。
(4)表が出る確率がpであるようなコインを 5 回投げる試行を考える。事象A, Bをそれぞれ、

$$A = \{1 \text{ 回目に表が出る}\}, \qquad B = \{\text{表の出る回数は } 3 \text{ 回}\}$$

とするとき、Bが与えられたときのAの条件付き確率$P(A|B)$を求めよ。

演習問題 12　確率分布(倉田 04)

(1)日本人の 30%は何らかの宗教を信仰している。6 人の日本人に宗教を信仰しているか否かを尋ね、信仰していると答える人数をXとするとき、$X = 2$である確率$P(X = 2)$を求めよ。
(2)日本人の 0.2%は自分を上流階級と考えている。1000 人の日本人に自分を上流階級と思うか否かを尋ね、上流階級と答える人数をXとするとき、$X = 3$である確率$P(X = 3)$を求めよ。
(3)日本人の 60%は北枕を嫌う。団体旅行者の添乗員が、宿泊先の都合上、客の誰かに北枕で寝てもらうことを順々に頼まなければならないとする。X人目で初めて北枕 OK に返事がもらえるとするとき、確率$P(X \leqq 3)$を求めよ。

演習問題 13　幾何分布(倉田 05)

(1)幾何分布の無記憶性を『災害発生時点』の例として、解釈せよ。
(2)幾何分布の無記憶性を表現する式を書き、証明せよ。

演習問題 14　ポアソン分布(倉田 06)

ある溶液は 1ml あたり平均 3 個のバクテリアを含む。この溶液 1ml 中のバクテリアの数はポアソン分布にしたがうと仮定して次の確率を求めよ。
(1)1ml の溶液を採取したとき、その中に 4 個以上のバクテリアが含まれる確率。
(2)1ml の溶液を 2 回採取したとき、どちらにもバクテリアが含まれない確率。
(3)1ml の溶液を 3 回採取したとき、3 回のうち 2 回に少なくとも 1 個のバクテリアが含まれる確率。

演習問題 15　正規分布(倉田(1)03&04,(3)05,(5)06)

(1)確率変数Xは正規分布N(50,100)にしたがうものとする。確率P(X ≧ 70)、P(40 ≦ X ≦ 60)、P(X ≦ 55)をそれぞれ求めよ。

(2)小学校 6 年生の男子の身長は、平均 145.2cm、標準偏差 7.1cm の正規分布で表わされることが知られている。156cm 以下の男子の割合を求めよ。

(3)出生男児の体重は、平均 3.2kg、標準偏差 0.4kg の正規分布にしたがうとする。1 人の出生男児を無作為に選ぶとき、「2.7kg 以上 3.7kg 以下」になる確率を求めよ。

~Tea Break Ⅱ~　多次元の確率分布[18]

教科書:第七章

1 基本事項覚え書き

同時確率分布

離散型確率変数X,Yに対し$X = x, Y = y$となる確率

$$P(X = x, Y = y) = f(x,y)$$

を 2 次元確率変数(X,Y)に対する同時確率分布という。規格化条件として以下の 2 式をみたす。

$$f(x,y) \geqq 0\ , \sum_x \sum_y f(x,y) = 1$$

・周辺確率分布…X,Y単独の確率分布はそれぞれ

$$g(x) = \sum_y f(x,y)\ , h(y) = \sum_x f(x,y)$$

で求められる。これを周辺確率分布という。

同時確率密度関数

連続型確率変数X,Yに対し$a \leqq X \leqq b, c \leqq Y \leqq d$となる確率は

$$P(a \leqq X \leqq b, c \leqq Y \leqq d) = \int_c^d \int_a^b f(x,y)dxdy$$

と表される。この$f(x,y)$を 2 次元確率変数(X,Y)に対する同時確率密度関数という。
規格化条件として以下の 2 式をみたす。

$$f(x,y) \geqq 0\ , \int_y \int_x f(x,y)dxdy = 1$$

・周辺確率密度関数…X,Y単独の確率密度関数はそれぞれ

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy\ , h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

で求められる。これを周辺確率密度関数という。

二次元の期待値

$$(\text{離散型}) \qquad E[(\Psi(X,Y)] := \sum_x \sum_y \Psi(x,y) f(x,y)$$

$$(\text{連続型}) \qquad E[\Psi(X,Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Psi(x,y) f(x,y) dxdy$$

共分散$(\mu_X = E[X], \mu_Y = E[Y]$とする$)$

$$Cov[X,Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

---

[18] 主に 2 次元の確率分布について述べていくが、数学的に追いついてないので(主に連続型)、特に細かい説明はしない。
イメージだけ掴んでもらえればよい。

相関係数

$$\sigma_{XY} = \frac{Cov[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}}$$

$$-1 \leqq \sigma_{XY} \leqq 1$$

条件付き確率分布[19]

$$g(x|y) = \frac{f(x,y)}{h(y)}, h(y|x) = \frac{f(x,y)}{g(x)}$$

条件付き期待値・分散に関しては、各自確認されたし。

独立性

$$\text{確率変数} X,Y \text{が独立} \overset{def}{\Longleftrightarrow} f(x,y) = g(x)h(y)$$

このとき

●$g(x|y) \equiv g(x)$　　(Yの出方によらない)

●$h(y|x) \equiv h(y)$　　(Xの出方によらない)

●$E[XY] = E[X]E[Y]$

$\therefore Cov[X,Y] = E[XY] - E[X]E[Y] = 0$

$\therefore \sigma_{XY} = 0$

(独立 ⇒ 無相関)

2　確率変数の和

Z = X + Yを考える。

2.1 期待値・分散

期待値

$$E[Z] = E[X] + E[Y]$$

これは常に成り立つ。

分散

$$V[Z] = V[X] + 2Cov[X,Y] + V[Y]$$

X,Yが独立のときは

$$Cov[X,Y] = 0$$

ゆえ

$$V[Z] = V[X] + V[Y]$$

が成立する。

[19] 申し訳ないが、著者は条件付き確率密度分布(連続型)についてあまりよく理解していない。

※確率変数をn個に拡張する。

$$Z = X_1 + X_2 + \cdots + X_n = \sum_i^n X_i$$

を考える。$X_1, X_2, \cdots, X_n$が全て独立のとき、

$$E[Z] = \sum_i^n E[X_i] \ , V[Z] = \sum_i^n V[X_i]$$

2.2 同一分布に従う確率変数

$X_1, X_2, \cdots, X_n$が全て独立で、平均μ、分散σ^2の同一の分布に従うとする。すなわち、

$$E[X_i] = \mu \ , V[X_i] = \sigma^2 \qquad (i = 1\sim n)$$

このとき、

$$Z = X_1 + X_2 + \cdots + X_n = \sum_i^n X_i$$

を考えると、

$$E[Z] = \sum_i^n E[X_i] = n\mu \quad , V[Z] = \sum_i^n V[X_i] = n\sigma^2$$

となる。

さらに、n個の確率変数$X_1, X_2, \cdots, X_n$の相加平均

$$\bar{X} = \frac{1}{n}\sum_i^n X_i = \frac{Z}{n}$$

について考える。このとき、

$$E[\bar{X}] = \frac{1}{n}\sum_i^n E[X_i] = \mu$$

$$V[\bar{X}] = \left(\frac{1}{n}\right)^2 \sum_i^n V[X_i] = \frac{\sigma^2}{n}$$

ここで、$n \to \infty$としたとき、$V[\bar{X}] \to 0$となるのは、重要である[20]。

[20] 標本のサイズが大きいほど、調査の精度がよくなることは、この事実による。このことを定理にしたのが、大数の法則である。

2.3 和の確率分布

確率変数X, Yが独立であるとする。X, Yはそれぞれ確率分布$g(x), h(y)$に従っているとする。

Z = X + Yの確率分布

(離散型)

$$k(z) = P(Z = X + Y) = \sum_{x} g(x)h(z - x)$$

(連続型)

$$k(z) = P(Z = X + Y) = \int_{-\infty}^{\infty} g(x)h(z - x)dx$$

このように、g, hから$k = g * h$を作る操作を『たたきこみ』という。

g, hが同一種類の確率分布に従っていて、たたきこみの結果として得られた$k = g * h$もそれと同一の確率分布となるとき、その確率分布は再生性を持つという。

今まで登場した確率分布では、

二項分布・ポアソン分布・指数分布・幾何分布・正規分布

などが再生性を持つ。特に大事なのは、二項分布・ポアソン分布・正規分布である。

$$Bi(n, p) * Bi(m, p) = Bi(n + m, p)$$
$$Po(\lambda) * Po(\mu) = Po(\lambda + \mu)$$
$$N(\mu_1, {\sigma_1}^2) * N(\mu_2, {\sigma_2}^2) = N(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$$

演習問題 16　正規分布からの標本平均(倉田(1)03,(2)05,(3)04&06)

(1)20 歳の男子日本人の胸囲は平均が 86.8cm、標準偏差 4.80cm の正規分布で表わされるとする。無作為に 16 人を選ぶとき、16 人の胸囲の平均が 85cm 以下となる確率を求めよ。

(2)小学校 6 年生の女子の身長は、平均 147.0cm、標準偏差 6.8cm の正規分布で表わされることが知られている。無作為に 9 人を選ぶとき、9 人の身長の平均が 145cm 以上 150cm 以下となる割合を求めよ。

(3)出生男児の体重は、平均 3.2kg、標準偏差 0.4kg の正規分布にしたがうとする。16 人の出生男児を無作為に選ぶとき、16 人の体重平均が 3.35kg 以下になる確率を求めよ。

演習問題 17　同時確率分布(倉田 06)

右の表は 2 次元離散型確率変数(X,Y)の同時確率分布を表す。以下の各問に答えよ。

(1)X と Yの周辺分布をそれぞれ求めよ。

(2)X の期待値 E[X]と分散 V[X]を求めよ。

(3)X と Y の共分散Cov[X,Y]を求めよ。

(4)X と Yは独立か。理由を付して答えよ。

X＼Y	1	2	3
0	3/20	1/10	3/20
2	1/10	0	1/10
3	3/20	1/10	3/20

演習問題 18　同時確率分布・二項分布(倉田 03)

壺の中に赤、青、緑の 3 種類の玉が 2 個ずつ計 6 個入っているものとする。その壺から 2 個取り出す試行を考える。取り出された赤玉の数をX、青玉の数をYとすると、X と Yはともに離散型の確率変数であり、各々0,1,2 のいずれかの値を取る。X と Yの同時確率分布は下の表の通りである。以下の各問に答えよ。

赤X＼青 Y	0	1	2	行和
0	1/15	4/15	1/15	2/5
1	4/15	4/15	0	8/15
2	1/15	0	0	1/15
列和	2/5	8/15	1/15	1

(1)X と Y の共分散Cov[X,Y]を計算せよ。(分数で)

(2)X と Y の相関係数ρ_{XY}を計算せよ。(途中計算はすべて分数でおこない、結果は小数とせよ)

(3)(2)の結果を解釈せよ。

(4)問題文中の試行を 90 回行ない、赤の個数と青の個数が等しくなる回数をZ とおくとき、Zの確率分布および平均E[Z]、分散 V[Z]を求めよ。

第五講 母集団と標本

教科書:第九章

§5.1　推測統計の基本概念

第一講・第二講で取り上げた記述統計は、分析対象集団の全体について調査することで得たデータを分析・要約することが主たる内容であった。

しかし、調べたい集団全体を調査することは困難が伴う場合が多い。

このようなときは、全体の中から『ランダム』に部分的なデータを取り、その部分的なデータを要約・分析して、全体の性質を推測する。このとき、全体からランダムに選ばれていないと、部分は全体を正しく反映しない。

5.1.1　母集団と標本

・母集団　…分析対象集団全体、またはそのデータ
・標本　　…母集団から選ばれた部分、またはそのデータ
・標本抽出…母集団から標本を選ぶこと。

標本の分析結果は、抽出の仕方に依存する[21]

5.1.2　母集団分布と標本分布

・推測統計の目的…母集団が従う分布、すなわち母集団分布の把握
・標本$X_1, X_2, \cdots, X_n$…母集団分布の中からランダムに選ばれる

⟶ 各標本X_iは母集団分布に従う確率変数である

また統計学では、ほとんどの場合、無限母集団を仮定するので、一回の抽出は他の抽出に影響されない。従って各X_iは互いに独立である。

標本$X_1, X_2, \cdots, X_n$は、同一の母集団分布にしたがう独立な確率変数である

[21] 教科書 P.178 の表 9.2、9.3 を参照

5.1.3　(ノン)パラメトリック・(非)復元抽出

パラメトリック
母集団分布がどのような分布であるかが、標本調査の前から分かっている状況のことを指す。
事前に『母集団分布は○○分布である』と分かっていれば、いくつかの定数を求めるだけでその母集団分布の様子を完全に把握できる。このような定数を母数という。

ex.(1)母集団分布:ポアソン分布$Po(\lambda)$　　(2)母集団分布:正規分布$N(\mu,\sigma^2)$
母数: λ　　母数: μ,σ^2

一方、事前に母集団分布の分かっていない状況をノン・パラメトリックという。

復元抽出
一度抽出した標本を元に戻して次の標本を抽出する、というのを繰り返す方法
一方、一度抽出した標本は戻さないで次の標本を抽出する方法を非復元抽出という。

基礎統計の範囲内では、ほぼ無限個の母集団を想定しているので、非復元抽出を考える。

§5.2　母数と統計量

5.2.1　統計量
・統計量…大きさnの標本データ$X_1,X_2,\cdots,X_n$を要約したもの。標本(⇒データとして手元にある)を要約したものゆえ未知のパラメータは含まない。

ex. ・標本$X_1,X_2,\cdots,X_n$の平均

$$\bar{X}=\frac{1}{n}\sum_{i}^{n}X_i$$

・標本$X_1,X_2,\cdots,X_n$の分散

$$S^2=\frac{1}{n}\sum_{i}^{n}(X_i-\bar{X})^2$$

標本$X_1,X_2,\cdots,X_n$は母集団分布に従う確率変数であり、統計量はそれらの関数であるので、この統計量も何らかの確率分布に従っている。この分布を標本分布という。

5.2.2　標本平均と標本分散

母平均μ,母分散σ^2とし、大きさnの標本データ$X_1, X_2, \cdots, X_n$(独立)を考える。

標本平均

$$\bar{X} = \frac{1}{n}\sum_{i}^{n} X_i$$

このとき、

$$E[\bar{X}] = \mu \quad \cdots(5-1), V[\bar{X}] = \frac{\sigma^2}{n} \quad \cdots(5-2)$$

である。

標本分散(不偏分散とも)

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(X_i - \bar{X})^2$$

このとき、

$$E[s^2] = \sigma^2 \quad \cdots(5-3)$$

である。

・式$(5-1), (5-3)$について

式$(5-1)$は、標本平均$\bar{X}$の期待値が母平均に一致することを示している。(式$(5-3)$も同様)

> 『ある母集団からn個の標本を抽出し、その平均$\bar{X}$を計算していく』という操作の繰り返しを考える。各標本抽出によって計算された平均$\bar{X}$は(抽出の仕方によって)母平均μより大きかったり、小さかったりする。しかし、全ての抽出パターンを考えれば、$\bar{X}$は母平均μを中心として均等にばらついている、ということを主張している。

このように、統計量$T(X_1, X_2, \cdots, X_n)$の期待値$E[T]$が母数tと一致するとき、この統計量Tは母数tの不偏推定量であるという[22]。

注意　5.2.1 の$ex.$で提示した

$$S^2 = \frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^2$$

について、$E[S^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$となり、$S^2$は$\sigma^2$の不偏推定量とはならない。

[22] 第八講の点推定を参照せよ。

・式$(5-2)$について

$$V[\bar{X}] = \frac{\sigma^2}{n} \to 0 \quad (n \to \infty)$$

標本のサイズnが大きくなればなるほど、標本平均$\bar{X}$のμからの散らばり具合は小さくなっていく。すなわち、各抽出における標本平均は母平均に近づく、ということを主張している。

第六講 大数の法則と中心極限定理

教科書:第八章

§6.1　大数の法則

第五講の式$(5-2)$について

$$V[\bar{X}] = \frac{\sigma^2}{n} \to 0 \quad (n \to \infty)$$

という事実を、定理の形にしたのが、大数の法則[23]である。

大数の法則

母平均μ,母分散σ^2とし、大きさnの標本データ$X_1, X_2, \cdots, X_n$(独立)について、その相加平均を$\bar{X}_n$とする。このとき以下の式が成り立つ。

$$n\text{に依らない勝手な}\varepsilon > 0 \text{ に対して、 } \lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$[24]

(証明)

$$E[\bar{X}_n] = \mu\,, V[\bar{X}_n] = \frac{\sigma^2}{n}$$

であるので、チェビシェフの不等式に当てはめると、

$$P\left(|\bar{X}_n - \mu| \geqq k\sqrt{\frac{\sigma^2}{n}}\right) \leqq \frac{1}{k^2} \qquad (k\text{は正の任意実数}) \quad \cdots(6-1)$$

ここで、

$$k\sqrt{\frac{\sigma^2}{n}} = \varepsilon \Leftrightarrow \frac{1}{k^2} = \frac{1}{\varepsilon^2}\cdot\frac{\sigma^2}{n}$$

とおくと、そもそもkはnに依らない任意正数として取ったので、εもnに依らない任意正数として取れる。このとき式$(6-1)$は

$$0 \leqq P(|\bar{X}_n - \mu| \geqq \varepsilon) \leqq \frac{1}{\varepsilon^2}\cdot\frac{\sigma^2}{n} \xrightarrow{(n\to\infty)} 0$$

はさみうちの原理より

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geqq \varepsilon) = 0$$
$$\Leftrightarrow \lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$

■

---

[23] ×「だいすうのほうそく」→○「たいすうのほうそく」

[24] $\bar{X}_n$はμに確率収束する、と言う。単なる「収束」ではないのは教科書 P.158 の図 8.2~8.5(特に 8.5)を見ていただきたい。「必ずしも収束するのではない」くらいの気持ちで良いだろう。(と思われる)

§6.2　中心極限定理[25]

平均μ、分散σ^2の任意の確率分布にしたがう標本[26]に対する、中心極限定理の大まかな主張は

> 標本サイズnが十分大きいとき、その標本の和は、近似的に
>
> $$N(n\mu, n\sigma^2)$$
>
> にしたがう

ということである。数式で書けば、

$$P\left(a \leqq \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sqrt{n}\sigma} \leqq b\right) \xrightarrow{(n\to\infty)} \int_a^b \left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right\} dx$$

となる。これの系として

> 標本平均
>
> $$\bar{X}_n = \frac{1}{n}\sum_i^n X_i$$
>
> はnが十分大きいとき、近似的に
>
> $$N(\mu, \frac{\sigma^2}{n})$$
>
> にしたがう

とも言える。

$$P\left(c \leqq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leqq d\right) \xrightarrow{(n\to\infty)} \int_c^d \left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right\} dx$$

---

[25] 証明は数理統計学 etc の教科書を参照せよ。

[26] ただし、同一の確率分布から、独立に抽出する。

演習問題 19　再生性・中心極限定理(倉田 05)

ある飛行機の乗客定員は 300 名で、そのうち 30 席はファースト・クラス、270 席はエコノミー・クラスである。この航空会社は完全予約制であり、30 名のファースト・クラスと 290 名のエコノミー・クラスの予約を受け付けている。予約をした人が現れない確率は(クラスに関わらず)0.1 であるとする。また、エコノミー・クラスの乗客をファースト・クラスに割り当てることはできるとする。

(1)現れる乗客の数をXとおく。Xの確率分布を導け。

(2)$E[X]$、$V[X]$はいくらか。(答えのみ)

(3)現れた乗客を全員収容できる確率を求めよ。(近似計算でよい)

※編者注：下線部追加(一応)

第七講 正規分布からの標本とその標本分布

教科書:第十章

§7.1　代表的な標本分布

確率分布$f(x)$において以下の式をみたすx_αのことを上側確率100α%点という。

$$\int_{x_\alpha}^{\infty} f(x)dx = \alpha$$

ex.　標準正規分布$N(0,1)$の上側確率100α%点Z_αは、

$$\int_{Z_\alpha}^{\infty} \left\{ \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right\} dx = \alpha$$

をみたす。

7.1.1　χ^2分布

$Z_1, Z_2, \cdots, Z_k$を独立な、標準正規分布$N(0,1)$にしたがう確率変数とするとき、

$$\chi^2 = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_k}^2$$

がしたがう確率分布を自由度kのχ^2分布$\chi^2(k)$という。確率密度関数は

$$f_k(\chi^2) = \frac{1}{2\Gamma\left(\frac{k}{2}\right)} \left(\frac{\chi^2}{2}\right)^{\frac{k}{2}-1} e^{-\frac{\chi^2}{2}}$$

であり、上側確率100α%点は$\chi_\alpha^2(k)$とかく。

7.1.2　t分布

二つの確率変数YとZが以下の条件を満たす。

①Zは標準正規分布$N(0,1)$にしたがう
②Yは自由度kのχ^2分布$\chi^2(k)$にしたがう
③YとZは独立である

このとき、

$$t = \frac{Z}{\sqrt{\frac{Y}{k}}}$$

がしたがう確率分布は自由度kのt分布$t(k)$という。確率密度関数は

$$f_k(t) = \frac{1}{\sqrt{k}B\left(\frac{1}{2}, \frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}$$

であり、上側確率100α%点は$t_\alpha(k)$とかく。

特徴

・正規分布に似ている

・$t = 0$で対称

・$k \to \infty$で$N(0,1)$に一致する。すなわち

$$\int_a^b f_k(t)\,dx \xrightarrow{(k\to\infty)} \int_a^b \left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right\} dx$$

7.1.3　F分布

二つの確率変数UとVが以下の条件を満たす。

①Uは自由度k_1のχ^2分布$\chi^2(k_1)$にしたがう

②Vは自由度k_2のχ^2分布$\chi^2(k_2)$にしたがう

③UとVは独立である

このとき、

$$F = \frac{\frac{U}{k_1}}{\frac{V}{k_2}} = \frac{k_2}{k_1} \cdot \frac{U}{V}$$

がしたがう確率分布は自由度(k_1, k_2)のF分布$F(k_1, k_2)$という。確率密度関数は

$$f_{k_1,k_2}(F) = \frac{1}{B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)} \left(k_1{}^{\frac{k_1}{2}}\right)\left(k_2{}^{\frac{k_2}{2}}\right)(k_2 + k_1 F)^{-\frac{k_2+k_1}{2}}$$

であり、上側確率100α%点は$F_\alpha(k_1, k_2)$とかく。

特徴

・Fが$F(k_1, k_2)$にしたがうとき、$1/F$は$F(k_2, k_1)$にしたがう

・$F_\alpha(k_1, k_2) \cdot F_{1-\alpha}(k_2, k_1) = 1$

・tが$t(k)$にしたがうとき、t^2は$F(1, k)$にしたがう

§7.2　統計量と標本分布

7.2.1　母分散が既知の状態での標本平均

標本$X_1, X_2, \cdots, X_n$は互いに独立であり、平均μ、分散σ^2の正規分布$N(\mu, \sigma^2)$にしたがうとする。このとき、標本平均

$$\bar{X} = \frac{1}{n}\sum_i^n X_i$$

は、正規分布$N(\mu, \frac{\sigma^2}{n})$にしたがう。よってこれを標準化した

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

は、標準正規分布$N(0,1)$にしたがう。

7.2.2　標本分散

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(X_i - \bar{X})^2$$

に対して、

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \sum_{i}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

が自由度$(n-1)$のχ^2分布$\chi^2(n-1)$にしたがう[27]。

7.2.3　母分散が未知の状態での標本平均

7.2.1,7.2.2 での文字の置き方を踏襲して、

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

とすれば、Z, χ^2はそれぞれ標準正規分布$N(0,1)$、χ^2分布$\chi^2(n-1)$にしたがう。
これより、

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}}$$

$$= \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}}$$

$$= \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

は自由度$(n-1)$のt分布$t(n-1)$にしたがう。

27 $(X_i - \bar{X})/\sigma$は標準正規分布にしたがっていないが、とにかくこのような事実がある。証明はしない。

§7.3　二標本問題

正規分布$N(\mu_X, {\sigma_X}^2)$からサイズmの標本$X_1, X_2, \cdots, X_m$を、正規分布$N(\mu_Y, {\sigma_Y}^2)$からサイズnの標本$Y_1, Y_2, \cdots, Y_n$をそれぞれ独立に抽出したとする。さらに、標本$\{X_1, X_2, \cdots, X_m\}, \{Y_1, Y_2, \cdots, Y_n\}$の標本平均、標本分散を順に

$$\bar{X} = \frac{1}{m}\sum_i^m X_i$$

$$\bar{Y} = \frac{1}{n}\sum_i^n Y_i$$

$${s_X}^2 = \frac{1}{m-1}\sum_i^m (X_i - \bar{X})^2$$

$${s_Y}^2 = \frac{1}{n-1}\sum_i^n (Y_i - \bar{Y})^2$$

とおく。

7.3.1　母分散が既知の状態での標本平均の差

$\bar{X}, \bar{Y}$はそれぞれ$N\left(\mu_X, \frac{{\sigma_X}^2}{m}\right), N\left(\mu_Y, \frac{{\sigma_Y}^2}{n}\right)$にしたがうので、$\bar{X} - \bar{Y}$は$N\left(\mu_X - \mu_Y, \frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}\right)$にしたがう。よって、これを標準化した

$$\mathrm{Z} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}}}$$

は標準正規分布$N(0,1)$にしたがう。

7.3.2　母分散は未知であるが等しいと分かっている状態での標本平均の差

仮定より、

$${\sigma_X}^2 = {\sigma_Y}^2 = \sigma^2$$

とおける。

二種の標本$\{X_1, X_2, \cdots, X_m\}, \{Y_1, Y_2, \cdots, Y_n\}$の合併した標本分散

$$s^2 = \frac{1}{m+n-2}\left\{\sum_i^m (X_i - \bar{X})^2 + \sum_i^n (Y_i - \bar{Y})^2\right\}$$

$$= \frac{(m-1){s_X}^2 + (n-1){s_Y}^2}{m+n-2}$$

を定義する。このとき

$$\chi^2 = \frac{(m+n-2)s^2}{\sigma^2}$$

は自由度$(m+n-2)$のχ^2分布$\chi^2(m+n-2)$にしたがう[28]。
よって、

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{m+n-2}}}$$

$$= \frac{\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma^2}{m}+\frac{\sigma^2}{n}}}}{\sqrt{\frac{\frac{(m+n-2)s^2}{\sigma^2}}{m+n-2}}}$$

$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

は自由度$(m+n-2)$のt分布$t(m+n-2)$にしたがう。

7.3.3　標本分散の比

7.2.2 より、

$$s_X{}^2 = \frac{1}{m-1}\sum_i^m (X_i-\bar{X})^2 \ , s_Y{}^2 = \frac{1}{n-1}\sum_i^n (Y_i-\bar{Y})^2$$

に対して

$\frac{(m-1)s_X{}^2}{\sigma_X{}^2}$は自由度$(m-1)$の$\chi^2$分布$\chi^2(m-1)$にしたがい、

$\frac{(n-1)s_Y{}^2}{\sigma_Y{}^2}$は自由度$(n-1)$の$\chi^2$分布$\chi^2(n-1)$にしたがう。

さらに、$s_X{}^2$と$s_Y{}^2$は独立であるので、

$$F = \frac{\frac{\frac{(m-1)s_X{}^2}{\sigma_X{}^2}}{(m-1)}}{\frac{\frac{(n-1)s_Y{}^2}{\sigma_Y{}^2}}{(n-1)}} = \frac{s_X{}^2}{s_Y{}^2}\cdot\frac{\sigma_Y{}^2}{\sigma_X{}^2}$$

は自由度$(m-1,n-1)$のF分布$F(m-1,n-1)$にしたがう。

[28] これも受け入れてほしい。

第八講 推定

教科書:第十一章

定数ではあるが未知である母数を推定することが統計学の目的である。この文脈において、統計量は推定量とも呼ばれる。母数θを推定量は一般に$\hat{\theta}$と表記する。

§8.1　点推定

母数θを推定するさい、ある一つの値$\hat{\theta}$で推定することを点推定という。$\hat{\theta}$はいくつかの標本、すなわち確率変数の関数として表現される。したがって、その確率変数の実現値が標本抽出によって明らかになれば、$\hat{\theta}$は値として得られることになる。

ここで問題となるのは、どのような統計量を推定量とするか、である。

8.1.1　モーメント法

k個の未知の母数$\theta_1, \theta_2, \cdots, \theta_k$をもつ母集団分布を考える。この母集団から抽出した標本は、この母集団分布にしたがう確率変数である。これをXとし、その 1 次からk次までの(母)モーメント

$$\mu_i = E\left[X^i\right] \quad (i = 1\sim k)$$

を考えると、モーメント母関数は$\theta_1, \theta_2, \cdots, \theta_k$に依存するので、各モーメントも$\theta_1, \theta_2, \cdots, \theta_k$に依存する。したがって、適当な関数$g_i$を用いて

$$\mu_i = g_i(\theta_1, \theta_2, \cdots, \theta_k) \quad (i = 1\sim k)$$

と書ける。

一方、標本$X_1, X_2, \cdots, X_n$を元にして計算した 1 次からk次までの標本モーメントを

$$\widehat{\mu_i} = \frac{1}{n}\sum_j^n {X_j}^i \quad (i = 1\sim k)$$

とする[29]。ここで、母モーメントμ_iと標本モーメント$\widehat{\mu_i}$とが等しいとすれば[30]

$$g_i(\theta_1, \theta_2, \cdots, \theta_k) = \frac{1}{n}\sum_j^n {X_j}^i \quad (i = 1\sim k)$$

この連立方程式を解くことによって得られた$\theta_1, \theta_2, \cdots, \theta_k$が、推定量$\hat{\theta}_i (i = 1\sim k)$となる。

[29] この式が標本モーメントとして選ばれた理由は、形を見てもらえば分かるであろう。

[30] このように仮定することが、モーメント法の最大のポイントである。

例題 12. モーメント法

指数分布$Ex(\lambda)$に従う母集団分布から，ランダムに大きさnの標本を抽出して，パラメータλをモーメント法によって推定せよ. (廣松)

解答

指数分布は平均$1/\lambda$なので,

$$\frac{1}{\hat{\lambda}} = \frac{\sum X}{n} \qquad \therefore \hat{\lambda} = \frac{n}{\sum X} \quad \cdots(\text{答})$$

8.1.2　最尤法

モーメント法では、k次のモーメントまでしか考えなかった。この部分が不十分であるが、それを考慮したのが、最尤法である。

最尤法は、以下の最尤原理という原理の下での考え方である。

最尤原理

現実の標本は、確率最大のものが実現した

再び、k個の未知の母数$\theta_1, \theta_2, \cdots, \theta_k$をもつ母集団分布を考える。やはり、この母集団から抽出した標本は、この母集団分布にしたがう確率変数である。これをXとすると、確率分布関数は

$$f(X, \theta_1, \theta_2, \cdots, \theta_k)$$

とかける。このとき、その標本Xが実現値xを取る確率は

$$f(x, \theta_1, \theta_2, \cdots, \theta_k)$$

である。同様にして、n個の標本$X_1, X_2, \cdots, X_n$として実現値$x_1, x_2, \cdots, x_n$が独立に抽出されたとすると、その確率は尤度関数と呼ばれていて、

$$L(\theta_1, \theta_2, \cdots, \theta_k) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, \cdots, \theta_k)$$

となる。$(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)$が実現したので、最尤原理により、その確率$L(\theta_1, \theta_2, \cdots, \theta_k)$が最大値となっていることが必要。さらに、

$L(\theta_1, \theta_2, \cdots, \theta_k)$が最大値を取っている

$$\Leftrightarrow \ln L(\theta_1, \theta_2, \cdots, \theta_k) = \sum_{\mathrm{i}}^{\mathrm{n}} \{\ln f(x_i, \theta_1, \theta_2, \cdots, \theta_k)\} \text{が最大値を取っている}$$

$$\Rightarrow \frac{\partial}{\partial \theta_i} \ln L(\theta_1, \theta_2, \cdots, \theta_k) = 0 \qquad (i = 1\sim k) \qquad \cdots(8-1)$$

ゆえ、式$(8-1)$のk本の連立方程式を解くことで、$\theta_1, \theta_2, \cdots, \theta_k$の最尤推定量を求めることができる。

8.1.3　点推定の基準[31]

不偏性

推定量$\hat{\theta}$が

$$\mathrm{E}\left[\hat{\theta}\right] = \theta$$

をみたすとき、$\hat{\theta}$は不偏性を持つという。(詳しくは 5.2.2 を見よ)

一致性

サイズnの標本から計算された推定量$\hat{\theta}_n$がθに確率収束する、すなわち

$$^{\forall}\varepsilon > 0\text{に対して、} \lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1$$

をみたすとき、推定量$\hat{\theta}_n$は一致性を持つ、という。

例題 13. 母平均の推定量・t分布

総世帯数9420の市で16世帯の標本を選んで世帯員数を調べたところ,次の結果を得た.

5　4　2　7　6　4　4　4　3　4　3　3　3　5　2　5

(1)市の人口を推定せよ.

(2)正確な人口がこの推定値の±10%の範囲内にある確率を求めよ.　(廣松)

解答

(1)標本平均は母平均の推定量である、すなわち$\hat{\mu} = \bar{\mathrm{X}}$ =4 なので

$$4 \times 9420 = 37680\text{ 人}$$

(2)

『正確な人口がこの推定値の±10%の範囲内にある』

$$\Leftrightarrow \quad \frac{\left|\text{真の人口} - \text{推定値 } 37680\right|}{\text{推定値 } 37680} < 0.1$$

左辺の分母分子を総世帯数 9420 でわって

$$\frac{\left|\text{母平均 } \mu - \text{標本平均 } 4\right|}{\text{標本平均 } 4} < 0.1$$

$$-0.4 < \mu - 4 < 0.4$$

この標本の不偏分散s^2は

$$s^2 = \frac{1}{16-1}\sum(X-4)^2 = \frac{28}{15}$$

よって

$$t = \frac{\bar{X}-\mu}{\sqrt{s^2/n}} = \sqrt{\frac{60}{7}}(\mu - 4)$$

---

[31] モーメント法と最尤法とでは、推定の結果が異なることがある。したがって、点推定における基準というものを考える必要がありそうである。

はt分布(15)にしたがい、いま

$$-\frac{24}{7} < \mathrm{t} < \frac{24}{7} = 1.17\cdots$$

である。$t_{0.15}\,(15) = 1.074$より、$t_{\alpha}(15) = 1.17$をみたすαは 0.15 としてよく、これは上側確率ゆえ,求める確率は$1-0.15\times 2 = 0.7$　…(答)

§8.2　区間推定(文字等の設定は第 7 講に準ずるものとする)

ある区間$[L,U]$が、$100(1-\alpha)$%以上の確率で、未知の定数である母数θを含むような確率変数L,Uを求める推定法を区間推定という[32]。

$1-\alpha$　：信頼係数
L,U：下側信頼限界、上側信頼限界
区間$[L,U]$：$100(1-\alpha)$%信頼区間

$$\mathrm{P}(L \leqq \theta \leqq U) = 1-\alpha$$

$$L \leqq \theta \leqq U$$

という不等式は、確率変数L,Uに関する不等式であって、θに関する不等式ではないことに注意せよ。

8.2.1　母分散が既知の状態での母平均

7.2.1 より、

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

は、標準正規分布$N(0,1)$にしたがう。よって、

$$P\left(-Z_{\frac{\alpha}{2}} \leqq Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

カッコ内の不等式を(μにとって見やすくするために)μに関して解けば

$$\bar{X} - Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \leqq \mu \leqq \bar{X} + Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$

となる。これが$100(1-\alpha)$%信頼区間である[33]。

---

[32] なお,同一の母集団から抽出した標本でも,標本ごとに信頼区間の推定値は変化する．θは未知であるが決まった定数である．したがって,一つの標本から信頼区間を具体的な数値として推定してやれば,これは信頼区間に含まれるか否かのいずれしかない．すなわち,具体的に数値として計算した現実の信頼区間にたいして“$1-\alpha$の確率でθを含む”ということはない．信頼区間の意味は,繰り返し多くの異なった標本について信頼区間を計算した場合,θを区間内に含むものの割合が$1-\alpha$となるということである

[33] 信頼区間の幅は$\sqrt{n}$に反比例していく。すなわち、同じ信頼係数でも、標本サイズが大きいほど信頼区間の幅は狭くなり、推定誤差は小さくなる。このことは、以下の節で求める信頼区間にも通じる、大事な性質である。

例題 14. 母分散が既知の状態での母平均の区間推定

母平均＝μ、母分散＝25の正規母集団について,以下の問いに答えよ.

(1)この母集団から大きさnの標本を取り出して,母平均μを信頼水準99%で区間推定したい.信頼区間の幅を4以内にするためにはnの大きさをどのようにとればよいか.

(2)この母集団から取り出した大きさ9の標本の観測値が

32，39，34，45，39，29，36，42，37，

であったとする.母平均μに関する 95%信頼区間を求めよ. (廣松)

解答

(1)

母平均μ,母分散 25 とわかっている

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{5/\sqrt{n}}$$

は標準正規分布にしたがう。$Z_{0.005} = 2.58$より母平均μの 99%信頼区間は

$$|Z| < Z_{0.005} = 2.58$$

$$\Leftrightarrow \bar{X} - 2.58\frac{5}{\sqrt{n}} < \mu < \bar{X} + 2.58\frac{5}{\sqrt{n}}$$

信頼区間の幅は

$$5.16\frac{5}{\sqrt{n}}$$

よって題意より

$$5.16\frac{5}{\sqrt{n}} < 4 \Leftrightarrow n > 41.6025 \quad \therefore n\text{のサイズは 42 以上} \quad \cdots(\text{答})$$

(2)

標本平均$\bar{X}$ =37

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{37 - \mu}{5/3}$$

は標準正規分布にしたがう。よって 95%信頼区間は

$$|Z| < Z_{0.025} = 1.96$$

$$\Leftrightarrow 33.733\cdots < \mu < 40.266\cdots \qquad \cdots(\text{答})$$

8.2.2　母分散が未知の状態での母平均

7.2.3 より、

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

は自由度$(n-1)$のt分布$t(n-1)$にしたがう。よって、

$$P\left(-t_{\frac{\alpha}{2}} \leqq t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leqq t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

カッコ内の不等式をμに関して解けば

$$\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leqq \mu \leqq \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

となる。これが$100(1-\alpha)$%信頼区間である。

例題 15. 母分散が未知の状態での母平均の区間推定

母平均,母分散が未知の正規母集団から取り出した大きさ9個の標本の観測値が

17,19,20,22,22,23,24,25,26

であったとする．母平均μに関する信頼水準95%の信頼区間を求めよ．　(廣松)

解答

標本平均$\bar{X} = 22$

不偏分散$s^2 = 72$

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} = \frac{22 - \mu}{2\sqrt{2}}$$

はt分布(8)に従う．$t_{0.025}(8) = 2.306$より 95%信頼区間は

$$|t| < t_{0.025}(8) = 2.306$$

$$\Leftrightarrow 15.477 < \mu < 28.522 \quad \cdots(\text{答})$$

8.2.3　母分散

7.2.2 より、

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \sum_i^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

が自由度$(n-1)$のχ^2分布$\chi^2(n-1)$にしたがう。よって、

$$P\left(\chi^2_{1-\frac{\alpha}{2}} \leqq \chi^2 = \frac{(n-1)s^2}{\sigma^2} \leqq \chi^2_{\alpha/2}\right) = 1 - \alpha$$

カッコ内の不等式をσ^2に関して解けば

$$-\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

となる。これが$100(1-\alpha)$%信頼区間である。

8.2.4　母分散が既知の状態での母平均の差(二標本問題)

7.3.1 より

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{m}+\frac{{\sigma_Y}^2}{n}}}$$

は標準正規分布$N(0,1)$にしたがう。よって

$$P\left(-Z_{\frac{\alpha}{2}} \leqq Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{m}+\frac{{\sigma_Y}^2}{n}}} \leqq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

カッコ内の不等式を$(\mu_X-\mu_Y)$に関して解けば

$$(\bar{X}-\bar{Y}) - Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{{\sigma_X}^2}{m}+\frac{{\sigma_Y}^2}{n}} \leqq \mu_X-\mu_Y \leqq (\bar{X}-\bar{Y}) + Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{{\sigma_X}^2}{m}+\frac{{\sigma_Y}^2}{n}}$$

となる。これが$100(1-\alpha)$%信頼区間である。

8.2.5　母平均が未知であるが等しいと分かっている状態での母平均の差(二標本問題)

7.3.2 より

$$t = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

は自由度$(m+n-2)$のt分布$t(m+n-2)$にしたがう。よって

$$P\left(-t_{\frac{\alpha}{2}} \leqq t = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s\sqrt{\frac{1}{m}+\frac{1}{n}}} \leqq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

カッコ内の不等式を$(\mu_X-\mu_Y)$に関して解けば

$$(\bar{X}-\bar{Y}) - t_{\frac{\alpha}{2}}\cdot s\sqrt{\frac{1}{m}+\frac{1}{n}} \leqq \mu_X-\mu_Y \leqq (\bar{X}-\bar{Y}) + t_{\frac{\alpha}{2}}\cdot s\sqrt{\frac{1}{m}+\frac{1}{n}}$$

となる。これが$100(1-\alpha)$%信頼区間である。

8.2.6　母分散の比(二標本問題)

7.3.3 より

$$F = \frac{{s_X}^2}{{s_Y}^2} \cdot \frac{{\sigma_Y}^2}{{\sigma_X}^2}$$

は自由度$(m-1, n-1)$のF分布$F(m-1, n-1)$にしたがう。よって

$$P\left(F_{1-\frac{\alpha}{2}} \leqq F = \frac{{s_X}^2}{{s_Y}^2} \cdot \frac{{\sigma_Y}^2}{{\sigma_X}^2} \leqq F_{\frac{\alpha}{2}}\right) = 1-\alpha$$

カッコ内の不等式を$\frac{{\sigma_Y}^2}{{\sigma_X}^2}$に関して解けば

$$F_{1-\frac{\alpha}{2}} \cdot \frac{{s_Y}^2}{{s_X}^2} \leqq \frac{{\sigma_Y}^2}{{\sigma_X}^2} \leqq F_{\frac{\alpha}{2}} \cdot \frac{{s_Y}^2}{{s_X}^2}$$

となる。これが$100(1-\alpha)$%信頼区間である。

§8.3　中心極限定理の応用

これまでは、正規母集団からの標本について考えた。しかし中心極限定理を用いれば、他の分布からの標本についても同様の議論ができる。

8.3.1　二項分布への応用：母比率の推定

サイズnの標本$X_1, X_2, \cdots, X_n$を互いに独立に、二項分布$Bi(1, p)$から抽出したとする。このとき、

$$E[X_i] = p\,, V[X_i] = p(1-p) \quad (i = 1,2,\cdots,n)$$

である。したがって中心極限定理より、標本比率

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

は、nが十分大きいとき、近似的に

$$\text{正規分布} N\left(p, \frac{p(1-p)}{n}\right)$$

にしたがう。よって、これを標準化した

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

は、正規分布$N(0,1)$にしたがう。さらにいま、nが十分大きいので、分母において$\hat{p} \fallingdotseq p$としてよく[34]、そうすれば、

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

が(近似的に)正規分布$N(0,1)$にしたがう。以上より、

[34] 点推定によって、標本比率$\hat{p}$は母比率pの、一致推定量であることが言える。

$$P\left(-Z_{\frac{\alpha}{2}} \leqq Z = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leqq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

よって母比率pの$100(1-\alpha)$%信頼区間は、

$$\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqq p \leqq \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

となる。

例題 16. 母比率の区間推定

(1)ある工場で多数の製品の中から400個を任意に取り出して調べたところ,16個の不良品があった. 製品全体についての不良品率pを信頼係数95%で区間推定せよ.

(2)この工場の不良品率が(1)で求めた標本における比率とほぼ同じであると仮定して,その信頼区間の長さを0.01以下にするためには,何個の標本を抽出する必要があるか. (廣松)

解答

(1)$\hat{p} = 0.04 \quad n = 400$ より

$$Z = \frac{0.04-p}{\sqrt{\frac{0.0384}{400}}}$$

は、標準正規分布に近似的にしたがう。よって 95%信頼区間は

$$|Z| < Z_{0.025} = 1.96$$

$$\Leftrightarrow \quad 0.02642072173\cdots < p < 0.0535792\cdots$$

(2)題意より $\hat{p} = 0.04$のままである。したがって、標本数をnとすれば、

$$Z = \frac{0.04-p}{\sqrt{\frac{0.0384}{n}}}$$

が標準正規分布に近似的にしたがう。よって 95%信頼区間は

$$|Z| < Z_{0.025} = 1.96$$

$$0.04 - 1.96\sqrt{\frac{0.0384}{n}} < p < 0.04 + 1.96\sqrt{\frac{0.0384}{n}}$$

これの区間幅が 0.01 以下なので

$$2 \times 1.96\sqrt{\frac{0.0384}{n}} \leqq 0.01$$

$n \geqq 2950.3\cdots$ よってnは 2951 個以上 $\cdots$(答)

8.3.2　ポアソン分布への応用

サイズnの標本$X_1, X_2, \cdots, X_n$を互いに独立に、ポアソン分布$Po(\lambda)$から抽出したとする。このとき、

$$E[X_i] = \lambda\,, V[X_i] = \lambda \quad (i = 1,2,\cdots,n)$$

である。したがって中心極限定理より、標本平均

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

は、nが十分大きいとき、近似的に

$$\text{正規分布} N\left(\lambda, \frac{\lambda}{n}\right)$$

にしたがう。よって、これを標準化した

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

は、正規分布$N(0,1)$にしたがう。さらにいま、nが十分大きいので、分母において$\hat{\lambda} \fallingdotseq \lambda$としてよく[35]、そうすれば、

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}}$$

が正規分布$N(0,1)$にしたがう。以上より、

$$P\left(-Z_{\frac{\alpha}{2}} \leqq Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} \leqq Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

よって母平均λの$100(1-\alpha)$%信頼区間は、

$$\hat{\lambda} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\lambda}}{n}} \leqq \lambda \leqq \hat{\lambda} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\lambda}}{n}}$$

となる。

[35] 点推定によって、標本平均$\hat{\lambda}$は母平均λの、一致推定量であることが言える。

例題 17. ポアソン分布における区間推定

1時間毎の電話の受信数を記録したところ

4,3,5,4,8,2,5,9,3,5

であった。ポワソン分布$Po(\lambda)$を仮定して,λに関する信頼係数99%の信頼区間を求めよ。 (廣松)

解答

$\hat{\lambda} = 4.8\,, n = 10$ より

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} = \frac{4.8 - \lambda}{\sqrt{\frac{4.8}{10}}} = \frac{4. - \lambda}{0.4\sqrt{3}}$$

が近似的に正規分布$N(0,1)$にしたがう。よって 99%信頼区間は

$$|Z| \leqq Z_{0.005} = 2.58$$

$$3.0125235744\cdots \leqq Z \leqq 6.5874764256\cdots \qquad \cdots (答)$$

例題 18. チェビシェフの不等式と中心極限定理の比較・中心極限定理を利用した区間推定の手法

細工のされていないサイコロをn回投げて 1 の目が出る回数をrとする。r/nと 1/6 の差が 1/100 以下となる確率を 0.99 より大きくするためにはnを何回以上にすればよいか。分布の形を考慮した場合とそうでない場合とを比較せよ。 (廣松)

解答

『r/nと 1/6 の差が 1/100 以下となる確率が 0.99 より大きい』とは、数式では

$$P\left(\left|\frac{r}{n} - \frac{1}{6}\right| \leqq \frac{1}{100}\right) \geqq 0.99$$

となる。

【分布の形を考慮しない場合】

rは，二項分布$Bi(n,\frac{1}{6})$にしたがうので、$\mathrm{E}[r] = \frac{n}{6}\,, \mathrm{V}[r] = n \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5n}{36}$である。

したがって$\frac{r}{n}$は$\mathrm{E}\left[\frac{r}{n}\right] = \frac{1}{6}\,, \mathrm{V}\left[\frac{r}{n}\right] = \frac{5}{36n}$をみたす。

これをチェビシェフの不等式に適応させて

$$P\left(\left|\frac{r}{n} - \frac{1}{6}\right| \leqq k\sqrt{\frac{5}{36n}}\right) \geqq 1 - \frac{1}{k^2}$$

$k = 10$とすれば

$$P\left(\left|\frac{r}{n}-\frac{1}{6}\right| \leqq 10\sqrt{\frac{5}{36n}}\right) \geqq 0.99$$

よって、題意より

$$10\sqrt{\frac{5}{36n}} \leqq \frac{1}{100} \Leftrightarrow n \geqq 138888.88\cdots$$

以上より、nは 138889 回以上　　$\cdots$(答)

【分布の形を考慮する場合】

$\frac{r}{n}$が $\mathrm{E}\left[\frac{r}{n}\right] = \frac{1}{6}$, $\mathrm{V}\left[\frac{r}{n}\right] = \frac{5}{36n}$をみたすのは同じ。

nが十分大きいとき,中心極限定理より$\frac{r}{n}$は正規分布$\mathrm{N}(\frac{1}{6}, \frac{5}{36n})$に従うと近似できる。

したがって,

$$Z = \frac{\frac{r}{n}-\frac{1}{6}}{\sqrt{\frac{5}{36n}}}$$

は標準正規分布N(0,1)にしたがう。

[$\frac{\mathrm{r}}{\mathrm{n}}$の 99%信頼区間を求める方法を応用]

$$|\mathrm{Z}| \leqq \mathrm{Z}_{0.005} = 2.58$$

$$\left|\frac{r}{n}-\frac{1}{6}\right| \leqq Z_{0.005}\sqrt{\frac{5}{36n}} = 2.58\sqrt{\frac{5}{36n}}$$

これは上側確率の定義より、

$$P\left(\left|\frac{r}{n}-\frac{1}{6}\right| \leqq Z_{0.005}\sqrt{\frac{5}{36n}} = 2.58\sqrt{\frac{5}{36n}}\right) = 1-0.005\times 2 = 0.99$$

を意味する。よって題意の条件より

$$2.58\sqrt{\frac{5}{36n}} \leqq \frac{1}{100} \Leftrightarrow n \geqq 9245 \quad \cdots(\text{答})$$

このとき,中心極限定理による近似は妥当である。

演習問題 20　区間推定(倉田 03)

ある爬虫類の体長を調べるため、20 頭を捕獲し、体長$X_1, X_2, \cdots, X_{20}$(単位は cm)を測定したところ、

$$\text{標本平均}\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X_i = 30.5\text{ , 標本分散}s^2 = \frac{1}{19}\sum_{i=1}^{20}(X_i - \bar{X})^2 = 17.64\text{ , 標本標準偏差}s = \sqrt{s^2} = 42$$

であった。正規母集団$N(\mu, \sigma^2)$を仮定して、以下の各問に答えよ。

(1)母平均μの点推定値として$\overline{X}$の実現値 30.5 を用いるのが通常であるが、このことの根拠は何か。簡潔に述べよ。(1 行程度)

(2)母分散σ^2は未知であるとして、母平均μの信頼係数 0.95 の信頼区間を作れ。

(3)母分散σ^2は$\sigma^2 = 16$であると分かっているものとして、母平均μの信頼係数 0.95 の信頼区間を作れ。

(4)母分散σ^2の信頼係数 0.95 の信頼区間を作れ。

演習問題 21　二標本問題(倉田 06)

タイヤメーカーが新製品のタイヤを装着したときの停止距離について調べるため、時速 100km でブレーキを踏んだ時の停止距離を計測したとする。計測は晴天時と雨天時にそれぞれ 14 回と 10 回であり、計測された値を晴天時$X_1, X_2, \cdots, X_{14}$、雨天時$Y_1, Y_2, \cdots, Y_{10}$と表す。(単位は m)

$$\text{晴天時の標本平均}\overline{X} = \frac{1}{14}\sum_{i=1}^{14} X_i = 44.2\text{ , 標本不偏分散}s_1^2 = \frac{1}{13}\sum_{i=1}^{14}(X_i - \bar{X})^2 = 4.2$$

$$\text{晴天時の標本平均}\overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i = 49.6\text{ , 標本不偏分散}s_2^2 = \frac{1}{9}\sum_{i=1}^{10}(Y_i - \bar{Y})^2 = 6.4$$

であった。晴天時、雨天時の停止距離はそれぞれ正規母集団$N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$からの無作為標本と仮定できるとする。以下の各問に答えよ。計算過程で適当に四捨五入してよい。

(1)晴天時の母平均μ_1に関する信頼係数 0.95 の信頼区間を作れ。

(2)晴天時の母分散${\sigma_1}^2$に関する信頼係数 0.95 の信頼区間を作れ。

(3)母分散は等しい(${\sigma_1}^2 = {\sigma_2}^2$)として、母平均の差$\mu_2 - \mu_1$に関する信頼係数 0.95 の信頼区間を作れ。

演習問題 22　母比率の推定(倉田 05)

(1)不偏推定量の定義を述べよ。(1,2 行)

(2)400 世帯を無作為に選んだところ、35%の世帯があるテレビ番組を視聴していた。この番組の視聴率の信頼係数 0.95 の信頼区間を求めよ。

演習問題 23　母比率の推定(倉田 04)

ある種子の発芽率をpとする。100 個の種子を観察したところ、発芽したものは 70 個であった。中心極限定理を用いてpに関する信頼係数 0.95 の信頼区間を作れ。

演習問題 24　チェビシェフの不等式・中心極限定理を利用した手法(倉田 04)

コインを 10000 回投げるとき、表の出る回数が 4850 回以上でかつ 5150 回以下である確率を
(ⅰ)チェビシェフの不等式
(ⅱ)中心極限定理
を用いてそれぞれ評価せよ。

第九講 検定

教科書:第十二章

§9.1　検定の基本的概念

検定の基本的思想は

区間推定で求めた信頼区間に、母数に関する仮説が当てはまっているかを調べる

ということである。

ex.

(再掲)例題 15. 母分散が未知の状態での母平均の区間推定

母平均,母分散が未知の正規母集団から取り出した大きさ9個の標本の観測値が

17,19,20,22,22,23,24,25,26

であったとする．母平均μに関する信頼水準95%の信頼区間を求めよ．　(廣松)

解答

標本平均$\bar{X} = 22$

不偏分散$s^2 = 72$

$$t(\mu) = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} = \frac{22 - \mu}{2\sqrt{2}}$$

はt分布(8)に従う．$t_{0.025}(8) = 2.306$より 95%信頼区間は

$$|t(\mu)| < t_{0.025}(8) = 2.306 \quad \cdots (※)$$
$$\Leftrightarrow 15.477 < \mu < 28.522 \quad \cdots (答)$$

例題 15 で求めた$15.477 < \mu < 28.522$という 95%信頼区間に対して…

・μ＝29という仮説は含まれていない　⇒　仮説は棄却される。
・μ＝20という仮説は含まれている　⇒　仮説は棄却されない。

用語の確認

帰無仮説：ここでのμ＝29、μ＝20という仮説のこと。単に仮説ともいう。

有意水準[36]：$100(1-\alpha)$%信頼区間に対して検定を行うとき、それは有意水準100α%の検定、という。

棄却域　：信頼区間に含まれない領域のこと。

[36] 有意水準が小さい検定で棄却された仮説ほど、信用できない。(逆は成立しない)

実際は、統計量の形のまま判断する。
例題 15 でいえば、(※)式においてμに仮説を代入して、成立するかで判断すればよい。

・仮説$\mu = 29$では…

$$|t(\mu = 29)| = 2.47487373\cdots > t_{0.025}(8) = 2.306$$

ゆえ、有意水準 5%で棄却される。

・仮説$\mu = 20$では…

$$|t(\mu = 20)| = 0.70710678\cdots < t_{0.025}(8) = 2.306$$

ゆえ、有意水準 5%では棄却されない。

§9.2　両側検定・片側検定

9.2.1　区間推定のおさらい

$$『統計量\ \mathrm{X}\ が確率分布\ \mathrm{K}\ にしたがう』 \overset{def}{\iff} P(a \leqq X \leqq b) = \int_a^b K(x)dx$$

したがって、『サイズnの標本を整理した統計量X_nが確率分布 K にしたがう』とは

$$P(a \leqq \mathrm{X}_n \leqq b) = \int_a^b K(x)dx \qquad \cdots(◎)$$

ということである。ここで例えば

$$\int_a^b K(x)dx = 0.95$$

となるようにa, bを上手く取れば、(◎)式は

$$\mathrm{X}_n が区間[a, b]にある確率は\ 95\%$$

ということを意味する。この

$$a \leqq \mathrm{X}_n \leqq b$$

を、定数であるが未知の母数θに関して整理したものを、θの 95%信頼区間というのであった。

さて第八講では、区間$[a, b]$を分布の中心部分にもってきていた。しかし、ここで見たとおり、a, bは

$$\int_a^b K(x)dx = 0.95$$

をみたしていればよいのであって、特に区間$[a, b]$を中心部分にもってくる論理的必然性はない。中心部分にあるのが一番分かりやすいので中心部分にする、程度である。(教科書図 11.6)

,

また、今まで紹介した統計量は、未知母数に関してみれば単調減少であったことに注意。
$ex.$ t統計量は母平均μに関して単調減少である。

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

9.2.2　棄却域(教科書図 12.1~12.2)[37]

検定において、棄却域とは信頼区間に含まれない区間のことなので、9.2.1 でいえば

$$\mathrm{X}_n < a, b < \mathrm{X}_n$$

のことである。棄却域をどこに取るかは、検定の目的によって異なってくる。

そもそも検定とは、帰無仮説を棄却するのが目的で行われる[38]。

前から持っていた帰無仮説を棄却して、別の仮説を採用する

のが目的である[39]。この『別の仮説』のことを対立仮説という。

・帰無仮説$\theta = \theta_0$を、対立仮説$\theta \neq \theta_0$に対して検定する：両側検定

このとき、帰無仮説で与えられたθ_0が、真の母数θより大きく出ているか小さく出ているかはわからない。したがって、とりあえずどちらの場合でも棄却できるように、両側に半分ずつ棄却域を設置しておく。このような検定を両側検定という。

・帰無仮説$\theta = \theta_0$を、対立仮説$\theta < \theta_0$に対して検定する：片側検定

このときも、帰無仮説で与えられたθ_0が、真の母数θより大きく出ているか小さく出ているかはわからない。しかし、対立仮説が$\theta < \theta_0$となっているので、θ_0が真の母数θより小さく出ている場合は、棄却する必要はないし、むしろ棄却したくない。したがって、棄却域を両側に均等に配置する必要もなくなる。では、どこに棄却域を設置するか。

未知母数θを含む統計量を$\mathrm{X}(\theta)$とすると、9.2.1 で見た通り、$\mathrm{X}(\theta)$はθに関して単調減少である。

前述のとおり、検定とは帰無仮説$\theta = \theta_0$を棄却し、対立仮説$\theta < \theta_0$を採用するのが本当の目的である。

本当に対立仮説$\theta < \theta_0$が正しいのであるならば、帰無仮説をウッカリ信じてしまった場合の$\mathrm{X}(\theta_0)$は

$$\mathrm{X}(\theta) > X(\theta_0)$$

をみたす。この、小さく出てしまった統計量が棄却域に入るように、設置すればいい。つまり、棄却域を下側に寄せて設置すれば、より帰無仮説を棄却しやすくなる。

---

37 棄却域をどのように設置するかは、方法論として覚えてしまってもかまわないが、できれば、その理由も理解していただきたい。こんがらがりやすい個所ではあるが。

38 そういう意味で『無に帰る仮説』である。

39 検定を行う人は『一応、帰無仮説を信じるフリをしているけど、実は対立仮説の方が正しいのでは??』と思っている

例題 19. 片側検定

正味220gと書いてある缶詰を30コ調べたところ,平均218g,標準偏差(不偏分散の正の平方根)6.01gであった. μ = 220という帰無仮説を対立仮説μ < 220に対して,有意水準1%で検定せよ. (廣松)

解答

標本平均$\bar{X} = 218$　不偏分散$s^2 = (6.01)^2$　標本サイズ$n = 30$

このとき

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} = \frac{\sqrt{30}}{6.01}(218 - \mu)$$

はt分布(29)に従う．仮説：$\mu = 220$を認めてしまったときのtが

$$t < -t_{0.01}(29) = -2.462$$

であれば仮説は対立仮説μ < 220に対して棄却される。実際は$\mu = 220$ のとき

$$t = -1.82\cdots > -t_{0.01}(29) = -2.462$$

であるので, $\mu = 220$は棄却されない。

※帰無仮説が正しいのに、棄却することを『第一種の誤り』
　帰無仮説が誤っているのに、採択することを『第二種の誤り』という。

§9.3　χ^2検定

χ^2分布は分散に関する推定・検定で利用した。実は、分散でなくとも、『ばらつき』に関する統計量のうち、χ^2分布にしたがうものがあり、それを利用した検定をχ^2検定という。

9.3.1　適合度検定

実験・観察によって観測された標本の分布の仕方が
理論的に仮定された確率分布に適合しているか

を調べる検定を『適合度検定』という。

n個の標本を観察し、k個のカテゴリーに分類するとしよう。以下のような結果が得られたとする。

カテゴリー	A_1	A_2	…	A_k	計
観測度数	f_1	f_2	…	f_k	n

一方、母集団のしたがう確率分布が理論的に仮定されていて、それによると、カテゴリーA_iに分類される確率はp_iであるとされている。

カテゴリー	A_1	A_2	…	A_k	計
理論確率	p_1	p_2	…	p_k	1

標本の総数はnなので、カテゴリーA_iに分類される理論度数はnp_iとなる。

カテゴリー	A_1	A_2	…	A_k	計
理論度数	np_1	np_2	…	np_k	n

観測度数と理論度数をまとめると以下のようになる。

カテゴリー	A_1	A_2	…	A_k	計
観測度数	f_1	f_2	…	f_k	n
理論度数	np_1	np_2	…	np_k	n

この理論度数の観測度数からのばらつきが、ある一定の基準以下ならば、仮定された確率分布は正しい

という思想の下で行うのが、適合度検定である。

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

は自由度$(k-1)$のχ^2分布にしたがうことが知られているので、

$$\chi^2 > \chi_\alpha^2(k-1)$$

であれば、仮定された確率分布は現実に適応していないとして、有意水準 100 α%で棄却され、

$$\chi^2 < \chi_\alpha^2(k-1)$$

であれば、仮定された確率分布は現実と矛盾しないとして、棄却されず、採択される。

適合度検定の原理は、

観測された度数を O(Observed)、理論により期待される度数を E(Expected)とすれば、

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

と要約できる。

例題 20. 適合度検定

3 人の子供がいる 5896 の世帯を男児の数によって分類したところ、次の表が得られた。以下の問いに答えよ

男児数	0	1	2	3	計
世帯数	638	2212	2250	796	5896

(1)この表から、男女の性比は 1：1 であるとみなしてよいか。

(2)性比は未知であるが、二項分布に従うとしてよいか。 (廣松)

解答

(1)

総人数 ：$5896 \times 3 = 17688$

総男児数：$2212 \times 1 + 2250 \times 2 + 796 \times 3 = 9050$

総女児数：$17688 - 9050 = 8638$

男女の出生性比が 1:1 だとすると、以下のような表が得られる

男女	男	女	計
観測度数	9050	8638	17688
理論確率	$1/2$	$1/2$	1
理論度数	8844	8844	17688

よって

$$\chi^2 = \frac{(9050-8844)^2}{8844} + \frac{(8638-8844)^2}{8844} = 9.59656264133\cdots$$

は自由度 1 のχ^2分布にしたがう。いま

$$\chi^2 = 9.59656264133\cdots > \chi^2_{0.01}(1) = 6.63490$$

より、有意水準 1%でこの仮説は棄却される。

(2)

1 回の出産で男児が生まれる比率は標本比率と一致するとできるので、$p = \frac{9050}{17688} = \frac{4525}{8844}$とできる。よって 3 回の出産で生まれる男児数が

$$\text{二項分布} Bi(3, \frac{4525}{8844})$$

にしたがうとすれば、

子供 3 人のうち男児が 0 人となる確率

$$p_0 = {}_3C_0 \left(\frac{4525}{8844}\right)^0 \left(\frac{4319}{8844}\right)^3 = \frac{4319^3}{8844^3}$$

子供 3 人のうち男児が 1 人となる確率

$$p_1 = {}_3C_1\left(\frac{4525}{8844}\right)^1\left(\frac{4319}{8844}\right)^2 = \frac{3\cdot 4525\cdot 4319^2}{8844^3}$$

子供 3 人のうち男児が 2 人となる確率

$$p_2 = {}_3C_2\left(\frac{4525}{8844}\right)^2\left(\frac{4319}{8844}\right)^1 = \frac{3\cdot 4525^2\cdot 4319}{8844^3}$$

子供 3 人のうち男児が 3 人となる確率

$$p_3 = {}_3C_3\left(\frac{4525}{8844}\right)^3\left(\frac{4319}{8844}\right)^0 = \frac{4525^3}{8844^3}$$

なので、観測度数・理論確率・理論度数に関して以下の表を得る。

男児数	0	1	2	3	計
観測度数	638	2212	2250	796	5896
理論確率	$\frac{4319^3}{8844^3}$	$\frac{3\cdot 4525\cdot 4319^2}{8844^3}$	$\frac{3\cdot 4525^2\cdot 4319}{8844^3}$	$\frac{4525^3}{8844^3}$	1
理論度数	$5896\times\frac{4319^3}{8844^3}$ $= 686.69$	$5896\times\frac{3\cdot 4525\cdot 4319^2}{8844^3}$ $= 2158.33$	$5896\times\frac{3\cdot 4525^2\cdot 4319}{8844^3}$ $= 2261.27$	$5896\times\frac{4525^3}{8844^3}$ $= 789.71$	5896

よって

$$\chi^2 = \frac{(683-686.69)^2}{686.69} + \frac{(2212-2158.33)^2}{2158.33} + \frac{(2250-2261.27)^2}{2261.27} + \frac{(796-789.71)^2}{789.71}$$

$$= 1.46067920902\cdots$$

は自由度 3 のχ^2分布にしたがう。いま

$$\chi^2 = 1.46067920902\cdots < \chi^2_{0.01}(3) = 11.3449$$

より、有意水準 1%で棄却されず、採択される。よって二項分布にしたがっているとしてよい。

9.3.2　独立性の検定

適合度検定の応用として、独立性の検定というものがある。以下のような分割表を考えよう。

カテゴリー	B_1	B_2	$\cdots$	B_l	計
A_1	f_{11}	f_{12}	$\cdots$	f_{1l}	f_{A1}
A_2	f_{21}	f_{22}	$\cdots$	f_{2l}	f_{A2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	f_{k1}	f_{k2}	$\cdots$	f_{kl}	f_{Ak}
計	f_{B1}	f_{B2}	$\cdots$	f_{Bl}	n

ただし

$$f_{Ai} = \sum_{j=1}^{l} f_{ij}\ , f_{Bj} = \sum_{i=1}^{k} f_{ij}\ , n = \sum_{i=1}^{k} f_{Ai} = \sum_{j=1}^{l} f_{Bj}$$

とした。すなわちf_{Ai}、f_{Bj}はカテゴリーA, Bの周辺度数分布である。

ここで

カテゴリーA, Bが独立である

という仮説を考える。Tea Break Ⅱより、

カテゴリーA, B が独立である

$$\overset{def}{\Leftrightarrow} \text{全ての} i, j \text{に関して、} P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$$

であり、周辺度数分布を用いて

$$P(A_i) = \frac{f_{Ai}}{n}\ , P(B_j) = \frac{f_{Bj}}{n}$$

と書ける。したがって、この仮説による、$A_i \cap B_j$の理論度数は

$$n \cdot P(A_i \cap B_j) = n \cdot P(A_i) \cdot P(B_j) = \frac{f_{Ai} \cdot f_{Bj}}{n}$$

となる。

理論度数分布の表

カテゴリー	⋯	B_j	⋯	計
⋮	⋱	⋮	⋱	⋮
A_i	⋯	$\frac{f_{Ai} \cdot f_{Bj}}{n}$	⋯	f_{Ai}
⋮	⋱	⋮	⋱	⋮
計	⋯	f_{Bj}	⋯	n

適合度検定と同様に

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

を計算する。すなわち、

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(f_{ij} - \frac{f_{Ai} \cdot f_{Bj}}{n}\right)^2}{\frac{f_{Ai} \cdot f_{Bj}}{n}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n \cdot f_{ij} - f_{Ai} \cdot f_{Bj}\right)^2}{n \cdot f_{Ai} \cdot f_{Bj}}$$

を考える。これは自由度$(k-1)(l-1)$のχ^2分布にしたがうことが知られているので、あとは適合度検定と同様に処理をする。

例題 21. 独立性の検定

下の分割表は、ある薬剤に関する効果を調べるため、ランダムにマウスを選び,薬剤を投与した場合としなかった場合で、病気を発症の有無を調べた実験の結果である。この実験結果から、薬剤の投与と病気の発症の間に関連が認められるか.有意水準5%で検定せよ。 (廣松)

	投与	非投与	合計
発症	19	27	46
非発症	63	51	114
合計	82	78	160

解答

薬剤の投与をX,病気の発症をYとする

$i = 1$(投与),2(非投与)　　$j = 1$(発症),2(非発症)として

仮説：

X と Y は独立

$$\Leftrightarrow \text{すべての} i,j \text{について} \quad P(X_i \cap Y_j) = P(X_i)P(Y_j) = \frac{f_{Xi} \cdot f_{Yj}}{160^2}$$

を有意水準 5%で検定する。

いま、

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(f_{ij} - 160\frac{f_{Xi} \cdot f_{Yj}}{160^2}\right)^2}{160\frac{f_{Xi} \cdot f_{Yj}}{160^2}}$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(160f_{ij} - f_{Xi} \cdot f_{Yj}\right)^2}{160f_{Xi} \cdot f_{Yj}}$$

$$= \frac{(19 \cdot 160 - 46 \cdot 82)^2}{160 \cdot 46 \cdot 82} + \frac{(63 \cdot 160 - 82 \cdot 114)^2}{160 \cdot 82 \cdot 114} + \frac{(51 \cdot 160 - 114 \cdot 78)^2}{160 \cdot 114 \cdot 78} + \frac{(27 \cdot 160 - 78 \cdot 46)^2}{160 \cdot 78 \cdot 46}$$

$$= 2.55605978\cdots$$

自由度2－1) (2－1) = 1 のχ^2分布にしたがう。いま、

$$\chi^2 = 2.55605978\cdots < \chi^2_{0.05}(1) = 3.84146$$

ゆえ、この仮説は有意水準 5%では棄却されない．つまり薬剤投与と発症には関連が認められない.

演習問題 25　母比率の検定(片側検定)（倉田(1)～(3)03,(4)05)

(1)中心極限定理の主張を書け(2,3 行)

(2)確率変数$X_1, X_2, \cdots, X_n$は互いに独立に同一のベルヌーイ分布$Bi(1,p)$にしたがうものとする。すなわち$P(X_i = 1) = p, P(X_i = 0) = 1 - p$ $(i = 1,2,\cdots,n)$が成立するものとする。この場合に中心極限定理を応用すると、どのような事実が得られるか。

(3)十二指腸虫の感染率が 10%であるとされていたある地域の環境が悪化したため、その地域から改めて、400 人を抽出して感染の有無を調べたところ 56 人の感染者がいた。感染率は変わらないと言えるか。

(4)ある市では一日当たり平均 9 件の交通事故が起こるものとする。市によって事故削減のための対策が行われたとする。対策後の 30 日間の事故件数の平均をとると、7 件であった。対策の効果について述べよ。

演習問題 26　二標本の推定・検定(倉田 04)

ある企業は生産工程で必要となる工業原料を A 社と B 社から購入しているとする。A 社、B 社から購入した工業原料の中からそれぞれ 10 袋を無作為に選び、1 袋当たりの不純物混入率X_i, Y_j(%)を調べたところ$(i = 1\sim10, j = 1\sim10)$、

$$\text{A 社の標本平均}\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 14.4\text{ , 標本不偏分散}s_1^2 = \frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X})^2 = 4.90$$

$$\text{B 社の標本平均}\bar{Y} = \frac{1}{10}\sum_{j=1}^{10} Y_j = 17.9\text{ , 標本不偏分散}s_2^2 = \frac{1}{9}\sum_{j=1}^{10}\left(Y_j - \bar{Y}\right)^2 = 2.10$$

であった。A 社、B 社の標本はそれぞれ正規母集団$N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$からの無作為標本と仮定できるとする。以下の各問に答えよ。計算過程では適当に四捨五入してよい。

(1)A 社の母平均μ_1に関する信頼係数 0.95 の信頼区間を作れ。

(2)A 社の母分散${\sigma_1}^2$に関する信頼係数 0.95 の信頼区間を作れ。

(3)母分散に関して${\sigma_1}^2 = {\sigma_2}^2$が成立するものとして、帰無仮説$\mathrm{H}_0: \mu_1 = \mu_2$を対立仮説$\mathrm{H}_1: \mu_1 \neq \mu_2$に対して有意水準 0.05 で検定せよ。

(4)帰無仮説$\mathrm{H}_0: {\sigma_1}^2 = {\sigma_2}^2$を対立仮説$\mathrm{H}_1: {\sigma_1}^2 > {\sigma_2}^2$に対して有意水準 0.05 で検定せよ。

演習問題 27　二標本の推定・検定(倉田 06)

(1)仮説検定における、第一種の誤りとは何か。定義を述べよ。

32 匹のマウスを 16 匹ずつ A 群、B 群に分け、A 群のマウスには生ピーナッツを与え、B 群のマウスには焼ピーナッツを与えて飼育し、一定期間後に体重X_i, Y_j(g)を調べたところ(i = 1~10, j = 1~10)、

$$\text{A 群の標本平均}\overline{\mathrm{X}} = \frac{1}{16}\sum_{\mathrm{i}=1}^{16} X_i = 59.9\text{ , 標本不偏分散}s_1^2 = \frac{1}{15}\sum_{\mathrm{i}=1}^{16}(X_i - \bar{X})^2 = 21.1$$

$$\text{B 群の標本平均}\overline{\mathrm{Y}} = \frac{1}{16}\sum_{j=1}^{16} Y_j = 55.8\text{ , 標本不偏分散}s_2^2 = \frac{1}{15}\sum_{j=1}^{16}\left(Y_j - \bar{Y}\right)^2 = 9.4$$

であった。A 社、B 社の標本はそれぞれ正規母集団$N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$からの無作為標本と仮定できるとする。以下の各問に答えよ。計算過程では適当に四捨五入してよい。

(2)A 群の母平均μ_1に関する信頼係数 0.95 の信頼区間を作れ。

(3)A 群の母分散${\sigma_1}^2$に関する信頼係数 0.95 の信頼区間を作れ。

(4)帰無仮説$\mathrm{H}_0: {\sigma_1}^2 = {\sigma_2}^2$を対立仮説$\mathrm{H}_1: {\sigma_1}^2 > {\sigma_2}^2$に対して有意水準 0.05 で検定せよ。

(5)(前問の結果にかかわらず)母分散に関して${\sigma_1}^2 = {\sigma_2}^2$が成立するものとして、帰無仮説$\mathrm{H}_0: \mu_1 = \mu_2$を対立仮説$\mathrm{H}_1: \mu_1 \neq \mu_2$に対して有意水準 0.05 で検定せよ。