基礎統計学

記述統計学（きじゅつとうけいがく）

調査対象の全体を**母集団**（population）といい、母集団から抽出されたメンバーを**標本**（sample）という。統計学では母集団から得られた**データ**を扱う。データは数字や文字の羅列であるが、原則として、すべて数字で表現可能である。たとえば男女という性別であれば、男を１、女を０のように表現できる。

統計学の主要な目的はデータを整理して見やすくまとめたり、データを生み出した母集団全体の特性を推論したりすることである。データ分析の目的をどこにおくかによって、記述統計と推測統計に分けることができる。

記述統計は、データのもっている情報を整理して見やすくまとめることを目的としている。記述統計では必ずしも母集団メンバー全員のデータで分析しなくてもよいが、得られたデータがすべてであるという前提で自己完結的に分析を行っている。この点において、母集団メンバーの一部を抽出して、母集団全体の特性を推論する推測統計とは異なっている。

記述統計の手法としては、まず、データのとりうる範囲をいくつかの階級に分け、それぞれの階級に入るデータ数をカウントして度数分布を作成することが挙げられる。また、度数分布をグラフ化したものとしてヒストグラムがある。これは度数に比例して棒の長さを描く棒グラフとは異なり、度数に比例するものは棒（長方形）の面積である。

データがもつ情報をさらに縮約するには、とりうる値のうち最も代表的な値を求めたり、とりうる値の散らばりを示す尺度を求めることである。位置の代表値としては最頻値、中央値、平均値が、散らばりを示す尺度としてはレンジ、四分位偏差、分散などがある。

２変数を同時に扱う場合には、各変数の特徴を表す平均、分散などのほかに、関連を示す尺度が必要である。２変数がどの程度直線関係に近いかという観点から、共分散や相関係数という尺度が用いられる。

推測統計学（すいそくとうけいがく）

推測統計では、母集団から抽出された標本をもとに母集団分布を推測することが目的である。母集団分布は平均や分散などのいくつかの数値によって表現される。そのような母集団分布を特徴づける数値を**母数**（**パラメーター**）という。

標本を用いて母数を推測するためには、標本を観測する前に、母数を推測するための計算式を標本から構成しておく必要がある。その計算式を**統計量**（statistic）という。すなわち、統計量とは標本からなる関数である。

母集団には無限に多くのメンバーがいるとし、そこからn個の標本$(X_1,\cdots,X_n)$を抽出しよう。n個の標本を抽出しても母集団分布は変わらない。i番目の標本をX_iとする。観測する前には、X_iの値は当然わからない。しかし母集団から抽出

されるので、どの値がどの程度出やすいかは母集団分布に従っている。すなわちX_iは母集団分布と同じ分布に従う確率変数である。（「確率変数」の項目を参照）。もし母集団のメンバーが有限であれば、X_iは母集団分布を少し修正した分布に従う。

標本からなる統計量を構成しよう。その統計量の値が、推測したい母数の周辺に多く分布しているならば、よい統計量と言える。統計量の分布を計算するためには、各標本が互いに独立であることが必要である。（その理由は「確率変数」の項目をみよ）。標本が互いに独立になるような抽出方法を**無作為抽出**（random sampling）という。

推測には推定（estimation）と検定（test）という２つの手法がある。推定は母数の値を直接いくらと表現する方法である。具体的に１つの値で表現する場合を点推定、幅をもって表現する場合を区間推定という。推定のための統計量を推定量という。また、標本を観測し、その観測値を推定量に代入して計算した結果を推定値という。

検定は、母数の値として否定したいと考えている仮説（帰無仮説）と主張したいと考えている仮説（対立仮説）を並べ、データから帰無仮説を強く棄却できるかどうかで判断する。判断するための統計量を検定量という。また、帰無仮説を棄却できると判断する検定量のとる値の領域を棄却域という。

度数分布（どすうぶんぷ）

得られたデータを整理する方法として、データのとりうる範囲をいくつかの階級に区分し、それぞれの度数をカウントして表にしたものを度数分布という。階級幅は等間隔にとることが多いが、度数の少ないところでは、階級幅を広げて表現することもある。階級数をいくつにするかについては決まった理論はない。目安としてスタージェス（Starjes）の公式が利用される。観測個数をnとした場合

$$\text{階級数} = 1+3.3\times\log_{10}n$$

である。n＝100であれば、階級数は７～８である。

例として、男子20人の体重（kg）は以下のようであった。

{57, 65, 56, 48, 54, 68, 72, 62, 56, 53, 59, 60, 64, 52, 58, 59, 63, 56, 55, 54}

表１のように３kgの階級幅でとったが、軽い階級と重い階級はそれぞれ広い幅でとった。

表1

階級	度数
46.5～52.5	2
52.5～55.5	4
55.5～58.5	5
58.5～61.5	3
61.5～64.5	3
64.5～73.5	3

この度数分布をグラフにしたものがヒストグラムである。ヒストグラムでは、度数の大きさを表現するのは棒の長さではなく、長方形の面積である。したがって、図１のようになる。図２のように、度数に比例して棒の長さを描くと誤ったグラフになる。

図1

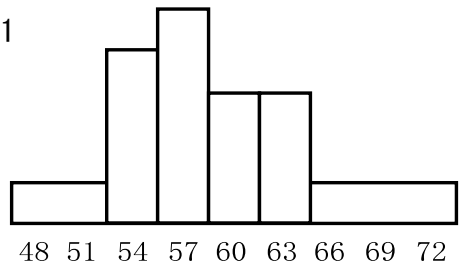
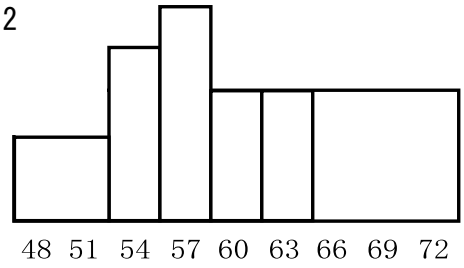


図2



位置の代表値（いちのだいひょうち）

データ全体の中心を表す尺度として最頻値（並み数、mode）、中央値（中位数、median）、平均値（mean）がある。例として、データが次の16個の観測値からなるとしよう。

｛ 4, 6, 3, 4, 5, 2, 5, 4, 6, 3, 5, 4, 3, 5, 3, 4 ｝

表2

値	度数
2	1
3	4
4	5
5	4
6	2
計	16

データを整理すると表２のようになる。**最頻値**は度数の一番多い値であるので、４である。**中央値**は大きさの順に並べて真ん中の値である。この場合は小さい方から数えて８番目と９番目の間である。どちらも値は４であるので、足して２で割っても４であり、それが中央値になる。平均値はすべての値を合計して得た６６を、観測数の１６で割った値である。したがって、4.125である。

N個のデータからなる一般式で平均値を表現してみよう。N個のデータを

$$\{x_1, x_2, \cdots, x_N\}$$

とする。**平均値**は

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

と定義される。

表3

値	度数
x_1	f_1
x_2	f_2
・	・
・	・
・	・
x_M	f_M
計	N

表３のような度数分布の一般的なケースでは、平均値は

$$\mu = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_M f_M}{f_1 + f_2 + \cdots + f_M}$$

$$= \frac{1}{N}\sum_{i=1}^{M} x_i f_i$$

$$(N = f_1 + f_2 + \cdots + f_M)$$

と定義される。母集団平均の記号はμを用いることが通例である。標本平均の場合には変数の上にバーをつけて、たとえば $\bar{x}$と表す。

階級幅のある例をみよう。

表4

階級	階級代表値	度数	累積度数
10ー20	15	3	3
20ー30	25	4	7
30ー40	35	8	15
40ー50	45	5	20
計		20	

階級３０～４０の度数が一番多い。したがって、最頻値はその階級中央の値３５とすればよい。中央値は全体の面積が度数の合計である２０となるヒストグラムを描き、面積を左右に二等分する座標である。累積度数をみると３０～４０の階級にあることがわかる。計算式は

$$30+10\times\frac{10-7}{8}=33.75$$

$$20+10\times\frac{5-3}{4}=25$$

である。平均値は階級代表値と度数を用いて計算すればよいので、32.5である。

最頻値は階級の分け方によっては複数存在したり、極端に変化することがあるので、さほど用いられない。中央値は外れ値（outlier）の影響を受けない長所はあるが、順序と中央値付近の値が分かればよいので、すべてのデータを用いているわけではないという短所がある。平均値はもっともよく使われる代表値であるが、外れ値に影響されやすい欠点がある。

散らばりの特性値（ちらばりのとくせいち）

広がりの尺度、ばらつき度などの表記がある。中央値と同様に大きさの順に並べていく考え方で作られた尺度として、**レンジ**（範囲）が挙げられる。これは最大値−最小値のことで、表２のケースでは６−２＝４である。レンジは外れ値があると影響を受けるので、通常は上下何％かを切り捨てた間の長さで測る。これを**パーセンタイル**という。とくに上下１０％ずつを切り捨てた間の長さを**十分位偏差**（じゅうぶんいへんさ）、上下２５％を切り捨てた尺度を**四分位偏差**（しぶんいへんさ）という。表２では第１四分位点は４番目と５番目の間の値であるから３であり、第３四分位点は１２番目と１３番目の間の値であるから５である。したがって、四分位偏差は5−３＝２である。

表４のケースでは、第１四分位点は小さい方からの面積が５である座標であるから、

である。第３四分位点は小さい方からの面積が１５である座標であるから、４０である。したがって、四分位偏差は４０−２５＝１５である。

平均と同様の考え方としては**分散**（variance）が挙げられる。その平方根を計算したものが**標準偏差**（standard deviation）である。考え方としては、分布の代表値（この場合は平均値）からの距離の平均を求めようとすることである。ただし、そのまま計算するとプラス・マイナス相殺されてゼロとなるので、代表値からの距離の２乗の平均を計算する。これを分散という。平均と同じ次元にするために分散の平方根を計算し、標準偏差を求める。

表２のケースでは、分散は 1.23 、標準偏差は 1.11 である。一般式では

$$\sigma^2=\frac{(x_1-\mu)^2+\cdots+(x_N-\mu)^2}{N}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

である。これは母集団分布における分散の定義式である。母集団分散の記号には σ^2 を用いる。もし n 個の標本から母集団分散を推定しようとするならば、

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

を用いる。（その理由については計量経営分析「誤差分散」の項目をみよ）。標本分散の記号にはs^2を用いる。この場合には、分数は1.32、標準偏差値は1.15に

なる。

度数がある表３のケースでは、母集団分散の定義は

$$\sigma^2 = \frac{(x_1-\mu)^2 f_1 + \cdots + (x_M-\mu)^2 f_M}{f_1 + \cdots + f_M}$$

$$= \frac{1}{N}\sum_{i=1}^{M}(x_i-\mu)^2 f_i$$

$$(N = f_1 + f_2 + \cdots + f_M)$$

である。標本から母集団分散を推定する場合には

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

$$(n = f_1 + f_2 + \cdots + f_m)$$

を用いる。

平均と標準偏差が与えられたとき、正規分布を基準にしてデータが従う分布を理解する。すなわち、平均から±３標準偏差離れた区間にほぼすべての観測値が入っている。±２標準偏差離れた区間には全体の約９５％が、±１標準偏差離れた区間には全体の２／３の観測値が入っている。（詳しくは「正規分布」の項目を参照）

５数要約（５すうようやく）

例として、以下に２１人の年間所得（単位：百万円）がある。

{7, 7, 5, 5, 9, 5, 3, 3, 3, 3, 11, 5, 3, 9, 7, 5, 7, 1, 5, 21, 5}

このデータを大きさの順に並べて整理しよう。

{1, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 7, 7, 7, 7, 9, 9, 11, 21}

分布の概形を知るには図３のようにヒストグラムを描くのが有効である。

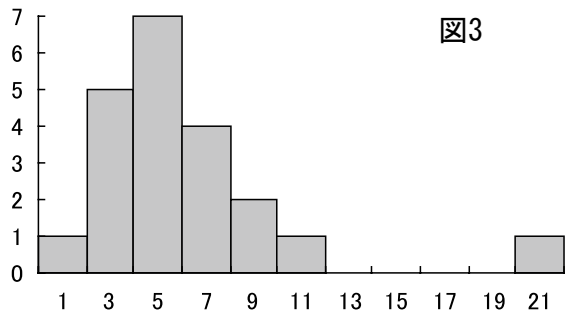


図3

所得分布は右に裾野が長い。５百万円の所得の人が一番多く、それが最頻値である。中央値は５百万円、平均値は６１４万円、n-1でわった標準偏差は４１７万円である。

所得分布のように左右対称ではなく歪んでいる分布の場合には、正規分布を基準にした平均、標準偏差による把握には無理が生じる。そこで直感的にも理解しやすく、概形をとらえる数値として５数要約がある。それは最小値、第１四分位点、中央値、第３四分位点、最大値の５個の値である。この例では、（1, 3, 5, 7, 21）である。これをグラフにしたのが図４のような**箱ひげ図**（ボックスプロット）である。

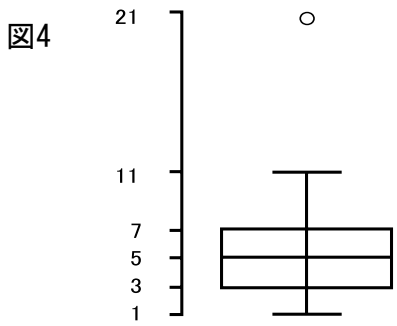


図4

長方形の箱の上辺が第３四分位点、下辺が第１四分位点、中の線が中央値（第２四分位点）である。四分位偏差は7－３＝４である。ひげの描き方はつぎの通りである。

上ひげは、第３四分位点＋1.5×四分位

偏差 より小さい実在の値

下ひげは、第１四分位点−1.5×四分位偏差 より大きい実在の値

である。このケースでは、上ひげは１３（＝7+1.5×4）より小さい実在の値であるので１１になる。下ひげは−１（＝3−1.5×4）より大きい実在の値であるので１になる。上ひげ、下ひげからはずれた値は小さい丸で示してある。

相関係数（そうかんけいすう）

１変数の場合、データを位置の代表値と散らばりの特性値に縮約して把握しようとするのは、正規分布が基準になっていると考えれば理解しやすい。多く見られる分布は、中心極限定理により正規分布に近い形をしているからである。正規分布の場合には、平均と分散で完全に記述される。（「正規分布」「中心極限定理」の項目を参照）

表5

	成績			
	良い	普通	悪い	計
A	235	884	170	1289
B	150	472	95	717
O	193	610	139	942
AB	65	224	39	328
計	643	2190	443	3276

表6

	成績			
	良い	普通	悪い	計
A	0.07	0.27	0.05	0.39
B	0.05	0.14	0.03	0.22
O	0.06	0.19	0.04	0.29
AB	0.02	0.07	0.01	0.10
計	0.20	0.67	0.13	1.00

表7

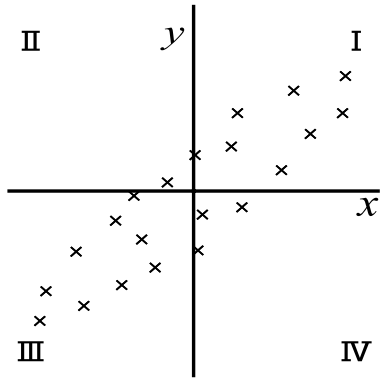
	成績			
	良い	普通	悪い	計
A B O AB	同時確率			周辺確率
計	周辺確率			1.00

２変数を同時に扱う場合を考えてみよう。表５は血液型と成績を同時に扱った度数分布である。表６は、全体を１としたときの相対度数である。全体の度数が大きい場合、相対度数を確率と考えてよい。（「確率」の項目を参照）。したがって、表７に示すように、これを**同時確率**と呼ぼう。それに対し、血液型や成績のみに注目した場合を**周辺確率**という。

２変数を同時に扱う場合には、それぞれの変数の平均・分散のほかに、両者の関連の程度を示す尺度が必要である。この場合にも、２変数の同時正規分布を基準に考えると理解しやすい。２変数同時正規分布の場合には、各変数の平均・分散のほかに、２変数の関連を示す相関係数を加えて、５個の数値で分布を完全に記述することができる。

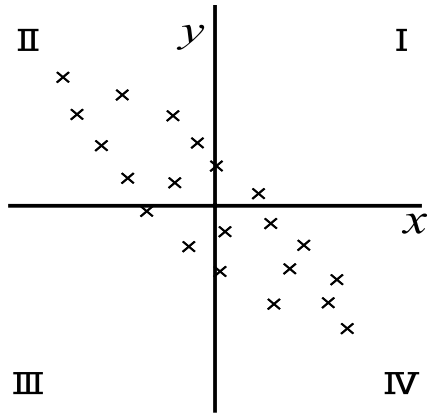
相関係数とは、２変数が直線関係にどの程度近いかを示す尺度である。変数X、YについてN個の観測値が得られたとしよう。Xの平均をμ_x、Y の平均をμ_yとする。座標 (μ_x,μ_y) を中心に平面を４分割して、それぞれⅠ～Ⅳ象限とよぼう。

図a

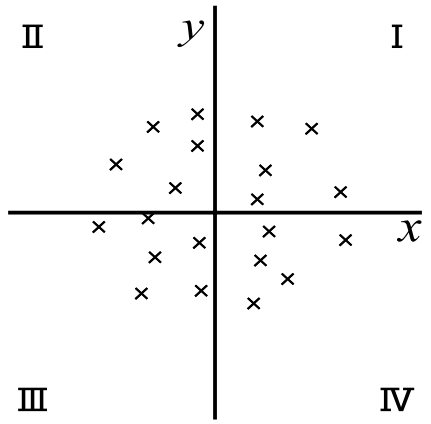


図ａでは、右上がりの直線に近く、ⅠとⅢ象限に多くの点がある。図ｂは右下がりの直線に近くⅡとⅣ象限に多くの点が、図ｃは直線関係からほど遠く、Ⅰ～Ⅳ象限に万遍なく点がある。この特徴か

図b



図c



らⅠとⅢに共通な性質と、ⅡとⅣに共通な性質を取り出してみる。i 番目の観測値を (x_i, y_i) としよう。それがどの象限にあるかによって、座標の中心 (μ_x, μ_y) からの偏差の符号と、その積の符号は以下のようになる。

	$x_i-\mu_x$	$y_i-\mu_y$	$(x_i-\mu_x)(y_i-\mu_y)$
Ⅰ象限	＋	＋	＋
Ⅱ象限	－	＋	－
Ⅲ象限	－	－	＋
Ⅳ象限	＋	－	－

したがって、偏差の積 $(x_i-\mu_x)(y_i-\mu_y)$ がⅠ・Ⅲ象限ではプラス、Ⅱ・Ⅳ象限ではマイナスになる。偏差の積を足した値がプラスに大きければ右上がりの直線に、マイナスに大きければ右下がりの直線に近く、０に近ければ直線からほど遠い。観測数で割って平均をとると

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$$

となる。これを**共分散**（covariance）という。

共分散は単位が変わると、変化する。たとえば、X は身長でｃｍの単位、Y は体重でｋｇの単位であるとする。身長をメートルで表現すれば、共分散は百分の１になる。そこで単位が変わっても変わらない尺度として相関係数（correlation coefficient）を考える。すなわち、平均からの偏差を標準偏差で割って基準化すればよい。相関係数は

$$\rho_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)$$

$$= \frac{Cov(X,Y)}{\sigma_x\sigma_y}$$

である。

相関係数の値は−１から１までの範囲にある。１であれば、点 (x_i, y_i) がすべて右上がりの直線上にある。−１であれば、すべて右下がりの直線上にある。０であれば、直線関係から遠い。

確率（かくりつ）

コインを２枚投げる実験を考えよう。実験の結果を**事象**（event）という。起こりうる事象は（表，表）、（表，裏）、（裏，表）、（裏，裏）の４つが基本である。それぞれをｅ１、ｅ２、ｅ３、ｅ４と名づけよう。そのさい基本事象の組合

せも事象と考えることにする。たとえば（$e_1 \cup e_2$）である。∪は「または：ｏｒ」という意味であり、e_1とe_2の和事象を表す。（$e_1 \cup e_2$）と（$e_2 \cup e_3$）の共通事象はe_2であるが、これは積事象とよばれ、∩を用いて表す。∩は「かつ：ａｎｄ」という意味である。

Ｓ＝（$e_1 \cup e_2 \cup e_3 \cup e_4$）は全事象とよばれる。全事象から（$e_1 \cup e_2$）を除いた残りの事象は（$e_3 \cup e_4$）であるが、これは（$e_1 \cup e_2$）の補事象（余事象）とよばれる。全事象の補事象は空事象とよばれ、φで表す。この例では、空事象も含めて、事象は全部で１６ある。

確率（probability）とは、事象の出やすさを０から１までの数字で表現したものである。ある事象Ａの確率はＰ（Ａ）と表す。そのさい次の２つの関係を満たすことが必要である。

(1)ある事象ＡとＢの積事象が空事象ならば（Ａ∩Ｂ＝φ、ＡとＢは**互いに排反**とよばれる）、

Ｐ（Ａ∪Ｂ）＝Ｐ（Ａ）＋Ｐ（Ｂ）

がなりたつ。

(2)ある事象Ａの補事象をA^Cとする。

Ｐ（Ａ）＋Ｐ（A^C）＝１。

(1)と(2)の規則が成り立てば確率と言ってよいが、ある現象を説明する適切な確率表現であるかどうかは分からない。コインを２枚投げるケースでは、たとえば１００回実験を繰り返し、（表，表）が２６回出現したならば、その相対度数である0.26を確率と考えたい。実験は無限回行うことにし、相対度数の極限値を確率とする。これを**経験的確率**という。しかし実際には、無限回実験をすることはできない。歪んでいないコインであれば、４つの基本事象の出やすさは区別ができない。したがって、みな等しい確率とみなすことにする。これを**等可能性の原理**という。

独立（どくりつ）

例として、表５にある血液型と成績の度数分布を取り上げよう。これは母集団メンバー全員を調べた結果であるとする。血液型Ａで成績良である同時確率は

$$P(\mathrm{A},\ \text{良}) = \frac{235}{3276} = 0.07$$

である。また、血液型Ａの人の確率は

$$P(\mathrm{A}) = \frac{1289}{3276} = 0.39$$

である。ここで、血液型Ａだけに限って、そのうちで成績良である確率を求めよう。

$$P(\text{良} \mid \mathrm{A}) = \frac{235}{1289}$$

$$= \frac{235/3276}{1289/3276} = \frac{P(\mathrm{A},\ \text{良})}{P(\mathrm{A})}$$

である。一般に、事象Ａに限って、そのうちで事象Ｂである確率を**条件付き確率**といい、$P(B \mid A)$と表現する。

$$P(B \mid A) = \frac{P(A,B)}{P(A)}$$

となる。同時確率を左辺にして変形すれば

$$P(A,B) = P(A)P(B \mid A)$$

である。

２枚のコインがある。コインＡを投げると表のでる確率は１／２、コインＢも表のでる確率は１／２であるとする。２枚のコインを同時に投げるとき、２枚とも表のでる確率はいくらか。この答は１／４とは限らない。２枚のコインを表が同じ側になるようにして棒に結びつけて投げれば、２枚とも表のでる確率は１／２であろう。すなわち、コインＡが表のでるとき、コインＢが表のでる条件付き確率が分からないと答えられない。一般的に、条件付き確率は分からないのが通常である。コインＡの結果によらず、コインＢが表のでる確率が１／２であれば、答は１／４になる。すなわち、

$$P(B \mid A) = P(B)$$

が成り立てばよい。このような状態を独立という。したがって事象Ａ、Ｂが独立であれば、同時確率は周辺確率の積になる。すなわち $P(A,B) = P(A)P(B)$ である。

表8

	好き	嫌い	計
男			110
女			90
計	120	80	200

表８にある性別による犬の好き嫌いを考えてみよう。性別によって犬の好き嫌いは影響されないとしよう。その場合、空欄の数字はいくらになるか。答は表９しかありえない。他のどのような数字を入れても、男で犬の好き嫌いの分布と、女で犬の好き嫌いの分布は異なってくる。表９の場合だけが、男女で犬の好き嫌いの分布が同じになる。

表9

	好き	嫌い	計
男	66	44	110
女	54	36	90
計	120	80	200

性別によって犬の好き嫌いが影響されないなら、

Ｐ（好き｜男）＝Ｐ（好き｜女）＝Ｐ（好き）、

Ｐ（嫌い｜男）＝Ｐ（嫌い｜女）＝Ｐ（嫌い）

が成り立つ。犬の好き嫌いの変数からみても、犬の好きな人の男女比率と、嫌いな人の男女比率は同じであることが言える。すなわち、独立とは条件の如何によって分布が左右されないことである。

確率変数（かくりつへんすう）

例として、３枚のコインを投げる実験を考えよう。

表10

記号	基本事象	確率
e_1	(表, 表, 表)	1/8
e_2	(表, 表, 裏)	1/8
e_3	(表, 裏, 表)	1/8
e_4	(表, 裏, 裏)	1/8
e_5	(裏, 表, 表)	1/8
e_6	(裏, 表, 裏)	1/8
e_7	(裏, 裏, 表)	1/8
e_8	(裏, 裏, 裏)	1/8

表１０は基本事象とその確率である。確率を棒の長さで表し、これをグラフに描こう。

図5

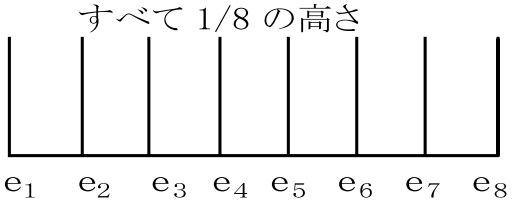


図５のように描きたいところである。しかし、絵グラフならともかく、横軸が実数軸となるグラフを描くとすれば、図５は意味をなさない。まずｅ1、ｅ2、・・・、ｅ8と並ぶ必然性がないし、等間隔である必要もない。図５では平均や分散という概念も意味を持たない。すなわち、事象そのものの確率を直接扱うのは、グラフが描けない等の不便さがあり、操作性が悪い。そこで事象を実数に対応させれば、横軸が実数軸となるグラフが描けるし、平均・分散という概念も意味を持つ。この対応を確率変数（random variable）とよぶ。

表１０の例ではいくつかの意味ある対応が考えられるが、ここでは表のでる数という対応をXとして、その確率を導こう。

表11

確率変数 X	
値	確率
0	1/8
1	3/8
2	3/8
3	1/8

図6

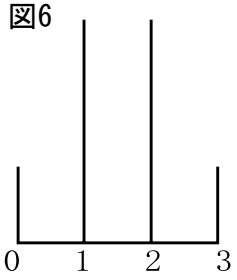


表１１は確率変数 Xの分布で、図６はそのグラフである。縦の棒の長さは確率をあらわしている。この場合、平均は1.5 、分散は 0.75 である。

つぎに確率変数の和の分布を考えよう。例として、X_1 、X_2の同時確率分布が表１２のように与えられているとしよう。

表12

		X_2		
		1	2	計
X_1	1	0.2	0.1	0.3
	2	0.1	0.6	0.7
	計	0.3	0.7	1

X_1+X_2 の分布は表１３のように導かれる。もし表１４のように X_1、X_2の周辺分布しか分からなかったとしたら、一般的には同時分布は計算できない。

表13

X_1+X_2	
値	確率
2	0.2
3	0.2
4	0.6

表14

		X_2		
		1	2	計
X_1	1			0.3
	2			0.7
	計	0.3	0.7	1

したがって和の分布も求められない。同時分布が計算できるためには、X_1 、X_2が独立であることが必要である。独立であれば、和の分布は表１５のように求められる。

表15

X_1+X_2	
値	確率
2	0.09
3	0.42
4	0.49

統計的推測においては、標本の関数である統計量を用いて母数を推測する。単純化のために母集団には無限に多くのメンバーがいるとし、そこからn個の標本 $(X_1,X_2\cdots,X_n)$ を抽出しよう。各標本は観測後には定まった値になるが、観測前には母集団分布と同じ分布に従う確率変数である。それゆえ、統計量の分布は各標本の同時確率$P(X_1,X_2,\cdots,X_n)$ が計算できないと求められない。一般的に、各標本の確率分布 $P(X_i)$は未知パラメーターを含んではいるが想定できる。したがって、同時確率は各標本が独立であることを前提にすれば、周辺確率の積として求められる。

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2)\cdots P(X_n)$$

以上の記述からわかるように、標本抽出では各標本が互いに独立であるような無作為抽出（ランダム・サンプリング）が原則である。**無作為抽出**であるためには、母集団の各メンバーが標本として抽出される確率が分かっていることが必要である。各メンバーが標本として抽出される確率がみな等しい場合が、無作為抽出の単純なケースである。

離散と連続（りさんとれんぞく）

確率変数のとる値の数が有限または加算無限の場合を**離散分布**、連続無限である場合を**連続分布**という。ここで、加算無限とは自然数と同じだけの数をいう。それ以上多くある場合を連続という。加算無限と連続無限の間にくるような数は存在しない。

表１１のケースでは０、１、２、３という４個の値をとる。したがって、これは離散分布である。グラフに描く場合、確率の大きさは図６のように棒の長さで表現する。それに対し、表１のケースでは46.5から73.5までのすべての実数を対象にしているので連続分布である。連続分布の場合、たとえば丁度５０ｋｇという確率は０と考えられる。というのは、少しでも違えば別の値ということになるので、丁度その値という一点の確率は０とする。グラフで描くとき、各値の確率を棒の長さで表現することはできない。確率は図１のように区間に対応する面積で表現する。任意の区間の面積でよいので、図７の太線で描いたような形がわかればよい。形を表すそのような線を**確率密度関数**という。

図7

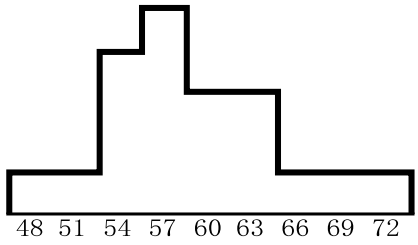


図８は、確率変数Xの確率密度関数であるとする。横軸がXのとりうる値、曲線 $f(x)$が確率密度関数である。区間(a,b)の確率は斜線を引いた面積であるが、積分を用いて$\int_a^b f(x)dx$で表される。

図8

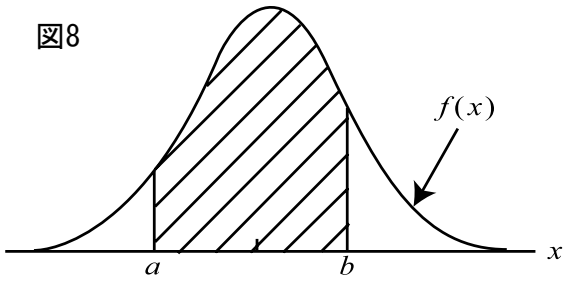
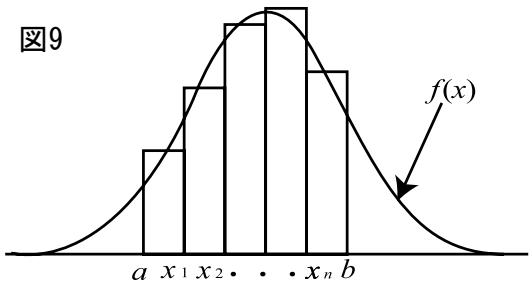


図9



積分の考え方は、区間(a,b)を図９のようにn等分する。
各区間の幅は $\Delta x = \dfrac{b-a}{n}$ である。各区間の中間点を$x_1, \cdots, x_n$とする。第i 区間の確率は、

矩形の面積＝縦 $f(x_i)$ ×横 Δx

で近似される。したがって、各区間の矩形をすべて合計した面積 $\sum_{i=1}^{n} f(x_i)\Delta x$ は区間(a,b)の確率に近似される。
ここで $n \to \infty$、$\Delta x \to 0$ にすると、矩形

面積の合計は区間 (a,b) の確率に収束する。すなわち、

$$P(a,b)=\lim_{n\to\infty}\sum_{i=1}^{n}f(x_i)\Delta x$$

であり、それを $\int_a^b f(x)dx$ と書くことにする。（積分の記述は考え方を示したものであり、数学的に厳密ではない）

期待値（きたいち）

離散分布の場合の期待値（expectation）を定義しよう。表１６のような確率変数 X の場合、その期待値は

表16

確率変数 X	
値	確率
x_1	p_1
x_2	p_2
・	・
・	・
・	・
x_n	p_n
	1

$$E(X)=x_1p_1+\cdots+x_np_n=\sum_{i=1}^{n}x_ip_i$$

であり、それは平均値 μ のことである。X の関数 $g(X)$ の期待値は

$$E\{g(X)\}=g(x_1)p_1+\cdots+g(x_n)p_n=\sum_{i=1}^{n}g(x_i)p_i$$

と定義される。分散は $(X-\mu)^2$ の期待値である。したがって

$$\sigma^2=E\{(X-\mu)^2\}=(x_1-\mu)^2p_1+\cdots+(x_n-\mu)^2p_n=\sum_{i=1}^{n}(x_i-\mu)^2p_i$$

分散の記号としては σ^2 、$V(X)$ 、$Var(X)$ などを用いる。

連続分布の場合には、確率密度関数を $f(x)$ とすると

$$\mu=E(X)=\int_{-\infty}^{\infty}xf(x)dx$$

と表現される。図９を用いて考え方を説明しよう。期待値の場合にはすべての値について考慮するので、区間 (a,b) に相当するのは区間 $(-\infty,\infty)$ になる。第 i 区間の確率は $p_i=f(x_i)\Delta x$ である。
期待値は、

$$\sum_{i=1}^{n}x_ip_i=\sum_{i=1}^{n}x_if(x_i)\Delta x$$

において $n\to\infty$ 、$\Delta x\to 0$ にした極限である。すなわち（数学的に厳密な表現ではないが）

$$\mu=E(X)=\lim_{n\to\infty}\sum_{i=1}^{n}x_ip_i=\lim_{n\to\infty}\sum_{i=1}^{n}x_if(x_i)\Delta x=\int_{-\infty}^{\infty}xf(x)dx$$

分散は

$$\sigma^2=V(X)=E\{(X-\mu)^2\}=\int_{-\infty}^{\infty}(x-\mu)^2f(x)dx$$

つぎに、確率変数のスカラー倍について、結果のみを記述しよう。

$$E(cX)=cE(X)$$
$$V(cX)=c^2V(X)$$

２つの確率変数 X、Y の和の平均と分数は

$$E(X+Y)=E(X)+E(Y)$$
$$V(X+Y)=V(X)+V(Y)+2Cov(X,Y)$$

もし X と Y が独立であれば

$$V(X+Y)=V(X)+V(Y)$$

中心極限定理（ちゅうしんきょくげんていり）

互いに独立な確率変数 $(X_1,X_2,\cdots,X_n)$ がある。その和を

$$Y=X_1+X_2+\cdots+X_n$$

としよう。各 X_i の分散が和の確率変数 Y の分散に比べていずれも十分小さいとする。適当な条件の下で、n が大きくなるに従って確率変数 Y は正規分布に近似するようになる。これを中心極限定理という。n の数はおおよそ３０以上であれば正規分布とみなしてよい。

例として３割打者のケースをみよう。i 回目の打席の結果を X_i とし、ヒットなら $X_i=1$ アウトなら $X_i=0$ とする。したがって

$$P(X_i=0)=0.7$$
$$P(X_i=1)=0.3$$

である。

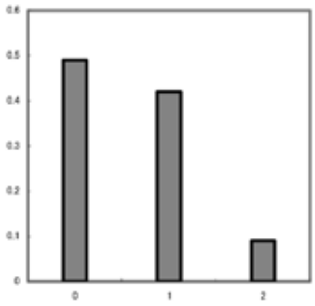
２回打席にたつとし、各回独立であるとする。同時確率 (X_1,X_2) およびヒット回数 X_1+X_2 の分布は以下の図表のようになる。

(X_1,X_2) の確率

		第2打席 X_2		
		0アウト	1 ヒット	計
第1打席	0 アウト	0.49	0.21	0.7
	1 ヒット	0.21	0.09	0.3
	計	0.7	0.3	1

X_1+X_2 の確率

値	確率
0	0.49
1	0.42
2	0.09



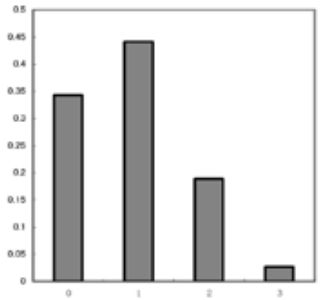
３回打席にたつ場合の同時確率とヒット回数 $X_1+X_2+X_3$ の分布は

(X_1+X_2,X_3) の確率

		X_3		
		0アウト	1 ヒット	計
X_1+X_2	0	0.343	0.147	0.49
	1	0.294	0.126	0.42
	2	0.063	0.027	0.09
	計	0.7	0.3	1

$X_1+X_2+X_3$の確率

値	確率
0	0.343
1	0.441
2	0.189
3	0.027



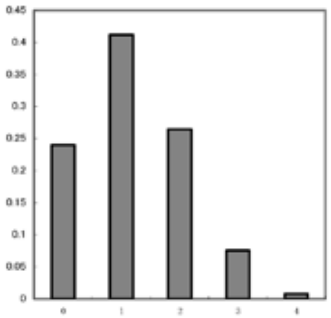
４回打席にたつ場合の同時確率とヒット回数 $X_1+X_2+X_3+X_4$ の分布は

$(X_1+X_2+X_3,X_4)$ の確率

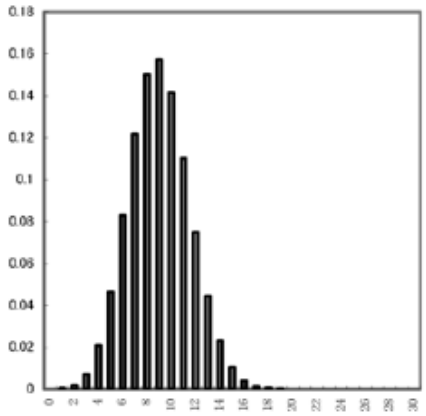
		X_4		
		0アウト	1 ヒット	計
$X_1+X_2+X_3$	0	0.2401	0.1029	0.343
	1	0.3087	0.1323	0.441
	2	0.1323	0.0567	0.189
	3	0.0189	0.0081	0.027
	計	0.7	0.3	1

$X_1+X_2+X_3+X_4$の確率

値	確率
0	0.2401
1	0.4116
2	0.2646
3	0.0756
4	0.0081



３０回打席にたつ場合のヒット回数 $X_1+X_2+\cdots+X_{30}$ の確率分布は、以下の図のように正規分布といわれる分布に近い。ここで挙げた例は、２項分布の正規近似といわれるものである。



一般に各 X_i がそれぞれ異なる分布の場合であっても、互いに独立であり、適度に分散が小さければ、和 $X_1+X_2+\cdots+X_n$ の分布は正規分布に近似してくる。

標本平均の分布（ひょうほんへいきんのぶんぷ）

母集団から n 個の標本を抽出しよう。母集団分布は平均 μ、分散 σ^2 であり、母集団には無限に多くのメンバーがいるとする。また、各標本は互いに独立であるように抽出すること（無作為抽出）が原則である。そのとき、標本平均 $\overline{X}$ の分布は

$$E(\overline{X})=E\left(\frac{X_1+\cdots+X_n}{n}\right)$$
$$=\frac{1}{n}\{E(X_1)+\cdots+E(X_n)\}=\mu$$

$$V(\overline{X})=V\left(\frac{X_1+\cdots+X_n}{n}\right)$$
$$=\frac{1}{n^2}\{V(X_1)+\cdots+V(X_n)\}=\frac{\sigma^2}{n}$$

である。分散の計算では、標本が互いに独立であることを用いている。

項目「中心極限定理」で述べた３割打者の例をとりあげよう。下表は、標本数 $n=1,2,3,4$ のケースにおける標本平均の分布である。

母集団分布は平均 $\mu=0.3$、分散 $\sigma^2=0.21$ である。標本平均の平均は μ、分散は $\frac{\sigma^2}{n}$ になっていることが確かめられる。

標本数 n が大きければ中心極限定理が成立するので、標本平均は正規分布をす

X の分布

値	確率
0	0.7
1	0.3

平均　0.3
分散　0.21

$\frac{X_1+X_2}{2}$ の分布

値	確率
0	0.49
0.5	0.42
1	0.09

平均　0.3
分散　0.105

$\frac{X_1+X_2+X_3}{3}$ の分布

値	確率
0	0.343
1/3	0.441
2/3	0.189
1	0.027

平均　0.3
分散　0.7

$\frac{X_1+X_2+X_3+X_4}{4}$ の分布

値	確率
0	0.2401
0.25	0.4116
0.5	0.2646
0.75	0.0756
1	0.0081

平均　0.3
分散　0.0525

ると考えてよい。すなわち

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

標本数nが無限に大きくなると標本平均の分散は０に収束する。したがって、標本平均は母平均μの値だけに分布するようになる。これを**大数法則**という。母集団分布をXとし、記号で示そう。母平均$\mu = E(X)$であるから、

$n \to \infty$のとき$\bar{X} \to E(X)$である。

２項分布（にこうぶんぷ）

２項分布は、離散確率分布のなかで最も基本的であって、以下の状況をモデル化した分布である。

次の２つの条件をみたす繰り返し実験を**反復試行（ベルヌーイ試行）**という：

・どの回も結果は2つ（成功Sと失敗 F ）で、その確率は

$$P(S) = p,\ P(F) = 1 - p = q$$

・どの回も成功確率pは一定不変で、それぞれの回は確率的に独立。

n回の反復試行における成功Sの回数Xは**２項分布**に従う離散確率変数で、その確率は

$$P(X = x) = f(x) = {}_nC_x p^x q^{n-x},$$
$$x = 0,1, \cdots, n$$

で与えられる。これを$X \frown B(n, p)$ とも表記する。この確率の式は、 n 回のうちx回成功（し､$n-x$回失敗）する確率は、成功確率不変かつ独立という条件から$p^x q^{n-x}$、そして n 回のうちx回成功という場合は、$1,2,\cdots,n$ のなかから成功回の番号を x 個とり出す組合せ${}_nC_x$だけある、ということから導かれる。２項分布$B(n, p)$の平均と分散は

$$E(X) = np, \quad V(X) = npq$$

である。平均$n \times p$は、例えばサイコロを60回振ったとき、1の目が出る回数は‘平均的に60×(1/6)＝10’とする直感と一致する。分散npqについては、反復数nが多くなれば X の値の範囲が比例的に広がるのだから、分散もnに比例すると考えればよい。反復数nを一定とすると、$p = q = 0.5$ のとき分散は最大になる。

反復数$n = 1$の２項分布 $B(1, p)$を、特に**ベルヌーイ分布**という。

$$P(X = x) = f(x) = p^x q^{1-x},\ \ x = 0,1$$
$$E(X) = 0 \times q + 1 \times p = p,$$
$$V(X) = (0-p)^2 \times q + (1-p)^2 \times p = pq$$

である。

n個の確率変数 $X_1, X_2, \cdots, X_n$ が独立で同じベルヌーイ分布$B(1, p)$に従う（**独立同分布**、independent and identically distributed；*i.i.d.*）とき、その和は反復試行n回の成功回数となるので２項分布$B(n, p)$に従う：$\sum_{i=1}^{n} X_i \ \sim B(n, p)$

これを利用すると、期待値の演算から

$$E\left(\sum X_i\right) = \sum E(X_i) = np,$$
$$V\left(\sum X_i\right) = \sum V(X_i) = npq$$

となって、上記の２項分布 $B(n, p)$の平均と分散が得られる。

正規分布（せいきぶんぷ）

中心極限定理からも予想されるように、正規分布は確率にかかわるすべての分野で中心的役割を果たす分布である。

連続確率変数Xの密度関数が

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

$$\left(=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right),$$

$$-\infty<x<\infty$$

ただし、$\mu,\ \sigma(>0)$はパラメータ（分布を決める定数）

で与えられるとき、Xは**正規分布**（または**ガウス分布**）に従うといい、

$X\backsim N(\mu,\sigma^2)$と表記する。ここで記号$\mu,\ \sigma\ (\sigma^2)$を用いているのは

$$E(X)=\mu,\ V(X)=\sigma^2$$

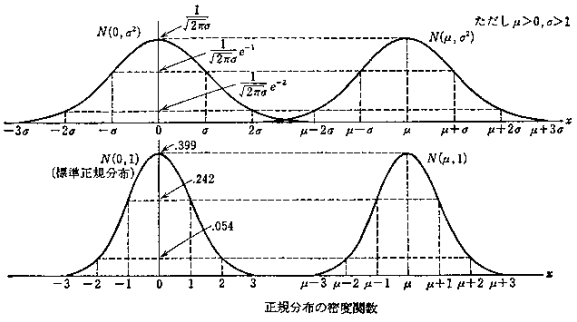
つまり、この正規分布の平均がμ、分散がσ^2（σは標準偏差）になる、という事実に基づく。この密度関数の曲線は次のような特徴をもつ。

①左右対称なベル型（bell shape）で、中心$x=\mu$で最大値 $1/\sqrt{2\pi\sigma^2}$ $\cong 0.399/\sigma$ をとる。

②$x\to\pm\infty$ のとき $f(x)\to 0$ であるが、$f(x)$は非常に早く０に近づく（スソが短い）。

③曲線の傾きが最も急（変曲点）なのは$x=\mu\pm\sigma$（平均±標準偏差）のとき。

④平均μの値が変化すると曲線の形は変わらずに左右に移動し、分散σ^2の値が大きくなると曲線は平べったく広がる。



平均μを中心とした正規分布の主要な区間確率を列挙すると、

$P(\mu-\sigma\le X\le\mu+\sigma)\quad=68.3\%$

（区間外、ほぼ1/3）

$90\%=P(\mu-1.645\sigma\le X\le\mu+1.645\sigma)$

$95\%=P(\mu-1.96\sigma\le X\le\mu+1.96\sigma)$

$P(\mu-2\sigma\le X\le\mu+2\sigma)=95.5\%$

（区間外、ほぼ1/20）

$99\%=P(\mu-2.576\sigma\le X\le\mu+2.576\sigma)$

$P(\mu-3\sigma\le X\le\mu+3\sigma)=99.7\%$

（区間外、ほぼ3/1000）

である。

平均$\mu=0$、標準偏差$\sigma=1$の正規分布$N(0,1)$を、特に**標準正規分布**という。

密度関数：$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2}\right\},$ $-\infty<z<\infty$

分布関数：$\Phi(u)=\int_{-\infty}^{u}\phi(z)dz$

であるが、分布関数は初等関数で表せないので数表が作成されている。例えば、**正規分布の上側確率**は

$Q(u)=1-\Phi(u)=\int_{u}^{\infty}\phi(z)dz$ を$u\ge 0.00$に対して示した表である。

正規分布$N(\mu,\sigma^2)$に従う確率変数Xは、標準化$Z=\frac{X-\mu}{\sigma}$すると標準正規分布$N(0,1)$に従うから、任意の正規分布の区間確率は正規分布表を使って簡単に求められる。

点推定（てんすいてい）

母集団比率p（例えば、ある時点における日本の有権者全体の内閣支持率）への推論を素材にして点推定の考え方を説明しよう。N＝約１億人の有権者（母集団）からn＝２千人程度を**無作為抽出**

（random sampling）するような場合は、抽出率が小さいので$N=\infty$と見做して差し支えない。そうすると標本抽出の際、内閣支持者($X=1$)に当る確率は p 、そうでない有権者（$X=0$） に当る確率は $1-p=q$ 、そして各有権者は独立に抽出されることになる。この状況は、母集団比率（母比率）p を未知パラメータ（未知母数）とするベルヌーイ分布からの無作為抽出であり、

母集団分布：ベルヌーイ分 $B(1,p)$ 、

（母平均のpは未知パラメータ）

標本変量：$X_1,X_2,\cdots,X_n$は独立同　ベルヌーイ分布 $B(1,p)$に従うと定式化される。

標本変量$(X_1,X_2,\cdots,X_n)$の和$\sum X_i$は２項分布$B(n,p)$に従い、その平均量（標本比率量）$\hat{p}=\dfrac{\sum X_i}{n}$ は、とる値が０から１に圧縮された２項分布

$$P(\hat{p}=r)={}_nC_{nr}p^{nr}q^{n(1-r)},\ r=\frac{0}{n},\frac{1}{n},\cdots,\frac{n}{n}$$

$$E(\hat{p})=p,\ \ V(\hat{p})=\frac{pq}{n}$$

に従う。これより、母比率pの自然で素朴な推定方式（**推定量**）として、標本比率量 $\hat{p}=\dfrac{\sum X_i}{n}$　は『「平均的には」母比率pを偏りなく推定する($E(\hat{p})=p$）』という意味で望ましい方式であり、更に『標本サイズnが大きくなると、$\hat{p}$ の精度は高くなる（分散$V(\hat{p})=(pq)/n$ が小さくなる）』ことがわかる。

この事実に基づいて標本比率量の実現値 $\hat{p}=\dfrac{\sum x_i}{n}$ を未知な母比率pの推定値としてよいと考える。

区間推定（くかんすいてい）

母比率pについて、n が大きい（n＞約30、**大標本**）ときを例にして説明する。

$$n\hat{p}=\sum x_i>5,$$

$$n(1-\hat{p})=n\hat{q}=n-\sum x_i>5$$

なら、中心極限定理により、標本比率量 $\hat{p}=\dfrac{\sum X_i}{n}$ の標準化変量

$Z=\dfrac{\hat{p}-p}{\sqrt{\dfrac{pq}{n}}}$は近似的に標準正規分布

$N(0,1)$に従う

と見做すことができる。そうすると、例えば平均pを中心とする区間確率について

$$P\left(\left|\frac{\hat{p}-p}{\sqrt{pq/n}}\right|\le 1.96\right)$$

$$=P\left(\hat{p}-1.96\sqrt{\frac{pq}{n}}\le p\le\hat{p}+1.96\sqrt{\frac{pq}{n}}\right)$$

$$\cong 0.95$$

が成り立つ。

これに基づき、$\hat{p}=\dfrac{\sum X_i}{n}$の実現値

$\hat{p}=\dfrac{\sum x_i}{n}$ を用いて作られる区間

$$\hat{p}-1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\le p\le\hat{p}+1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

（ただし、未知な pq は、推定値$\hat{p}\hat{q}$ に置き換えている。注を参照。）

を**信頼係数**95％の**信頼区間**といい、区間推定の方式とする。このとき、両端の $\hat{p}\pm1.96\sqrt{\hat{p}\hat{q}/n}$　を**信頼限界**という。なお、信頼係数を99%にしたければ、1.96を2.576にすればよい。一般に信頼係数が $100\times(1-\alpha)$ ％ならば、標準正規分布の上

側 $100\times(\alpha/2)$ %点とする。

点推定値 $\hat{p}$ を中心とする

$$\hat{p}\pm 1.96\sqrt{\hat{p}\hat{q}/n}$$

の区間ができるわけであるが、その意味は次のように解釈される。母比率 p は未知の定数であって、推定値から作られる区間 $\hat{p}\pm 1.96\sqrt{\hat{p}\hat{q}/n}$ 内にある（正しい）か、ない（誤り）かのどちらかである。しかしながら、サイズ n の無作為抽出を多数回行ってその度に区間推定を行えば、そのうちの95%は正しいこと述べていることになる。その意味でサイズ n の無作為抽出１回の結果で作られる区間を「確率区間」と言わず、「信頼区間」と言う。また、95%に「確率」という言葉を使わず、「信頼」係数と呼ぶ。

（注） p に関する２次不等式

$$\left(\frac{\hat{p}-p}{\sqrt{pq/n}}\right)^2 \le z_k^2 \quad （但し、 z_k=1.96）$$

を解くと

$$\frac{1}{1+(z_k^2/n)}\left[\left(\hat{p}+\frac{z_k^2}{2n}\right)-z_k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_k^2}{4n^2}}\right]$$

$$\le p \le \frac{1}{1+(z_k^2/n)}\left[\left(\hat{p}+\frac{z_k^2}{2n}\right)+z_k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_k^2}{4n^2}}\right]$$

が得られるから、n が大で、$\frac{z_k^2}{n}\cong 0$ と見做してよければ、$\frac{z_k^2}{n^2}\cong 0$ でもあり、近似としての信頼区間

$$\hat{p}-z_k\sqrt{\frac{\hat{p}\hat{q}}{n}} \le p \le \hat{p}+z_k\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

が導かれる。

最尤法（さいゆうほう）

標本データ $(x_1,x_2,\cdots,x_n)$ が得られたものとして、標本変量 $(X_1,X_2,\cdots,X_n)$ の同時分布を未知パラメータの関数とみなすとき、これを一般に**尤度関数**（likelihood function）という。

いま、

母集団分布：ベルヌーイ分布 $B(1,p)$、

（母平均の p は未知パラメータ）

標本変量： $X_1,X_2,\cdots,X_n$ は独立同ベルヌーイ分布 $B(1,p)$ に従う

の場合を例にすると、標本変量 $(x_1,x_2,\cdots,x_n)$ の同時分布（結合分布）は

$$\begin{aligned} &f(x_1,x_2,\cdots,x_n|\ p)\\ &=f(x_1\mid p)f(x_2\mid p)\cdots f(x_n\mid p)\\ &=p^{x_1}q^{1-x_1}p^{x_2}q^{1-x_2}\cdots p^{x_n}q^{1-x_n}\\ &=p^{\sum x_i}q^{n-\sum x_i}\end{aligned}$$

であるから、尤度関数を $L(p)$ で表記すれば、

$$L(p)=p^{\sum x_i}(1-p)^{n-\sum x_i}$$

この $L(p)$ は、標本データ $(x_1,x_2,\cdots,x_n)$ が観測されたとき、$p\quad(0\le p\le 1)$ の各値に対する「尤もらしさ（likelihood）」の程度（尤度）を表す関数である。

最尤法は、「現実の標本データは尤度（確率）最大のものが実現した」という仮定（最尤原理）に基づいて、「尤もらしさ」$L(p)$ を最大にする p を、$(x_1,x_2,\cdots,x_n)$ の関数として求め、それを p の推定値とする方法（R.A.Fisher）である。

$\sum x_i=y$ とおくと、$L(p)=p^y(1-p)^{n-y}$ で、積の微分公式により

$$\frac{dL(p)}{dp}=p^{y-1}(1-p)^{n-y-1}(y-np)$$

だから、p に関する方程式 $\frac{dL(p)}{dp}=0$ を解いて、$L(p)$ の最大値を与える p を $\hat{p}$ と表

記すれば、

$$\hat{p}=\frac{y}{n}=\frac{1}{n}\sum x_i$$

これがpの**最尤推定値**である。標本データ$(x_1,x_2,\cdots,x_n)$を標本変量$(X_1,X_2,\cdots,X_n)$として扱えば、**最尤推定量** $\hat{p}=\frac{1}{n}\sum X_i$ となる。

一般に最尤法は、不偏推定量が求められない場合にも使え、漸近的には望ましい性質をもつ推定方法であるという意味で良い方式である。なお、対数尤度$l(\theta)=\ln L(\theta)$を未知パラメータθに関して微分するほうが計算は簡単になることが多い。

検定（けんてい）

ここでいう検定は**統計的仮設検定**の略称であり、統計的推論のタイプの一つとして実証研究でも頻繁に使われている。以下、簡単な例で、検定の考え方を説明しよう。

手元にある十円硬貨の１枚が歪んでいるように見えるので、その硬貨を多数回はじいてみて、歪んでいるか、歪んでないかをチェックすることにする。

そこで、「歪みがない」を仮説にして、「歪みがある」を比較対象の仮説とする。後者の比較対象仮説は、仮説を否定したときに想定されるもので**対立仮説**という。これに対し、前者の仮説を特に**帰無仮説**ということもあるが、以下では単に**仮説**と呼ぶ。もし、硬貨に歪みがなければ、表の出る確率pは0.5である（$p=0.5$）。一方、歪みがあれば、$p\neq 0.5$である。そうすると、仮説と対立仮説は、それぞれ、統計的仮説として

仮説 $H_0:p=0.5$

対立仮説 $H_1:p\neq 0.5$

と表現される。ここでの記号H_0、H_1は慣用記号であり、「H」はhypothesis（仮説）の頭文字に由来する。

「仮説 H_0が真」か、「対立仮説 H_1が真」のどちらかであるが、硬貨を多数回はじいた結果でどちらかを判定する、つまり**仮説を検定**するというわけである。このとき、「仮説H_0が真」なのに「仮説H_0が偽（対立仮説 H_1が真）」と判定する誤り（**第1種の誤り**）と、「仮説H_0が偽（対立仮説 H_1が真）」なのに「仮説H_0が真」と判定する誤り（**第２種の誤り**）を犯す可能性がある。両者の誤りを犯す確率は共に小さいほどよいことはいうまでもないが、一般には不可能なので、ふつう第1種の誤りを犯す確率α（これを**有意水準**という）を５％か１％に定めておいて、第２種の誤りを犯す確率をできるだけ小さくするような方式にする。

さて、もし「仮説H_0が真」なら、この硬貨をn回はじくと、標本比率$\hat{p}$は0.5近辺の数値が出やすく、一方もし「対立仮説 H_1が真」なら、標本比率$\hat{p}$は0.5から離れた数値が出やすい。そこで、標本比率$\hat{p}$の実現値が0.5から離れた値のとき、仮説 H_0を否定する（仮説 H_0を**棄却する**）。そうでないときは仮説　を棄却しない。

大標本（標本サイズnが大きいとき）の場合、「仮説H_0が真（$p=0.5$）」のもとでは標本比率$\hat{p}$を標準化した変量

$$Z_0=\frac{\hat{p}-0.5}{\sqrt{\frac{0.5\times 0.5}{n}}}$$

は近似的に標準正規分布$N(0,1)$に従うから、

$$P\left(|Z_0| \geq 1.96\right)$$

$$= P\left(\left|\frac{\hat{p}-0.5}{\sqrt{(0.5\times 0.5)/n}}\right| \geq 1.96\right) \cong 0.05$$

したがって、仮説 H_0を棄却する範囲（これを**棄却域**という）をあらかじめ

$z \leq -1.96$ または $z \geq 1.96$ と定めておき、

変量 $Z_0 = \dfrac{\hat{p}-0.5}{\sqrt{\dfrac{0.5\times 0.5}{n}}}$ の実現値がこの棄却域に入ったとき、仮説 H_0を棄却し、そうでないときは仮説 H_0を棄却しない、という方式をとれば、有意水準５％の仮説検定となる（両側検定）。この判定で仮説 H_0が棄却される（棄却されない）とき、仮説 H_0は有意水準５％で**有意である**（有意でない）という表現も使われる。検定で用いた変量 Z_0を特に**検定統計量**という。もし、この硬貨を60回はじいてそのうち38回表が出たなら、$\hat{p} = 0.633$、$z_0 = 2.066$ だから、仮説 H_0 は棄却される。

なお、$P\left(|Z_0| \geq 2.066\right) = 0.0388$ であるが、この0.0388を p 値という。p 値が有意水準より小さければ、仮説 H_0は棄却される。

硬貨が歪んでいて表が出やすいのではないかという状況ならば、

仮説 $H_0 : p = 0.5$

対立仮説 $H_1 : p > 0.5$

また、硬貨が歪んでいて裏が出やすいのではないかという状況ならば、

仮説 $H_0 : p = 0.5$

対立仮説 $H_1 : p < 0.5$

となり、有意水準５％の棄却域は、それぞれ、$z \geq 1.645$ 、$z \leq -1.645$ となる。前者を右片側検定、後者を左片側検定といい、両者まとめて片側検定という。

以上から、検定の一般的手順は次のとおり：仮説 H_0と対立仮説 H_1を設定→有意水準 α の値を決定→検定統計量とその実現値の棄却域を選定→標本データから検定統計量の実現値を計算→実現値が棄却域に入っていれば仮説 H_0を棄却、そうでなければ仮説 H_0を棄却しない（消極的に H_0を受容）。