【セミナー】

連続型確率変数とその分布

――社会調査の推測統計（3）――

小林　久高
KOBAYASHI Hisataka

1　はじめに

今回は連続型の確率変数について解説する。連続型確率変数とは、たとえば、0 から 3 までの値をとる確率変数を考えた場合、0,1,2,3 といった離散的な数ではなく、0～3 にあるすべての数を値とする確率変数だ。

連続型確率変数について理解するためには、離散型確率変数についての知識があったほうがいい。そこで本題に入る前に離散型確率変数とその分布についておさらいしておこう。

表 1　離散型確率変数の確率分布

確率変数X（ カードに書かれた数）

値x	1	2	3	計
確率p	1/4	2/4	1/4	1

図 1　離散型確率分布のグラフによる表現

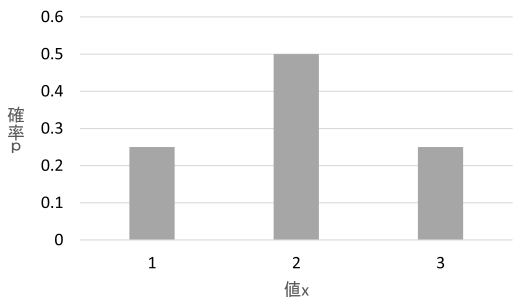


表 1 は、裏に 1 と書いてあるものが 1 枚、2 と書いてあるものが 2 枚、3 と書いてあるものが 1 枚の計 4 枚のカードからランダムに 1 枚選ぶ時、そのカードに書かれた数 X についての確率分布表である。図 1 はそれをグラフで表したものだ。横軸が値、縦軸が確率になっている。離散型の確率変数についてはこのような形で確率分布を表すことができる。

離散型確率変数については、次のような表記で確率を表す。

$$P(X=x_i)=f(x_i) \qquad (i=1,2,\cdots t)$$

これは言葉で言うと「X という確率変数が xi という値をとる確率が f(xi)である」となる。f(x)は確率質量関数とよばれる。

X は確率変数であるから、どの値についての確率も 0 以上であり、確率の計は 1 となる。

$$f(x_i)\geq 0 \qquad (i=1,2,\cdots t)$$

$$\sum_{i=1}^{t} f(x_i)=1$$

離散型確率変数 X の期待値と分散は次のようになる。

$$E(X)=\sum_{i=1}^{t} x_i p_i$$

$$V(X)=\sum_{i=1}^{t} (x_i-E(X))^2 p_i$$

確率変数の期待値は、「とりうる値×その確率」の総和で求められ、分散は「(とりうる値－期待値)の 2 乗×その確率」の総和で求められる。

4 枚のカードの例では次の通り。

$$E(X)=1\times\frac{1}{4}+2\times\frac{2}{4}+3\times\frac{1}{4}=2$$
$$V(X)=(1-2)^2\frac{1}{4}+(2-2)^2\frac{2}{4}+(3-2)^2\frac{1}{4}=\frac{2}{4}$$

確率変数と期待値だけを用いて分散を表すと次のようになる。

$$V(X)=E\left((X-E(X))^2\right)$$

以上が離散型確率変数についての基本事項だ。以下、連続型確率変数について説明していくが、離散型とどう違うのかということを確認していってほしい。

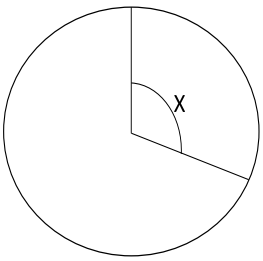
2　連続型確率変数の基礎

2.1 連続型確率変数の例

回る円盤に矢を放ち、基準の線から何度離れているのか（X）を当てると賞金がもらえるような賭けを考えてみよう（図 2)。「ばっちり当てたら 100 万円だよ。賭け金は 100 円だよ」などと言われても、このような賭けに決してのってはいけない。仮に 90 度に賭け、矢が 90 度に当たったように見えても、店の主人は目盛りの細かい分度器を持ち出して「90.1 度だからはずれ」と言うかもしれないからだ。こんなとき、矢は正確には 90.1345253652…のところにあったりする。この角度は連続変量なのである。連続変量である角度がばっちり 90 度になる確率は 0 だ。だからこんな賭けはかならず負けてしまう。

この角度 X のような連続型確率変数の分布を離散型で用いた表 1 のような確率分布表で表すことはできない。また、高さで確率を表す図 1 のようなグラフも使えない。

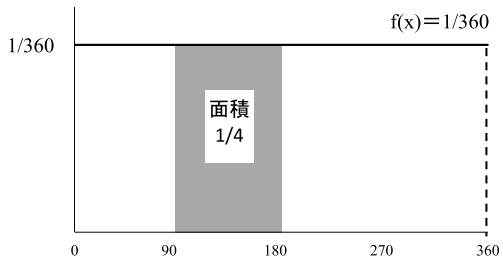
図 2　回る円盤の賭け



2.2 連続型確率変数の表し方

そこで面積で確率を表すことを考える。図 3 は、「円盤の賭け」の確率分布を示している。縦軸の 1/360 のところから横線が引かれているのは、こうしておくと横軸の 0 度～360 度の範囲で面積が 1 になるからだ(確率は全体で 1 でなければならない)。これを見ると、たとえば、90 度～180 度をとる確率（面積）は 1/4 などということがわかる。

図 3　回る円盤の確率は面積で表せる



面積で確率を表すときに役に立つのは積分だ。それゆえ連続型確率変数の分布は積分を使って示される（補参照）。

$$P(a \le X \le b) = \int_a^b f(x)dx$$

これを言葉で言うと、「X という確率変数が a 以上 b 以下の値をとる確率は $\int_a^b f(x)dx$ である」となる。

f(x)は確率密度関数と呼ばれ、図 3 の縦軸は確率密度を表している。確率密度とは何かということにこだわる必要はない。面積で確率を表すための縦軸と考えておけばいい。

「円盤の賭け」の確率密度関数は次のとおりである。

$$\begin{cases} f(x) = \dfrac{1}{360} & (0 \le x < 360) \\ f(x) = 0 & (x < 0,\ x \ge 360) \end{cases}$$

この確率密度関数から 90 度～180 度の確率を求めると次のようになる。

$$\int_{90}^{180} f(x)dx = \int_{90}^{180} \frac{1}{360}dx = \left[\frac{1}{360}x\right]_{90}^{180}$$
$$= \frac{180}{360} - \frac{90}{360} = \frac{90}{360} = \frac{1}{4}$$

確率密度関数の X は確率変数であるから、次の２つのことが成り立つ。

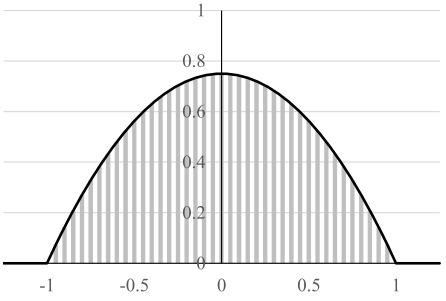
$$f(x) \ge 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

すなわち、確率密度関数を表す線はいつも横軸上かそれより上にあり、それと横軸で作られる面積は 1 になる。

確率密度関数はいろんな形をとる。たとえば次のような関数も確率密度関数と考えられる。

$$\begin{cases} f(x) = -\dfrac{3}{4}x^2 + \dfrac{3}{4} & (-1 \le x < 1) \\ f(x) = 0 & (x < -1,\quad x \ge 1) \end{cases}$$

図 4　確率密度関数の例



というのは、この関数の曲線は横軸上かそれより上にあり（図 4）、それと横軸で作られる面積は下の積分を見るとわかるように 1 になるからである。

$$f(x)dx = \int_{-1}^{1}\left(-\frac{3}{4}x^2 + \frac{3}{4}\right)dx = \frac{3}{4}\int_{-1}^{1}\left(-x^2 + 1\right)dx$$
$$= \frac{3}{4}\left[-\frac{1}{3}x^3 + x\right]_{-1}^{1} = \frac{3}{4}\left[\left(-\frac{1}{3} + 1\right) - \left(\frac{1}{3} - 1\right)\right]$$
$$= \frac{3}{4}\left(\frac{4}{3}\right) = 1$$

2.3 連続型確率変数の期待値と分散

連続型確率変数の期待値は次の式で表される。

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

$f(x)$は確率密度関数。

離散型の確率変数の期待値は、

$$E(X)=\sum_{i=1}^{t} x_i p_i = \sum_{i=1}^{t} x_i f(x_i)$$

というように、「値×確率」の総和だったのに対し、連続型では「値×確率密度」の積分で期待値が求められるのである。

「円盤の賭け」の例では、

$$\int_0^{360} xf(x)dx = \int_0^{360} x\frac{1}{360}dx = \frac{1}{360}\int_0^{360} xdx$$
$$= \frac{1}{360}\left[\frac{x^2}{2}\right]_0^{360} = \frac{1}{360}\left(\frac{360^2}{2}\right) = 180$$

連続型確率変数の分散は次の式で表される。

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

$f(x)$は確率密度関数。

離散型の確率変数の分散は、

$$V(X)=\sum_{i=1}^{t} (x_i - E(X))^2 p_i$$
$$= \sum_{i=1}^{t} (x_i - E(X))^2 f(x_i)$$

というように、「(値－期待値) の 2 乗×確率」の総和で求められるのに対し、連続型では「(値－期待値) の 2 乗×確率密度」の積分で分散が求められる。

「円盤の賭け」の例では、

$$\int_0^{360} (x-E(X))^2 f(x)dx = \int_0^{360} (x-180)^2 \frac{1}{360}dx$$
$$= \frac{1}{360}\int_0^{360} (x-180)^2 dx$$
$$= \frac{1}{360}\int_0^{360} (x^2 - 360x + 32400)dx$$
$$= \frac{1}{360}\left[\frac{1}{3}x^3 - \frac{360}{2}x^2 + 32400x\right]_0^{360}$$
$$= \frac{1}{360}\left(\frac{46656000}{3} - \frac{46656000}{2} + 1166400\right)$$
$$= \frac{129600}{3} - \frac{129600}{2} + 32400$$
$$= 43200 - 64800 + 32400 = 10800$$

2.4 期待値と分散の重要公式

以下の公式は離散型でも連続型でも成り立つ（証明は小林,2018a 参照）。

$E(c)=c$　（c は確率変数ではないきまった値）
$E(X+c)=E(X)+c$
$E(cX)=cE(X)$

$V(c)=0$
$V(X+c)=V(X)$
$V(cX)=c^2V(X)$

$E(X+Y)=E(X)+E(Y)$
$E(XY)=E(X)E(Y)$　（独立のとき）
$V(X+Y)=V(X)+V(Y)$　（独立のとき）

3　正規分布

3.1 正規分布の確率密度関数

連続型確率変数の分布の中で、統計学で最も重要な分布は正規分布だ。μ を平均、σ^2 を分散、π を 3.14…、e を 2.718…（自然対数の底）とするとき、正規分布の確率密度関数は次のようになる。

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

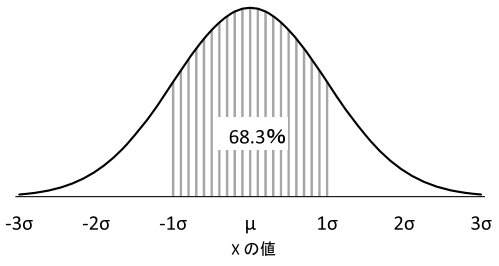
確率変数 X がこの分布に従うとき、平均μと分散σ^2を使って次のように表す（ここでの「～」は、左の確率変数が右の分布に従うという意味だ）。

$$X \sim N(\mu, \sigma^2)$$

正規分布のグラフは図 5 のように平均μを中心に左右対称の形になる。横軸は X の値であり縦軸は確率密度である。確率は面積で表される。曲線の下部の面積＝確率は全体で 1 だ。

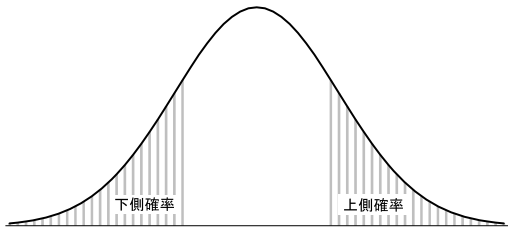
平均μを中心に両側に 1σ（標準偏差）の広がりをとると、面積は全体の約 68.3%になる（図 5）。平均から 2σをとると面積は約 95.5%になり、3σをとると約 99.7%になる。

図 5　正規分布



正規分布に限らず連続型の確率分布では、ある値より大きな値の範囲の確率を上側確率、ある値より小さな値の範囲の確率を下側確率という（図 6）。したがって、正規分布においてμの点の上側確率は 0.5、1σの点の上側確率は 0.158（(1－0.683)÷2）ということになる。

図 6　上側確率と下側確率



3.2 標準得点

ここでちょっと話題を変えて「標準得点」について述べておこう。変数 x が平均μ,分散σ^2の分布を持つとする。このとき、ケースの値 x_i を次のように変換したものを標準得点 z_i という。z の分布は平均 0、分散 1 となる。

$$z_i=\frac{x_i-\mu}{\sigma}$$

表 2 は元のデータであり、そこから表 3 の標準得点化されたデータが作成されている。たとえば C さんの教育年数の標準得点は次のように計算されたものだ。

$$\frac{18-14.2}{3.2}=1.17$$

表 2　元のデータ

得点

	年齢	教育年数	収入
Aさん	28	9	550
Bさん	35	16	560
Cさん	36	18	650
Dさん	49	16	630
Eさん	50	12	700
n	5	5	5
平均	39.6	14.2	618.0
分散	73.0	10.6	3176.0
標準偏差	8.5	3.2	56.4

表 3　標準得点化されたデータ

標準得点

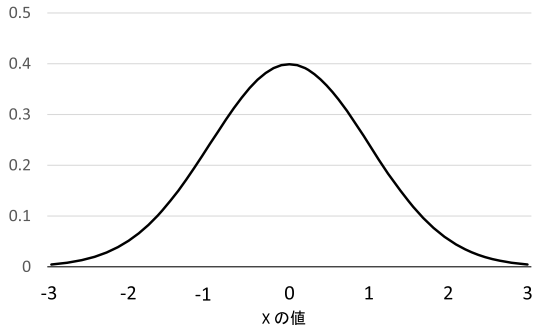
	年齢	教育年数	収入
Aさん	-1.36	-1.60	-1.21
Bさん	-0.54	0.55	-1.03
Cさん	-0.42	1.17	0.57
Dさん	1.10	0.55	0.21
Eさん	1.22	-0.68	1.46
n	5	5	5
平均	0.0	0.0	0.0
分散	1.0	1.0	1.0
標準偏差	1.0	1.0	1.0

点数を標準得点化すると、個々の得点の相対的な位置を把握しやすくなるし、変数間の値の比較も行いやすくなる。いわゆる偏差値は、平均を 50、標準偏差を 10 に置き換えて標準化した得点であり、標準得点が－1 のとき偏差値は 40、標準得点が 0 のとき偏差値は 50、標準得点が 1 のとき偏差値は 60、標準得点が 2 のとき偏差値は 70 となる。偏差値によって各人の試験での相対的な位置が把握しやすくなるとともに、さまざまな時に行われるさまざまな科目の試験成績の比較も容易になるということは、学生諸君もよく知っていることだろう。

3.3 標準正規分布

正規分布の話に戻ろう。正規分布の中で、特に、平均 0、分散 1（したがって標準偏差 1）の正規分布のことを標準正規分布といい、N (0, 1) で表す。標準正規分布のグラフは図 7 のようになる。

図 7　標準正規分布



確率密度関数は $\mu=0$、$\sigma^2=1$ をもとの式に代入して次のようになる。

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

ある確率変数 X が正規分布に従うとき、次の確率変数 Z は標準正規分布に従う。ここで行っている変換は、元の変数から平均を引き、標準偏差で割るというもので、標準得点の計算で行ったことと同じだ。

$$X\sim N(\mu,\sigma^2)$$
$$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$

このような変換によってどんな正規分布も標準正規分布になる。さまざまな正規分布の確率を積分で計算するのはたいへんな作業だが、標準正規分

布に変換するならば「数表」が利用でき、それを見ることで簡単に確率がわかる。

3.4 標準正規分布の数表

標準正規分布の数表は本稿の末尾に掲載されている。簡単にその読み方を説明しておこう。数表には２種類のものがある。１つは値 z から確率 α を求めるものである。これを抜粋したものが表 4 である。この数表には図 8 のような絵が描かれており、どの部分の確率の表なのかがわかるようになっている。ここで示されているのは、z より大きな値の範囲に対応する確率（上側確率）の表だということである。

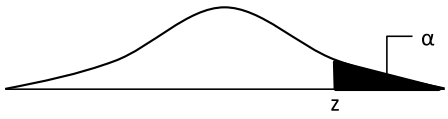
たとえば値が 1.96 のときの上側確率を求めるためには次のようにする。まず一番左の列から 1.9 を探し、そこから一番上の行を見ながら 0.06 の列まで右に移動していく。するとそこに示されている確率 0.0250 が 1.96（1.9＋0.06）という値の点の上側確率となる。

表 4　標準正規分布の数表（値→確率）

標準正規分布の値zに対応する上側確率 α

z	0.00	…	0.06	…	0.09
0.0	0.5000	…	0.4761	…	0.4642
0.1	0.4602	…	0.4365	…	0.4247
:	:	…	:	…	:
1.9	0.0288	…	0.0250	…	0.0233
:	:	…	:	…	:
3.0	0.0014	…	0.0012	…	0.0011

図 8　どの範囲の確率かを示す絵



もう１つの数表は確率 α から値 z を求めるものであり、表 5 のようなものである。

表 5　標準正規分布の数表（確率→値）

標準正規分布の上側確率 α に対応する点の値z(α)

α	0.000	…	0.005	…	0.009
0.00		…	2.5759	…	2.3657
0.01	2.3264	…	2.1701	…	2.0749
0.02	2.0538	…	1.9600	…	1.8957
0.03	1.8808	…	1.8120	…	1.7625
0.04	1.7507	…	1.6954	…	1.6547
0.05	1.6449	…	1.5982	…	1.5633

こちらの表で上側確率 0.025 の点の値を求めるときは、まず左の端の列を見て 0.02 を探し、そこから一番上の行を見ながら 0.005 の列まで右へ移動していく。そこに書かれている 1.96 が上側確率 0.025 に対応する点の値である。

正規分布は左右対称だから、1.96 の上側確率が 0.025 であるなら－1.96 の下側確率も 0.025 になる。したがって、－1.96 から＋1.96 の範囲の確率は 0.95（1－0.025×2）であることがわかる。ちなみに、－2.58 から＋2.58 の範囲の確率は 0.99 になる。これらの数値（1.96 と 2.58）は統計的検定ではよく利用される。

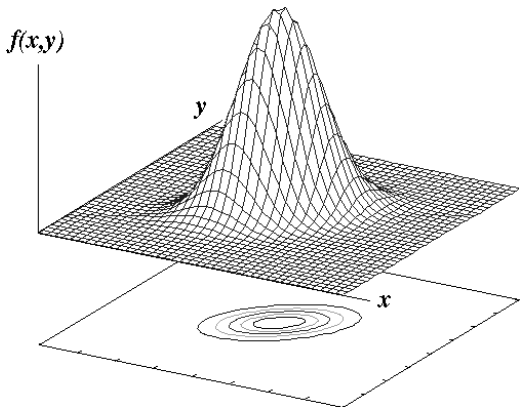
3.5 ２次元正規分布

図 9 のような曲面を同時確率密度関数とする分布を２次元正規分布という。x 軸に垂直な切り口も、y 軸に垂直な切り口もともに正規曲線になる。 ρ（相関係数）＝0 のときは、X,Y は独立に正規分布に従う。XY 平面の下にある楕円は等確率を示す等高線である。X と Y が独立のとき、この楕円は円になる。

具体的には、x 軸を身長、y 軸を体重とするとき、XY 平面上の a＜x＜b、c＜y＜d と曲面で囲まれる体積は、身長が a～b の範囲にあり、体重が c～d の範囲にあるものの割合を示す。確率として考えれば、任意に 1 人を取り出すとき、その者の身長

が a～b の範囲にあり、体重が c～d の範囲にある確率ということになる。離散型 2 次元確率分布のグラフと見比べておいてほしい（小林 2018b）。

図 9　２次元正規分布



4　正規分布に関わる重要事項

4.1 ラプラスの定理

（1）二項分布

二項分布とは次のような離散型の確率分布だ。１回の試行で、注目している事象の生じる確率が p だとする。その試行を独立に n 回繰り返したとき、注目している事象の生じる回数を X とする。X の分布は二項分布とよばれる。二項分布に従う確率変数は、回数という値をとり、その範囲は「0～試行回数」までの整数値である。

p を注目事象が１回の試行で生じる確率とするとき、注目事象が１回の試行で生じない確率は（1－p）である。二項分布は p と試行回数 n を用いて Bin（n, p）と表記する。確率変数 X（出現回数）がこの分布に従うとき下のように表す。

$$X \sim Bin(n, p)$$

p は注目している事象の 1 回の試行での出現確率だから、注目している事象の母比率とも考えられる。そうすると、二項分布が述べているのは、注目対象の母比率が p である母集団から、サイズ n の標本を取り出したとき、標本の中にいる注目対象の個数 X についての確率分布ということになる。

二項分布の期待値は np であり、分散は np(1－p)である。

（2）ラプラスの定理の内容

ラプラスの定理によると、n が十分大きいとき、二項分布に従う確率変数 X は、同じ平均・分散の正規分布に近似的に従う。二項分布の期待値と分散は np と np(1－p)なのでこの定理は結局次のことを言っていることになる。

$$X \sim Bin(n, p)$$

で n が大きいなら

$$X \sim N(np, np(1-p))$$

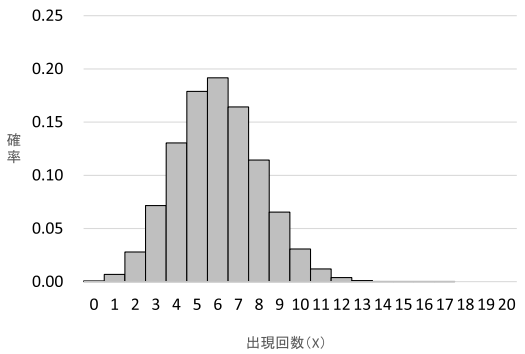
実用的な観点からは、np＞5 かつ n(1－p)＞5 ぐらいならば近似できると言われている（蓑谷 2003:483）。

ラプラスの定理をはじめて知ったとき筆者はとても不思議な気がした。というのは、二項分布は離散型の確率分布で縦軸は確率なのに対して、正規分布は連続型の確率分布で縦軸は確率密度で、確率は面積で表される。だから縦軸の異なる２つの分布がなぜ近似できるのだろうと疑問をもったのだ。

このことについて「はっ」と思ったのは、二項分布に従う確率変数がとりうる値を思い出した時だ。二項分布の値は 0,1,2,…といった 0 以上の整数

値をとる。整数値だから間隔が 1 である。これがミソだ。整数値を横軸に書き、棒グラフの棒の幅を 1 にしてとなりの棒にひっつけて並べると（ヒストグラム)、なんと確率は棒の高さだけでなく面積でも表されいるではないか（図 10)。このようにした場合、縦軸は確率だけでなく確率密度とも考えられる。だから近似もおかしくないのか…。そんな風にして筆者の素朴な疑問は解消したのだ。

図 10　二項分布 Bin(20, 0.3)



この疑問の解消は筆者の別の疑問の解消にもつながった。「○○が正規分布をしている」という表現はよくなされる。疑問が生じていたのはこの○○が離散型の変数のときである。このとき○○の確率はグラフの高さで表されるのに対し、正規分布の確率は面積だからこの表現はよくわからないと筆者は感じていたのだ。しかし二項分布の場合と同様、他の離散型の確率変数でも面積で確率を表すことはできる。そのためには変数の値をきれいに数直線上に並べてヒストグラムを描けばいい。ただ、グラフの棒の幅は二項分布と異なり 1 とは限らないので、面積が全体で 1 になるように縦軸の目盛をつけ替える必要がある。新しい縦軸の目盛は確率密度に準じたものになる。こうすることによって、その離散型の分布は正規分布に似ているかどうかが確認できるのだ。

（3）比率で考える

さて、話を二項分布に戻そう。二項分布は注目している事象が生じる回数についての分布であった。ここで、生じる回数ではなく生じる比率に焦点を置き、その比率を P’としよう。P’は注目している事象が標本で生じる比率（標本比率）を示す確率変数だ。

$$P' = \frac{X}{n}$$

この標本比率の期待値と分散はどうなるのかを考えよう。二項分布における出現回数 X の期待値と分散は次の通りだった。

$$E(X) = np$$
$$V(X) = np(1-p)$$

ここから出現比率 P’の期待値と分散次のようになる。

$$E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$
$$V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$
$$\because V(cX) = c^2V(X)$$

ここで、n が大きく、ラプラスの定理によって次のことが成り立っているとしよう。

$$X \sim N(np, np(1-p))$$

このとき、出現比率 P'については、上の期待値と分散を使って次のようになる。

$$P' = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n} \right)$$

すなわち、標本比率 P'は注目している事象の母比率 p を中心に分散 p(1－p)/n の広がりをもって正規分布する。

ここからさらに標準正規分布に従う確率変数を求めると次のようになる。

$$\frac{P' - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

すなわち、この比率に関する左の式は標準正規分布に従うのである。

4.2 正規分布の重要な性質

（1）aX＋b

正規分布には 2 つの重要な性質がある。1 つめは、正規分布に従う平均μ,分散σ^2 の確率変数 X があるとき、定数 a、b を用いた aX＋b という合成変数も正規分布に従うという性質だ。その時、平均は aμ ＋b、分散は $a^2\sigma^2$ になる。式で表すと次のようになる。

$$X \sim N(\mu, \sigma^2)$$

のとき、

$$aX + b \sim N(a\mu + b,\ a^2\sigma^2)$$

（2）正規分布の再生性

もう 1 つの重要な性質は正規分布の再生性と呼ばれている。それは X と Y が正規分布に従う独立の確率変数であるとき、X＋Y や X－Y という合成的確率変数も正規分布に従うという性質である。

一般に、確率変数の合成変数の期待値や分散については次のことが成り立つ。

$$E(X+Y) = E(X) + E(Y)$$
$$V(X+Y) = V(X) + V(Y) \quad \text{（独立のとき）}$$
$$E(X-Y) = E(X) - E(Y)$$
$$V(X-Y) = V(X) + V(Y) \quad \text{（独立のとき）}$$

X－Y の分散が足し算になる理由は次の通りだ。

$$V(X-Y) = V(X) + V(-Y)$$
$$= V(X) + (-1)^2 V(Y) = V(X) + V(Y)$$

したがって、X、Y が独立で、X が平均μ_A、分散σ_A^2の分布に従い、Y が平均μ_B、分散σ_B^2の分布に従うとき、X＋Y は平均$\mu_A + \mu_B$、分散$\sigma^2_A + \sigma_B^2$の分布に従うことになる。また、X－Y は平均$\mu_A - \mu_B$、分散$\sigma^2_A + \sigma^2_B$の分布に従う。このことはどんな確率分布でも成り立つ。

正規分布の場合は正規分布であるという特徴も引き継がれる。すなわち、

$$X_A \sim N(\mu_A, {\sigma_A}^2)$$
$$X_B \sim N(\mu_B, {\sigma_B}^2)$$

のとき、

$$X_A + X_B \sim N(\mu_A + \mu_B,\ \ {\sigma_A}^2 + {\sigma_B}^2)$$
$$X_A - X_B \sim N(\mu_A - \mu_B,\ \ {\sigma_A}^2 + {\sigma_B}^2)$$

これが正規分布の再生性だ。

4.3 中心極限定理

同一の分布（平均μ,分散σ^2）に従う独立な確率変数X_1～X_nの平均$\overline{X}$について、次のことが成り立つというのは推測統計の基本中の基本だ。

$$\overline{X}=\frac{1}{n}\left(X_1+X_2+\cdots+X_n\right)$$

とすると、$\overline{X}$は、平均μ,分散σ^2/nの分布に従う。

このことは、母集団と標本という言葉を使うと次のように表現できる。平均μ,分散σ^2の母集団からサイズ n の標本をとるとき、標本平均の期待値と標本平均の分散はそれぞれμ、σ^2/nになる。

さらに、同一の正規分布に従うX_1～X_nの場合は、$\overline{X}$も正規分布に従う。すなわち、平均μ,分散σ^2の正規母集団からのサイズ n の標本について、標本平均$\overline{X}$は、平均μ,分散σ^2/nの正規分布に従うのである。これは 4.2 で述べた「正規分布の重要な性質」による。

中心極限定理はさらにすごい定理である。これが言うのは次のことだ。n が大きいとき、同一の分布（平均μ,分散σ^2）に従う独立な確率変数 X_1～X_nの平均$\overline{X}$は、平均μ,分散σ^2/n の「正規分布」に従う。

どこがすごいのか見落としてしまいそうなのだが、母集団が「どんな分布でも」標本平均$\overline{X}$の分布は正規分布になるというところがすごいのだ。

検定や推定では標本平均などの標本統計量が正規分布することが利用されることが多い。母集団が正規分布している場合は標本平均も正規分布することがわかっているのでいい。しかし、母集団が正規分布しない場合は困ってしまう。だが、この中心極限定理があるおかげで、標本サイズさえ大きくすれば標本平均は正規分布をするので心配はいらないのである。

5　その他の連続型確率分布

5.1 二乗和の分布のイメージ

今まで正規分布について述べてきたが、初歩的な推測統計学では正規分布のほかにχ^2分布（カイ二乗分布）、t 分布、F 分布という連続型の確率分布をよく利用する。それらは標準正規分布に従う確率変数をもとに作られた合成的確率変数の分布である。これらの確率分布の中で、標準正規分布に従う確率変数の二乗和の分布であるχ^2分布はもっとも基本的なものである。しかし、確率変数の二乗和の分布についてのイメージがすぐにわく読者は少ないだろう。そこで、ここではまず離散型確率変数の二乗和を例に、二乗和の分布のイメージを獲得することにしよう。

3 枚のカードの裏に−1、0、1 という数字が書いてあり、そこからランダムに 1 枚選んだカードの裏の数を確率変数 X とする。このとき確率分布表は表 6 のようになる。

表 6　X の確率分布

X				
値	-1	0	1	計
確率	1/3	1/3	1/3	1

この分布の期待値と分散は次のようになる。

$$E(X)=(-1)\left(\frac{1}{3}\right)+0\left(\frac{1}{3}\right)+1\left(\frac{1}{3}\right)=0$$

$$V(X)=(-1-0)^2\left(\frac{1}{3}\right)+(0-0)^2\left(\frac{1}{3}\right)+(1-0)^2\left(\frac{1}{3}\right)$$

$$=\frac{2}{3}$$

次にこの確率変数の 2 乗についての確率分布を考えると表 7 のようになる。

表 7　X^2 の確率分布（1）

X^2				
値	$(-1)^2$	0^2	1^2	計
確率	1/3	1/3	1/3	1

これを整理すると表 8 になる。

表 8　X^2 の確率分布（2）

X^2			
値	0	1	計
確率	1/3	2/3	1

この分布の期待値と分散は次のようになる。

$$E\left(X^2\right)=\frac{2}{3}=\frac{18}{27}$$
$$V\left(X^2\right)=\frac{2}{27}=\frac{6}{81}$$

X の分布に従う独立した確率変数が 2 つあるとき（すなわち、2 回カードを復元抽出で選ぶとき）、その二乗和の値と確率を計算すると表 9 の同時確率分布表が得られる。

表 9　二乗和の値と確率（1）

$X_1^2+X_2^2$				
		X_1^2		
		0	1	
X_2^2	0	0(1/9)	1(2/9)	1/3
	1	1(2/9)	2(4/9)	2/3
		1/3	2/3	

値（確率）

表では「$x_1^2+x_2^2$ の値（確率）」という形で示されている。たとえば、表右下の「2(4/9)」は、1 枚目も 2 枚目も 2 乗したら 1 になるカードが出るなら二乗和は 2 になり、そうなる確率が 4/9 であることを示している。2 つの確率変数 X_1^2 と X_2^2 は独立しているので、それらの和の値（2）に対応する確率は周辺にある確率の掛け算（2/3×2/3＝4/9）で得られる。これを確率分布表にまとめると表 10 のようになる。

表 10　2 個の二乗和の確率分布

$X_1^2+X_2^2$				
値	0	1	2	計
確率	1/9	4/9	4/9	1

$$E\left(X_1^2+X_2^2\right)=\frac{4}{3}=\frac{36}{27}$$
$$V\left(X_1^2+X_2^2\right)=\frac{20}{81}$$

さらに、もうひとつの確率変数 X_3^2 も加えることを考えるとき、その値と確率は表 11 のように計算できる。

表 11　二乗和の値と確率（2）

$X_1^2+X_2^2+X_3^2$					
		$X_1^2+X_2^2$			
		0	1	2	
X_3^2	0	0(1/27)	1(4/27)	2(4/27)	1/3
	1	1(2/27)	2(8/27)	3(8/27)	2/3
		1/9	4/9	4/9	1

値（確率）

これを整理すると表 12 のようになる。

表 12　3 個の二乗和の確率分布

$X_1^2+X_2^2+X_3^2$

値	0	1	2	3	計
確率	1/27	6/27	12/27	8/27	1

$$E\left(X_1^2+X_2^2+X_3^2\right)=2=\frac{54}{27}$$
$$V\left(X_1^2+X_2^2+X_3^2\right)=\frac{14}{27}=\frac{42}{81}$$

二乗和の確率変数のイメージがつかめただろうか。－1、0、1 の値をとる確率変数 X について、2 乗の確率変数 X^2 は負の値をとらないし、$X_1^2+X_2^2+\cdots+X_n^2$ も負の値をとらない。加える二乗の数が増えると分布は変わる。そして、加えるごとに期待値や分散が大きくなっている。以下、連続型確率変数の分布である χ^2（カイ二乗）分布、t 分布、F 分布について述べるが、このイメージを頭の隅に置いておいてほしい。

5.2 χ^2（カイ二乗）分布

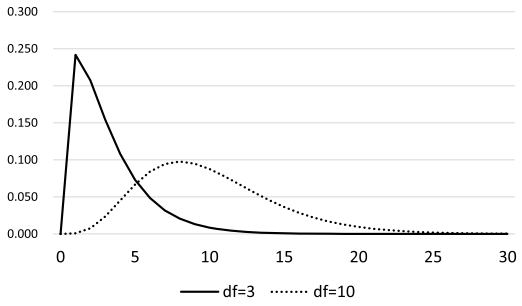
（1）確率密度関数

$$\chi^2{}_n(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & (x>0)\\ 0 & (x\le 0)\end{cases}$$

Γはガンマ関数

（2）分布の形

図 11　χ^2 分布



自由度によって分布の形は変わる。

（3）性質

互いに独立な n 個の確率変数、$Z_1,Z_2\cdots,Z_n$ が標準正規分布 N(0, 1)に従うとき、確率変数 $X=Z_1^2+Z_2^2+\cdots+Z_n^2$ は、自由度 n の χ^2 分布に従う。

これは標準正規分布に従う確率変数（Z_i）の二乗和の分布である。自由度とはこの二乗和のもととなった二乗（Z_i^2）の数である。自由度は df で表すことが多い。また、自由度 n の χ^2 分布は χ_n^2 と表される。

X は自由度 m の、Y は自由度 n の χ^2 分布に従うとき、確率変数 X＋Y は、自由度（m＋n）の χ^2 分布に従う。再生性があるのである。

自由度 n の χ^2 分布に従う確率変数 X の期待値と分散は次のようになる。

$$E(X)=n$$
$$V(X)=2n$$

期待値については、下のように証明できるが、分散の証明はなかなかやっかいだ。

$$X=Z_1^2+Z_2^2+\cdots+Z_n^2$$
$$E(X)=E\left(Z_1^2+Z_2^2+\cdots+Z_n^2\right)$$

$$= E\left(Z_1^2\right) + E\left(Z_2^2\right) + \ldots + E\left(Z_n^2\right)$$
$$= \underbrace{1+1+\cdots+1}_{n} \quad = n$$
$$\because E\left(Z_i^2\right) = V\left(Z_i\right) + \left(E\left(Z_i\right)\right)^2 = 1 + 0 = 1$$

5.3 t 分布

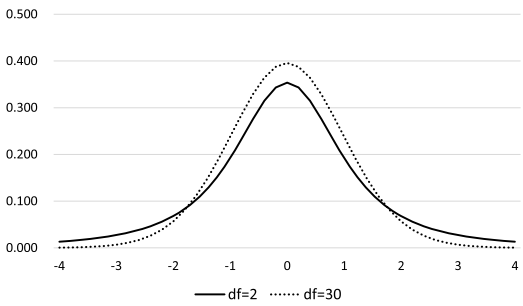
（1）確率密度関数

$$t_n(x) = \frac{1}{\sqrt{n}B\left(\frac{n}{2},\frac{1}{2}\right)}\left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}} \quad \left(-\infty < x < \infty\right)$$

B はベータ関数

（2）分布の形

図 12　t 分布



自由度によって分布の形は異なる

（3）性質

２つの独立な確率変数 Y と Z があり、Z は標準正規分布 N(0,1)に、Y は自由度 n の χ^2 分布に従うとき、確率変数 $X = \frac{Z}{\sqrt{Y/n}}$ は、自由度 n の t 分布に従う。

t 分布は左右対称の分布である。n≧30 で t 分布は標準正規分布とほぼ等しくなる。 n→∞になるとき、 t 分布は標準正規分布と同じ分布になる。

自由度 n（n＞2）の t 分布に従う確率変数 X の期待値と分散は次のようになる。

$$E(X) = 0$$
$$V(X) = \frac{n}{n-2} \quad (n > 2)$$

5.4 F 分布

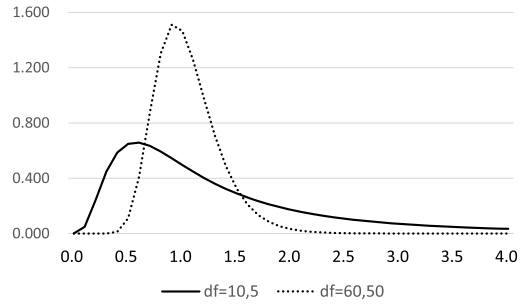
（1）確率密度関数

$$F_n^m(x) = \begin{cases} \dfrac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2},\frac{n}{2}\right)} \left(\dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} \right) & (x > 0) \\ 0 \quad (x \le 0) \end{cases}$$

B はベータ関数

（2）分布の形

図 13　F 分布



F 分布の形は 2 つの自由度によって決定される。

（3）性質

２つの独立な確率変数 Y と Z があり、Y は自由度 m の、Z は自由度 n の χ^2 分布に従うとき、確率

変数 $X=\dfrac{Y/m}{Z/n}$ は、自由度(m,n)の F 分布に従う。

自由度（m,n）の F 分布に従う確率変数 X の期待値と分散は次のようになる。

$$E(X)=\frac{n}{n-2}\qquad (n>2)$$
$$V(X)=\frac{2n^2(n+m-2)}{m(m-2)^2(n-4)}\quad (n>4)$$

F 分布の自由度を（m,n）とし、分布の上側確率を $F_n^m(\alpha)$ とするとき、次の式が成り立つ。

$$F_n^m(\alpha)=\frac{1}{F_m^n(1-\alpha)}$$

この性質は検定で利用することになる。

5.5 ４つの確率分布の相互関係

上で見てきた標準正規分布、χ^2 分布、t 分布、F 分布は相互に関係している。

今、Z を標準正規分布に従う確率変数とし、K_n を自由度 n の χ^2 分布に従う確率変数とする。また、T_n を自由度 n の t 分布に従う確率変数とし、F_n^m を自由度 m,n の F 分布に従う確率変数としよう。このとき次のようになる（右辺の変数はすべて独立）。

$$K_1=Z^2$$
$$K_n=\sum_{i=1}^{n}Z_i^{\,2}=Z_1^2+Z_2^2+\cdots+Z_n^2$$
$$T_n=\frac{Z}{\sqrt{K_n/n}}$$
$$F_n^m=\frac{K_m/m}{K_n/n}$$

これらの関係から、いろいろなことが見えてくる。たとえば、K は二乗和だ。だから負になることはない。χ^2 分布の期待値は n、分散は 2n と書いてあった。だから加える二乗の数が増えるほど分布の中心は右へ移動し、どんどん広がった分布になる。χ^2 分布はあんな形をしているのはこういうわけだ。

χ^2 分布には再生性があると述べたが、それも理解できる。K は Z の二乗和だから当然再生性はあるわけだ。

$$K_m+K_n=\sum_{i=1}^{m}Z_{ai}^{\,2}+\sum_{j=1}^{n}Z_{bj}^{\,2}$$
$$=\left(Z_{a1}^2+Z_{a2}^2\cdots+Z_{am}^2\right)+\left(Z_{b1}^2+Z_{b2}^2\cdots+Z_{bn}^2\right)$$
$$=K_{m+n}$$

分散というものは、「期待値からのズレの 2 乗の和」をもとに求められていた。だから二乗和である K は分散に関係するはずだ。おそらく、χ^2 分布は分散に関わる検定で利用されるのだろう。

F は K の比に関わる。だから F も負になることはない。だから F 分布はあんな形をしているのだ。K が分散に関係するなら、F は分散の比に関係するだろう。そういった検定で F 分布は利用されるのだろう。

関係を見ていると、T と Z が深い関係にあることもわかる。そして、n→∞のときに T は Z に等しくなる理由も見えてくる

χ^2 分布の期待値は n、分散は 2n であると述べた。すなわち、

$$E(K_n)=n$$
$$V(K_n)=2n$$

確率分布間の関係を見ていると、ここから次のよ

うになっていることがわかる。

$$E\left(Z^2\right)=E\left(K_1\right)=1$$
$$V\left(Z^2\right)=V\left(K_1\right)=2$$

ここでこの Z^2 の分布に従う n 個の確率変数を考えよう。そのときその平均は、

$$\overline{Z_n^2}=\frac{1}{n}\left(Z_1{}^2+Z_2{}^2+\cdots+Z_n{}^2\right)$$

ここで、標本平均の平均が母平均になり、標本平均の分散が母分散を標本数で割ったものになるということを思い出すと、次のようになることがわかる。

$$E\left(\overline{Z_n^2}\right)=1$$
$$V\left(\overline{Z_n^2}\right)=\frac{2}{n}$$

ところで、T_n という確率変数は次のようなものだった。

$$T_n=\frac{Z}{\sqrt{K_n/n}}$$

この分母の根号内は、$\overline{Z_n^2}$ と同じだ。だから根号内の分布の期待値は 1 で分散は 2/n である。ここで n がどんどん大きくなると、根号内の分布の分散はどんどん小さくなっていく。そして n→∞のとき根号内の分散は 0 になり、根号内の値は期待値である 1 に等しくなる。したがって、T は Z になるのだ。なるほどそうだったのか…、となるわけである。

少々ややこしい話をしたが確率分布相互の関係をイメージできるようになるのは重要だ。そのイメージがあると、将来、推測や検定を行う際に、なぜその確率分布を使うのかということがなんとなくわかるようになるのである。

6　おわりに

以上で連続型確率変数の話は終わりである。連続型確率変数の問題に限らず、統計学の理解においてまず重要なのは母集団の確率分布と標本平均の確率分布をきちんと区別することだ。そして「標本平均の期待値が母平均に一致し、標本平均の分散が母分散を標本サイズで割ったものになる」ことを必ず知っておく必要がある。これは本稿を読む前に学んでおくべき知識なのでここでは十分解説しなかったが、きちんと理解しておいてほしい。小林（2018a）が役に立つはずだ。

今回、連続型の確率分布は確率密度関数の曲線で表現され、確率はそれを使って面積で表されることについて説明した。有名な正規分布はこういう確率分布だったのだ。「母集団の確率分布がどのようなものでも、標本サイズが大きければ標本平均は正規分布をする」という中心極限定理はとても重要だ。この定理があるため、さまざまな分布に従う母集団について、統計的な推定や検定ができるのである。

基礎的な推測統計で主として用いる確率分布として、標準正規分布、χ^2 分布、t 分布、F 分布があるが、それらについても今回簡単に解説した。それぞれの確率密度関数の数式の導き方などは数学者にまかせておいていい。ただ、それらがおおよそどんな形をしており、お互いにどういう関係をもっているのかは知っておいたほうがいいだろう。こういった知識は推定や検定の原理の理解に役に立つからだ。

文献

馬場敬之・久池井茂, 2003『確率統計キャンパスゼミ』マセマ
小林久高,2018a「母集団・標本・確率変数」『同志社社会学研究』22
小林久高,2018b「離散型確率変数とその分布」『同志社社会学研究』22
蓑谷千凰彦, 2003『統計分布ハンドブック』朝倉書店

補：微分と積分について

1　微分

関数 f(x)の微分とは、その関数の導関数 f'(x)を求めること。

$$f(x)=3x^2+4x+5$$

のとき、

$$f'(x)=6x+4$$

となる。

高校の数学Ⅱで学ぶ基本的な公式は、次のようなもの。

$$\left(x^a\right)'=ax^{a-1}$$
$$\left(f(x)+g(x)\right)'=f'(x)+g'(x)$$

数学Ⅲや大学の数学では、下の公式を含め多数の関数の微分公式を学ぶ。

$$(px+q)^n=n(px+q)^{n-1}$$

2　積分

関数 f(x)の積分とはその関数の原始関数 F(x)を求めること。積分は微分と逆のことをしている。

$$\int f(x)dx=\int(6x+4)dx$$
$$=\frac{6}{2}x^2+4x+C=3x^2+4x+C$$
$$F(x)=3x^2+4x+C$$

高校の数学Ⅱで学ぶ基本的な公式は、次のようなもの。

$$\int x^a dx=\frac{1}{a+1}x^{a+1}+C$$
$$\int\left(f(x)+g(x)\right)dx=\int f(x)dx+\int g(x)dx$$

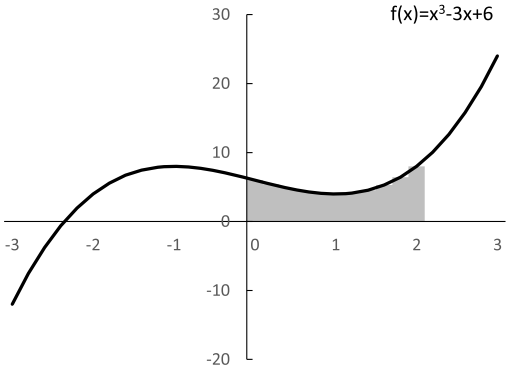
3　定積分

$$\int_a^b f(x)dx=\left[F(x)\right]_a^b=F(b)-F(a)$$

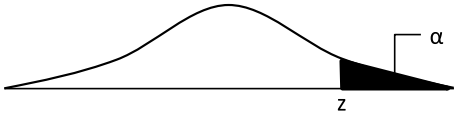
例

$$\int_0^2\left(x^3-3x+6\right)dx=\left[\frac{1}{4}x^4-\frac{3}{2}x^2+6x\right]_0^2$$
$$=\left(\frac{16}{4}-\frac{12}{2}+12\right)-\left(\frac{0}{4}-\frac{0}{2}+0\right)=10$$

これは下の図の網掛け部の面積を表す。

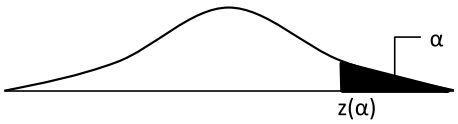


数表（標準正規分布）



標準正規分布の値zに対応する上側確率 α

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4961	0.4921	0.4881	0.4841	0.4801	0.4761	0.4721	0.4682	0.4642
0.1	0.4602	0.4563	0.4523	0.4483	0.4444	0.4404	0.4365	0.4326	0.4286	0.4247
0.2	0.4208	0.4169	0.4130	0.4091	0.4052	0.4013	0.3975	0.3936	0.3898	0.3860
0.3	0.3821	0.3783	0.3745	0.3707	0.3670	0.3632	0.3595	0.3557	0.3520	0.3483
0.4	0.3446	0.3410	0.3373	0.3336	0.3300	0.3264	0.3228	0.3192	0.3157	0.3121
0.5	0.3086	0.3051	0.3016	0.2981	0.2946	0.2912	0.2878	0.2844	0.2810	0.2776
0.6	0.2743	0.2710	0.2677	0.2644	0.2611	0.2579	0.2547	0.2515	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2267	0.2237	0.2207	0.2177	0.2148
0.8	0.2119	0.2090	0.2062	0.2033	0.2005	0.1977	0.1949	0.1922	0.1895	0.1868
0.9	0.1841	0.1815	0.1788	0.1762	0.1737	0.1711	0.1686	0.1661	0.1636	0.1611
1.0	0.1587	0.1563	0.1539	0.1516	0.1492	0.1469	0.1446	0.1424	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1293	0.1272	0.1251	0.1231	0.1211	0.1191	0.1171
1.2	0.1151	0.1132	0.1113	0.1094	0.1075	0.1057	0.1039	0.1021	0.1003	0.0986
1.3	0.0969	0.0951	0.0935	0.0918	0.0902	0.0886	0.0870	0.0854	0.0838	0.0823
1.4	0.0808	0.0793	0.0779	0.0764	0.0750	0.0736	0.0722	0.0708	0.0695	0.0682
1.5	0.0669	0.0656	0.0643	0.0631	0.0618	0.0606	0.0594	0.0583	0.0571	0.0560
1.6	0.0548	0.0537	0.0527	0.0516	0.0506	0.0495	0.0485	0.0475	0.0465	0.0456
1.7	0.0446	0.0437	0.0428	0.0419	0.0410	0.0401	0.0393	0.0384	0.0376	0.0368
1.8	0.0360	0.0352	0.0344	0.0337	0.0329	0.0322	0.0315	0.0308	0.0301	0.0294
1.9	0.0288	0.0281	0.0275	0.0269	0.0262	0.0256	0.0250	0.0245	0.0239	0.0233
2.0	0.0228	0.0223	0.0217	0.0212	0.0207	0.0202	0.0197	0.0193	0.0188	0.0184
2.1	0.0179	0.0175	0.0171	0.0166	0.0162	0.0158	0.0154	0.0151	0.0147	0.0143
2.2	0.0140	0.0136	0.0133	0.0129	0.0126	0.0123	0.0120	0.0117	0.0114	0.0111
2.3	0.0108	0.0105	0.0102	0.0100	0.0097	0.0094	0.0092	0.0089	0.0087	0.0085
2.4	0.0082	0.0080	0.0078	0.0076	0.0074	0.0072	0.0070	0.0068	0.0066	0.0064
2.5	0.0063	0.0061	0.0059	0.0058	0.0056	0.0054	0.0053	0.0051	0.0050	0.0048
2.6	0.0047	0.0046	0.0044	0.0043	0.0042	0.0041	0.0040	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0029	0.0028	0.0027
2.8	0.0026	0.0025	0.0025	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020
2.9	0.0019	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014
3.0	0.0014	0.0014	0.0013	0.0013	0.0012	0.0012	0.0012	0.0011	0.0011	0.0011



標準正規分布の上側確率 α に対応する点の値z(α)

α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.00		3.0903	2.8782	2.7478	2.6521	2.5759	2.5122	2.4573	2.4090	2.3657
0.01	2.3264	2.2904	2.2572	2.2263	2.1973	2.1701	2.1445	2.1201	2.0970	2.0749
0.02	2.0538	2.0336	2.0141	1.9954	1.9774	1.9600	1.9432	1.9269	1.9111	1.8957
0.03	1.8808	1.8663	1.8522	1.8385	1.8251	1.8120	1.7992	1.7867	1.7744	1.7625
0.04	1.7507	1.7392	1.7280	1.7169	1.7061	1.6954	1.6850	1.6747	1.6646	1.6547
0.05	1.6449	1.6353	1.6258	1.6165	1.6073	1.5982	1.5893	1.5805	1.5718	1.5633