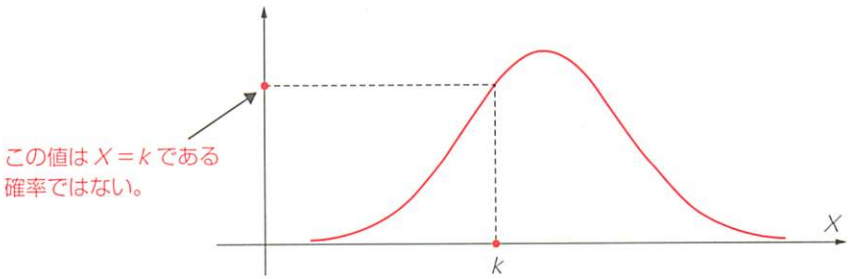
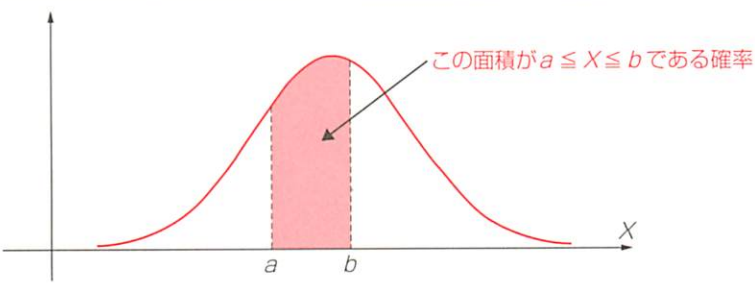
連続型と離散型では確率分布の解釈が異なる

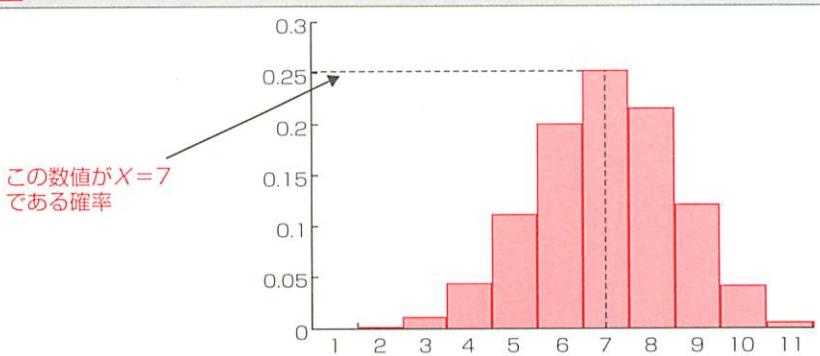
1 連続型の確率分布ではグラフの高さは確率そのものではない



2 連続型の確率分布では面積で確率を表す



3 離散型ではグラフの高さが確率



§65 サイコロの確率分布は二項分布

ここが要点！ サイコロを10回振れば1の目の出る回数は，0回から10回までまちまちです。しかし，1の目が何回出るかの確率は決まっています。じつは，このときの確率分布が二項分布と呼ばれるものです。

やさしい解説

サイコロを10回振る試行は，互いに独立です。したがって，10回中r回1の目が出る確率は，次のようになります。

まず10回中r回，1の目の出る出方は，10回中どのr回で1の目が出るかということなので${}_{10}C_r$通りあります。

10個の席からr個選んで，そこは1の目と考える

①②③④⑤⑥⑦⑧⑨⑩

このおのおのの場合の確率は1の目がr回，1でない目が$10-r$回なので，

$$\frac{1}{6}\times\frac{5}{6}\times\cdots\times\frac{1}{6}\times\cdots\times\frac{5}{6}=\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{10-r}$$

となります。

したがって10回中r回，1の目が出る確率は，

$${}_{10}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{10-r}$$

となります。

注）記号${}_nC_r$は，異なるnのものからr個選び出す選び出し方の総数のことです。

$${}_nC_r=\frac{n!}{(n-r)!\,r!}\quad,\quad {}_nC_0={}_nC_n=1$$

一般に，1回の試行である事象Aの起こる確率がpであるとき，この試行を独立にn回繰り返したとき，事象Aがr回起こる確率は，

$${}_nC_r p^r(1-p)^{n-r}\quad\cdots\cdots\cdots\cdots①$$

となります。

「確率変数Xがrという値をとる確率が式①で与えられるとき，この確率分布」を**二項分布**といいます。

この分布は，同じことを何回も繰り返したとき，ある事柄が何回起こるかの確率分布であり，日常生活に多く見られます。

注）二項分布はnとpだけで決まりますから，この分布を簡単に$B(n,\ p)$と表すことがあります。このBは，Binominal distribution（二項分布）の頭文字です。

二項分布の確率は ${}_nC_r p^r(1-p)^{n-r}$

確率変数 X が r である確率が，${}_nC_r p^r(1-p)^{n-r}$ である確率分布を二項分布という。

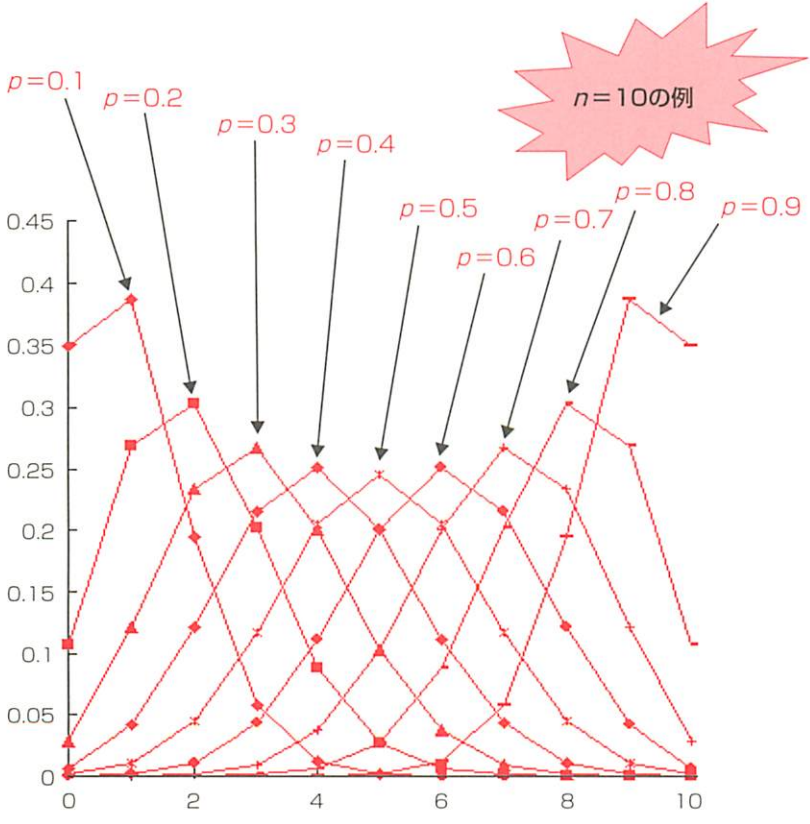
1 二項分布は次の確率分布表に従う

Xのとる値	0	1	2	……	r	……	n
確率	${}_nC_0 p^0 q^n$	${}_nC_1 p^1 q^{n-1}$	${}_nC_2 p^2 q^{n-2}$	……	${}_nC_r p^r q^{n-r}$	……	${}_nC_n p^n q^0$

ただし，$q=1-p$

確率変数 X の平均は np
確率変数 X の分散は npq

2 二項分布 $B(n,\ p)$ のグラフは離散型



§66 何回目に初めて表が出るかは幾何分布

ここが要点！ コインを何度も投げたとき，早くも1回目に表が出てしまうこともあれば，なかなか表が出ないこともあります。そこで，1回目，2回目，3回目，……に初めて表の出る確率を求めてみましょう。じつは，この場合の確率分布が「幾何分布」と呼ばれるものなのです。

やさしい解説

1回目に初めて表が出る確率 $P(1)$ は $1/2$ です。2回目に初めて表が出る確率 $P(2)$ は，1回目が裏で2回目が表ですから，

$$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)=\left(\frac{1}{2}\right)^2$$

となります。

3回目に初めて表が出る確率 $P(3)$ は，1回目，2回目が裏で3回目が表ですから，

$$\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)=\left(\frac{1}{2}\right)^3$$

となります。

このように考えていくと，X回目に初めて表が出る確率 $P(X)$ は，1回目から $X-1$ 回目までがすべて裏で，X 回目が表ですから，

$$\left(\frac{1}{2}\right)^{X-1}\left(\frac{1}{2}\right)=\left(\frac{1}{2}\right)^X$$

となります。なお，このとき数列の

$P(1)$，$P(2)$，$P(3)$，$\cdots P(X)$，$\cdots$

は初項1，公比1/2の等比（幾何）数列

$$\left(\frac{1}{2}\right),\ \left(\frac{1}{2}\right)^2,\ \left(\frac{1}{2}\right)^3,\ \cdots,\ \left(\frac{1}{2}\right)^X,\ \cdots$$

となります。

一般に，1回の試行で事象 A の起こる確率が p であるとき，X回目で初めて事象 A が起こる確率 $P(X)$ は，コインのときと同様に考えると，次のようになります。

$$P(X)=(1-p)^{X-1}p \quad \cdots\cdots\cdots\cdots ①$$

このとき「$P(1)$，$P(2)$，$P(3)$，$\cdots P(X)$，$\cdots$ は，初項 p，公比 $(1-p)$ の等比（幾何）数列」になります。

$P(X)\geqq 0$ で，

$P(1)+P(2)+P(3)+\cdots+P(X)+\cdots=1$

となるので，$P(X)$ は確率関数になります。この確率分布は，式①が等比（幾何）数列になることにちなんで**幾何分布**と呼ばれています。

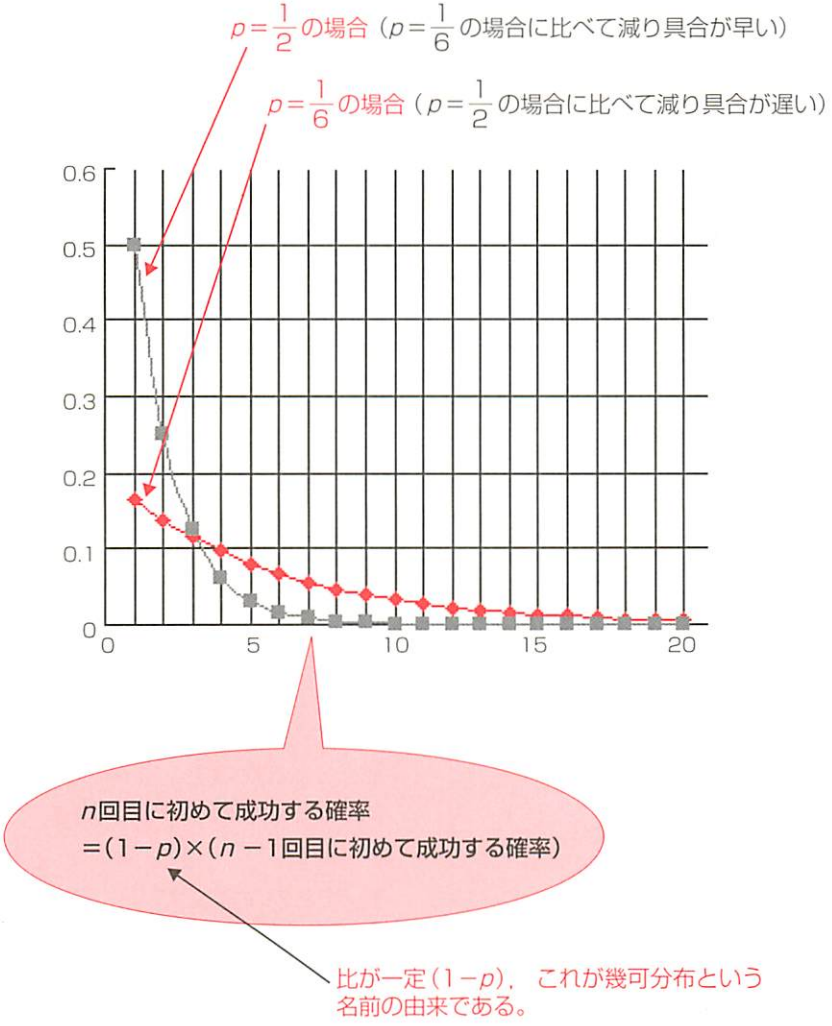
なお，事象 A の起こることを成功とみなせば，X回目で初めて成功する確率が $P(X)=(1-p)^{X-1}p$ となります。

注1）等比数列のことを幾何数列といいます。

2）離散的確率変数 X に対して $P(X)\geqq 0$ で，$P(X)$ の総和が1である関数を確率関数といいます。X が連続的確率変数の場合は確率密度関数と呼ばれています。

x回目に初めて成功する確率 $(1-p)^{x-1}p$ は幾何分布

幾何分布のグラフ



注1）超幾何分布の平均値は $E(X)=\frac{1}{p}$，分散は $V(X)=\frac{1-p}{p^2}$

2）超幾何分布というものがあるが，幾何分布とは直接関係はない。

§67 希な現象はポアソン分布

ここが要点！ 交通事故の件数や機械の故障回数，電話の呼び出し数など，起こることが希な現象が一定時間内に生起する回数の確率分布は，ポアソン分布をとります。

やさしい解説

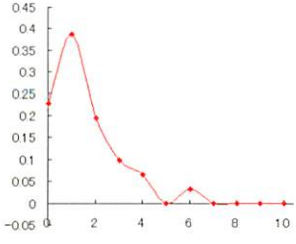
下表は平成15年度の8月における北海道全体の交通事故の死亡者数の生起日数です。

死亡者数	日数
0	7
1	12
2	6
3	3
4	2
5	0
6	1
7	0
8	0
9	0
10	0

死亡者数の平均 m は，

$$0\times\frac{7}{31}+1\times\frac{12}{31}+\cdots+6\times\frac{1}{31}+\cdots\cdots+10\times\frac{0}{31}$$

より，約1.52となります。



この表をもとに**相対度数分布グラフ**を描いたのが前段の下図です。

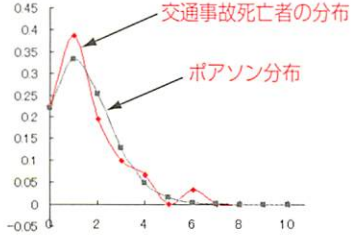
交通事故死は頻繁に起きる現象ではありません。そのため，この分布はポアソン分布という分布に従うことになります。

ポアソン分布とは，「確率変数 x の確率関数が，

$$P(x)=\frac{\lambda^x}{x!}e^{-\lambda}\quad(x=0,\ 1,\ 2,\ \cdots\cdots)$$

と表される分布」です。λ と e は定数で，e はネイピアの数2.71828……です。この分布の平均と分散は，ともに λ となります。

平均は λ なので，λ として先の交通事故の死亡者数の平均1.52を代入し，確率関数のグラフを描いてみましょう。それが下図です。

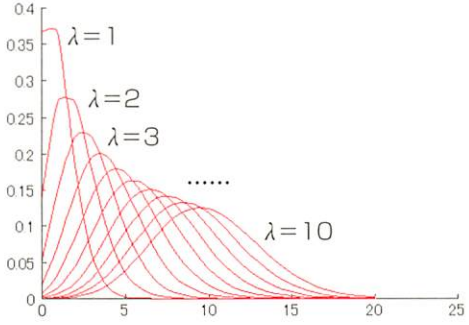


まさに，交通事故の死亡者はポアソン分布と見なせます。

ポアソン分布は平均値λのみで決まる

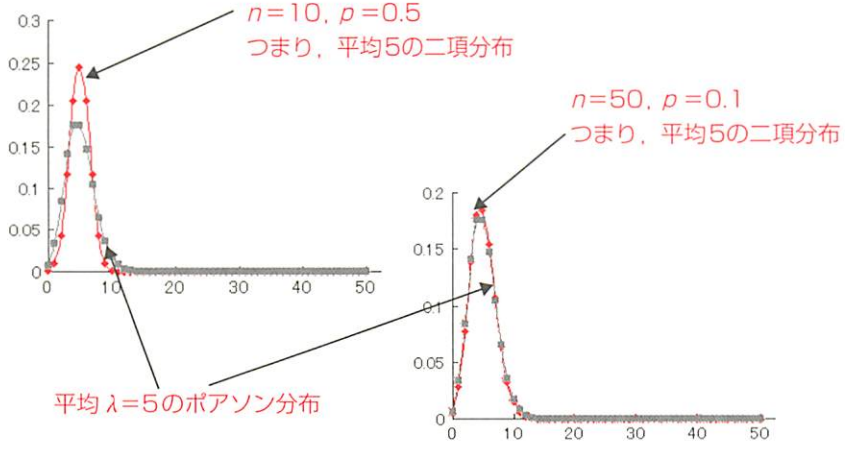
1 「希に起こる現象」の確率分布

平均値λさえわかれば，「希に起こる現象」の確率分布がわかる。



2 ポアソン分布は二項分布の極限分布

ポアソン分布は二項分布$B(n,\ p)$のpを小さくし，nを大きくしたときの極限分布 ---▸ **ポアソン分布は起こることが希な現象によく似合う。**

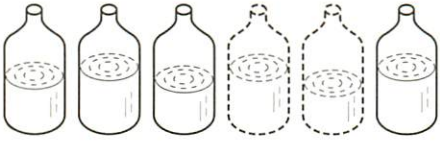


§68 誤差の研究から正規分布

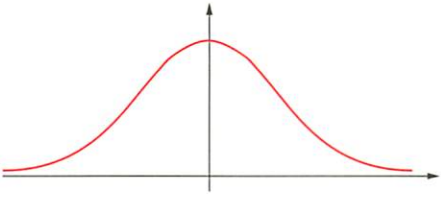
ここが要点！ 19世紀に，ドイツの数学者であるガウス（Gauss：1777～1855）が，天体観測のバラツキから正規分布という確率分布を発見しました。彼は正規分布は誤差の分布法則であると考えましたが，その後の研究で多くの確率分布の原型であることがわかりました。つまり，正規分布はいろいろな確率分布の女王なのです。

やさしい解説

たとえば，ある工場で機械を使って水200ccをボトルに詰めるとします。ところが，どんな優秀な装置を使ってもピッタリ200ccの水をボトルに詰めることは困難なのです。実際には200ccより，多かったり少なかったりまちまちです。



このとき，ボトルに詰めた水の量と200ccとの差が，いわゆる誤差になります。水の詰められたたくさんのボトルについて，誤差を求めて誤差の分布グラフを作ると，下図のように左右対称な山型の分布になります。



この確率分布がガウスの見つけた正規分布なのです。この分布を式を使って，もう少し正確に述べると次のようになります。

正規分布は，連続分布の一種で確率密度関数が，

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

となるものです。

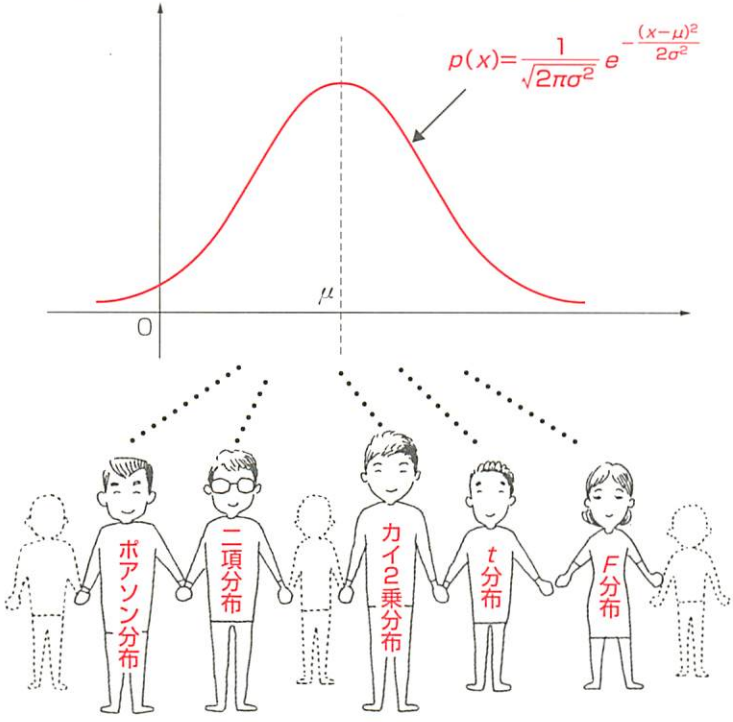
じつに凄い式です。ここで，πは円周率3.14159…，eはネイピアの数2.71828…です。また，σとμは定数で，μは正規分布の平均，σは標準偏差になります。

正規分布は，このμとσによって完全に特徴づけられた分布なので，この分布を記号で$N(\mu, \sigma^2)$と書きます。

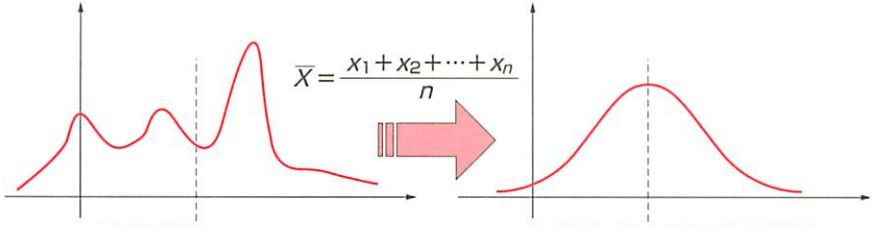
この正規分布は，いろいろな確率分布と密接な関係を持っています。また，次のセクション（§69）で紹介する中心極限定定理における正規分布の役割はきわめて重要です。

正規分布は確率分布の女王

1 正規分布 $N(\mu, \sigma^2)$ のグラフ



2 もとの分布が何でも標本平均の分布は正規分布……中心極限定理



母集団分布

平均 m，標準偏差 σ

標本平均 $\overline{X}$ の分布は正規分布

平均 m，標準偏差 $\frac{\sigma}{\sqrt{n}}$

§69 「正規分布」作成マシーン

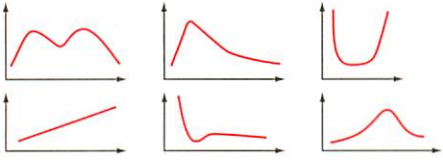
ここが要点！ 左右対称な山型の美しい正規分布は分布の女王です。この正規分布と世の中のすべての分布は他人ではありません。「母集団がどんな分布であっても，そこから抽出して得られる標本平均の分布は，標本が大きくなれば正規分布に従う」という中心極限定理によって仲のいい兄弟になります。

（各科目の得点分布は正規分布でなくても，英，国，数，理，社の平均点の分布は正規分布に近い!!）

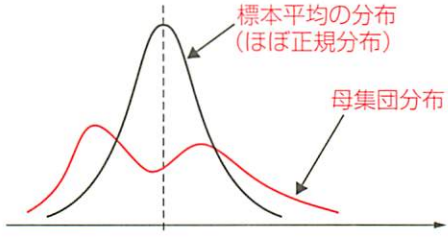
やさしい解説

身長や体重，人間の寿命など世の中には，たくさんの統計資料が存在し，それらの分布はじつにいろいろです。

【分布の形はいろいろ】



しかし，これらの母集団から抽出して得られる標本平均の分布となると，そこには共通の性質があります。それは，「母集団から得た標本平均の分布は，標本の大きさが大きければ，ほぼ正規分布になる」ということです。このとき，標本平均の平均は母集団の平均と同じで，分散はもとの母集団の分散を標本の大きさで割った値となります。したがって，標本の大きさが大きいほど標本平均の分散は小さくなり，分布は平均値の周りに集まってくることになります。

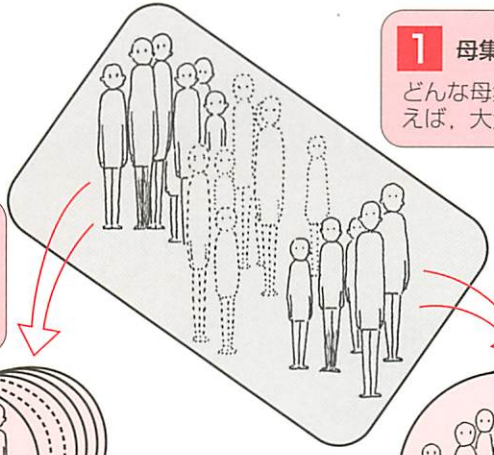


標本平均に関するこの性質は，**中心極限定理**と呼ばれています。

たとえば，ここに1万人の分の身長データがあり，平均値は165，分散は15であるとしましょう。ここから大きさ50の標本をとり，これらの身長の平均，つまり標本平均を考えてみます。

この標本平均は，標本を抽出するたびに変動します。しかし，中心極限定理によると，この標本平均の分布は，平均が165で，分散が15/50の正規分布に近い分布になります。

もとがどんな分布でも標本平均の分布は正規分布



1 母集団がある

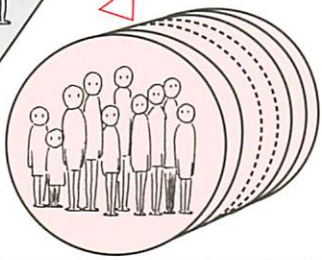
どんな母集団でもよい。たとえば，大量の身長データ

2

たとえば，大きさ5の標本をたくさん抽出

4

たとえば，大きさ10の標本をたくさん抽出

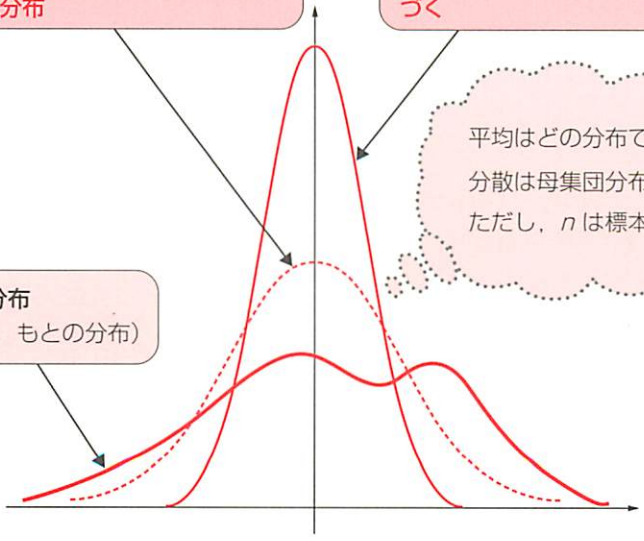


3

大きさ5の標本の標本平均（ここでは5人の平均身長，標本を抽出するたびに変動する）の分布を描いてみる。
ほぼ正規分布

5

大きさ10の標本の標本平均の分布を描いてみる。大きさが5の場合より平均付近に集中した**正規分布に近づく**



平均はどの分布でも同じ。
分散は母集団分布の $\frac{1}{n}$ 倍。
ただし，n は標本の大きさ

母集団分布
（つまり，もとの分布）

§70 標本分散の分布はχ^2分布

ここが要点！ χ^2（カイ２乗）分布は正規分布から導かれる分布で，簡単にいえば「標本分散に関する分布」です。この分布は利用価値が高く，とくに「適合度検定」や「独立性の検定」などのχ^2検定で頻繁に利用されます。

やさしい解説

母集団は，分散がσ^2の正規分布をしているものとします。平均は何でもかまいません。また，この母集団からランダムに抽出した大きさnの標本を，

$\{x_1, x_2, x_3, \cdots, x_n\}$

とします。ここで，χ^2（「カイ２乗」と読む）という，次の統計量に着目してみます。

$$\chi^2=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{\sigma^2} \quad \cdots\cdots\cdots\cdots①$$

ただし，$\overline{x}$は標本平均，つまり，

$$x=\frac{x_1+x_2+\cdots+x_n}{n}$$

です。すると，このχ^2は自由度$f=n-1$のカイ２乗（χ^2）分布と呼ばれる分布に従うことが知られています。

自由度fのχ^2分布とは，「確率変数Xの確率密度関数が次式で与えられる分布」をいいます。

$$g_f(x)=kx^{\frac{f}{2}-1}e^{-\frac{x}{2}}$$

（$x\geqq 0$，kは定数，eはネイピアの数）

この分布は「自由度fのみによって決まり，平均はfで分散は$2f$」となります。なお，式①は標本分散の

$$s^2=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}$$

を用いると，

$$\chi^2=\frac{(n-1)s^2}{\sigma^2}=\frac{(n-1)\text{標本分散}}{\sigma^2} \quad \cdots\cdots\cdots②$$

と書けます。

したがって，χ^2（カイ２乗）分布とは，標本分散の分布ということができます。この分布は母平均と関係ありません。つまり，母平均が何であっても式②は自由度fのχ^2分布をすることになります。

注）χ^2（カイ２乗）分布の確率密度関数を詳しく表現すると，

$$g_f(x)=\frac{x^{\frac{f}{2}-1}}{2^{\frac{f}{2}}\Gamma\left(\frac{f}{2}\right)}e^{-\frac{x}{2}}$$

となります。ここで，$\Gamma(p)$はガンマ関数の$\Gamma(p)=\int_0^\infty x^{p-1}e^{-x}dx$　のことです。こういう難しいことを理解する必要はありません。

χ^2（カイ２乗）分布は標本分散の分布

1 χ^2（カイ２乗）分布は自由度 f のみで決まり，母平均には無関係

平均■，分散 σ_1^2 の正規母集団

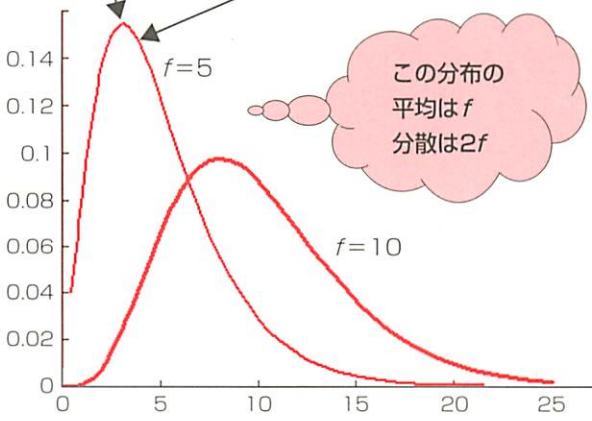
大きさ6の標本 → 算出

$$\chi^2 = \frac{(6-1)\text{標本分散}}{\sigma_1^2}$$ の分布

平均●，分散 σ_2^2 の正規母集団

大きさ6の標本 → 算出

$$\chi^2 = \frac{(6-1)\text{標本分散}}{\sigma_2^2}$$ の分布



2 χ^2（カイ２乗）検定は χ^2（カイ２乗）分布を利用した検定

「適合度検定」，「独立性の検定」，「母比率の差の検定」……いろいろある

§71 分散比の分布はF分布

ここが要点！ F分布は正規分布から導かれる分布で，簡単にいえば「標本分散の比の分布」です。この分布を利用した検定をF検定といいます。例として，分散分析における検定や，２組の観測資料の分散の比に関する検定などがあります。

やさしい解説

平均がμ_1，分散が${\sigma_1}^2$の正規母集団から抽出した大きさmの標本を，

$\{x_1, x_2, x_3, \cdots, x_m\}$

とし，これから算出した標本分散を${s_1}^2$とします。つまり，

$${s_1}^2=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_m-\overline{x})^2}{m-1}$$

とします。

また，平均がμ_2，分散が${\sigma_2}^2$の正規母集団から抽出した大きさnの標本を$\{x_1, x_2, x_3, \cdots, x_n\}$とし，これから算出した標本分散を${s_2}^2$とします。つまり，

$${s_2}^2=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}$$

とします。このとき，

$$F=\frac{{s_1}^2{\sigma_2}^2}{{s_2}^2{\sigma_1}^2}$$

は，F分布と呼ばれる分布に従うことが知られています。

それでは，F分布とはどんな分布なのでしょうか。

確率変数Xの確率密度関数が次式で与えられる分布を**自由度 m，n のF分布**といいます。

$$\begin{cases} x>0 \text{ のとき } f(x)=\dfrac{kx^{\frac{m}{2}-1}}{\left\{1+\left(\dfrac{m}{n}\right)x\right\}^{\frac{m+n}{2}}} \\ x\leqq 0 \text{ のとき } f(x)=0 \end{cases}$$

$$\text{平均は } E(X)=\frac{n}{n-2}$$

$$\text{分散は } V(X)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

となります。

注）F分布の確率密度関数を詳しく表現すると，次のようになります。

$$f(x)=\frac{\Gamma\left(\dfrac{m+n}{2}\right)\left(\dfrac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)\left\{\left(\dfrac{m}{n}\right)x+1\right\}^{\frac{m+n}{2}}}$$

となります。

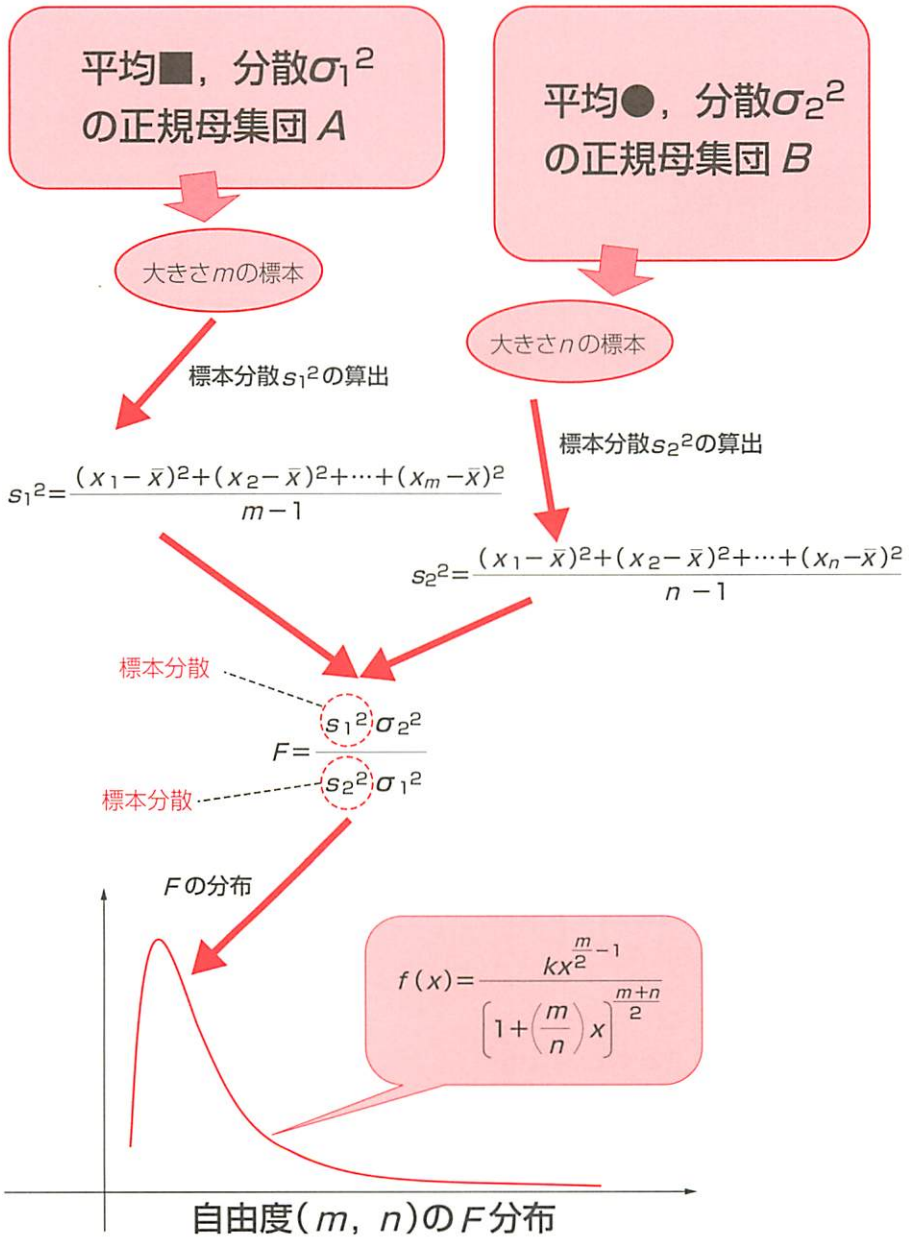
ここで，$\Gamma(p)$ はガンマ関数

$$\Gamma(p)=\int_0^\infty x^{p-1}e^{-x}dx$$

のことです。なお，こういう難しいことを理解する必要はありません。

F分布は標本分散の比の分布

F分布は自由度m，nのみで決まり，母平均には無関係



§72 標本平均の分布は t 分布

ここが要点！ 標本平均を論じるとき，母集団の分散（母分散）がわかっているのか，それとも未知なのかは議論の分かれ目です。t 分布は，母分散が未知のときの標本平均に関する分布なのです。

やさしい解説

平均が μ，分散が σ^2 の正規母集団からランダムに抽出した大きさ n の標本 $\{x_1, x_2, x_3, \cdots, x_n\}$ から算出した標本平均を $\overline{x}$，標本分散を s^2 とします。つまり，

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n}$$

$$s^2=\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}$$

このとき，標本平均 $\overline{x}$ に関して，次の2つのことが成立します。

(1) $\dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ は平均0，分散1の正規分布に従う。

(2) $\dfrac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ は，自由度 $f=n-1$ の t 分布に従う。

標本調査では，一般に母分散は未知です。したがって，(2) の理論は重要です。

正規分布については§68を参照してもらうことにして，t 分布とはどのような分布かを調べてみましょう。

確率変数 X の確率密度関数 $g_f(x)$ が次式で与えられる分布を**自由度 f の t 分布**（または，**ステューデント分布**）といいます。

$$g_f(x)=k\left(1+\frac{x^2}{f}\right)^{-\frac{f+1}{2}} \quad (k\text{は定数})$$

平均は $E(X)=0$

分散は $V(X)=\dfrac{f}{f-2}$

となります。

参考 t 分布の確率密度関数を詳しく表現すると，

$$g_f(x)=\frac{1}{\sqrt{f\pi}}\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}\left(1+\frac{x^2}{f}\right)^{-\frac{f+1}{2}}$$

となります。

ここで $\Gamma(p)$ はガンマ関数

$$\Gamma(p)=\int_0^\infty x^{p-1}e^{-x}dx$$

のことです。難しいですね。この関数そのものを理解する必要はありません。

母分散が未知ならば t 分布

1 母分散か，標本分散かが分布の分かれ目

平均 μ，分散 σ^2 の正規母集団

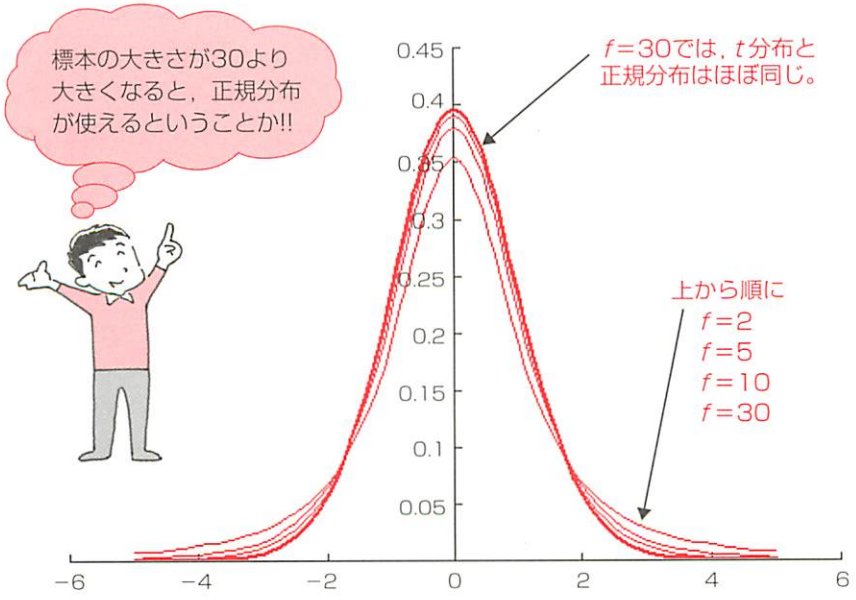
大きさの n 標本

$\dfrac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ の分布は正規分布

$\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$ の分布は t 分布

ただし，s は標本標準偏差 $\sqrt{\text{標本分散}}$

2 自由度が 30 を超すと t 分布と正規分布はほぼ同じ



§73 分布の非対称性や尖り具合

ここが要点！ 分布の非対称性を調べる指標として歪度(わいど)があります。また，分布の裾(すそ)の長短を調べる指標として尖度(せんど)があります。いずれも，それらの値が正か負かによって非対称性や裾の長短を判定できる優れものです。

やさしい解説

歪度(わいど)は，次式で定義されています。つまり，観測値x_1，x_2，…，x_nの分布の歪度をα_3とすると，

$$\alpha_3=\frac{1}{n}\left\{\left(\frac{x_1-\overline{x}}{s}\right)^3+\left(\frac{x_2-\overline{x}}{s}\right)^3+\cdots+\left(\frac{x_n-\overline{x}}{s}\right)^3\right\} \quad \cdots\cdots①$$

ただし，$\overline{x}$は平均値，sは標準偏差です。これは，「個々の観測値を標準化した値を3乗して平均値との違いをきわだたせ，それらの平均をとったもの」です。

したがって，分布の裾が右に伸びているほど歪度α_3の値は大きな正の値となります。分布の裾が左に伸びているほど，歪度α_3の値は絶対値の大きな負の値となります。

なお，確率変数Xの確率分布の歪度α_3は，次式で定義されます。

$$\alpha_3=\frac{E\{(X-\mu)^3\}}{\sigma^3} \quad \cdots\cdots②$$

ただし，μはXの期待値，σは標準偏差です。式②による歪度α_3の表す分布の特徴は，観測値の場合と同じです。

尖度(せんど)は，次式で定義されています。つまり，観測値x_1，x_2，…，x_nの分布の尖度をα_4とすると，

$$\alpha_4=\frac{1}{n}\left\{\left(\frac{x_1-\overline{x}}{s}\right)^4+\left(\frac{x_2-\overline{x}}{s}\right)^4+\cdots+\left(\frac{x_n-\overline{x}}{s}\right)^4\right\} \quad \cdots\cdots③$$

ただし，$\overline{x}$，sは式①と同じです。

式③は，「個々の観測値を標準化した値を4乗し，平均値との違いをきわだたせ，それらの平均をとったもの」です。したがって，分布の裾が短いほど尖度α_4の値は小さくなり，分布の裾が長いほど尖度α_4の値は大きくなります。

なお，確率変数Xの確率分布の尖度α_4は，次式で定義されます。

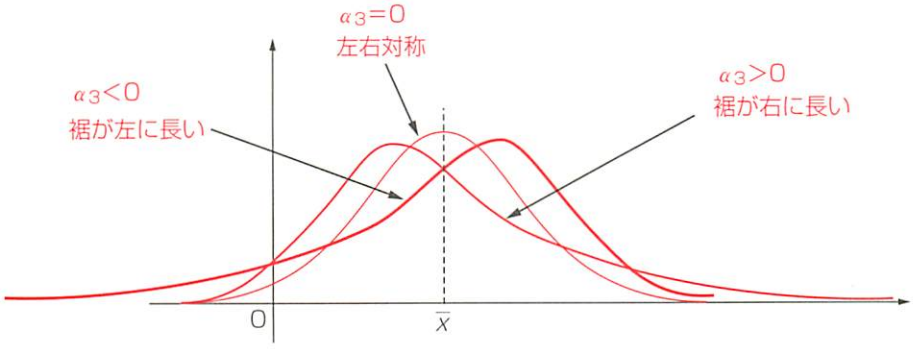
$$\alpha_4=\frac{E\{(X-\mu)^4\}}{\sigma^4} \quad \cdots\cdots④$$

ただし，μ，σは式②と同じで，μはXの期待値，σは標準偏差です。

分布の非対称性は歪度α_3で，分布の広がり具合は尖度α_4

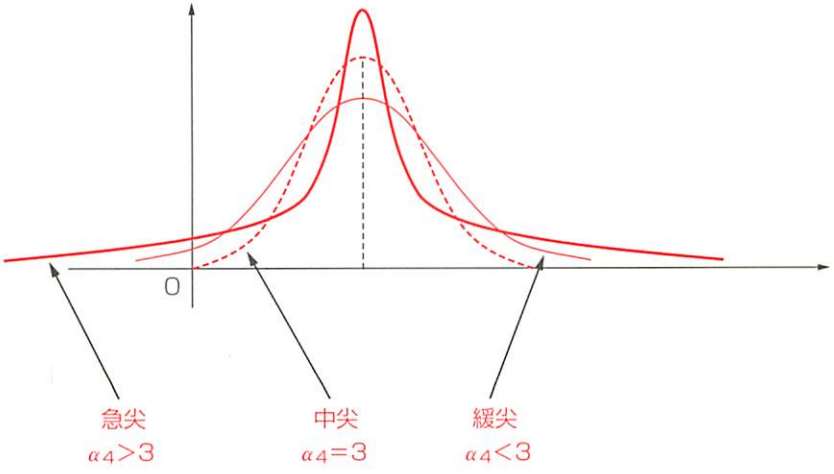
1 歪　度

$$\alpha_3=\frac{1}{n}\left\{\left(\frac{x_1-\bar{x}}{s}\right)^3+\left(\frac{x_2-\bar{x}}{s}\right)^3+\cdots+\left(\frac{x_n-\bar{x}}{s}\right)^3\right\}$$



2 尖　度

$$\alpha_4=\frac{1}{n}\left\{\left(\frac{x_1-\bar{x}}{s}\right)^4+\left(\frac{x_2-x}{s}\right)^4+\cdots+\left(\frac{x_n-\bar{x}}{s}\right)^4\right\}$$



§74 チェビシェフは万能

ここが要点！

平均が m，分散が σ^2（標準偏差 σ）である確率変数 X について，

「$m-k\sigma \leqq X \leqq m+k\sigma$ である確率は $1-\frac{1}{k^2}$ 以上」

ただし，k は正の数。このことをチェビシェフの不等式といいます。平均が m，分散が σ^2 でありさえすれば，どんな確率変数でも成り立ちます。

やさしい解説

確率変数の分布には二項分布や正規分布，t 分布，……などいろいろあります。分布が違えば確率変数のとる値も違ってきますが，どんな分布でも表題の不等式が成立するのです。

（例）平均が 50，分散が 5^2 である確率変数 X について，$k=2$ の場合はチェビシェフの不等式より，

$$50-2\times5 \leqq X \leqq 50+2\times5$$

つまり，

$$40 \leqq X \leqq 60$$

を満たす確率は，

$$1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}=0.75$$

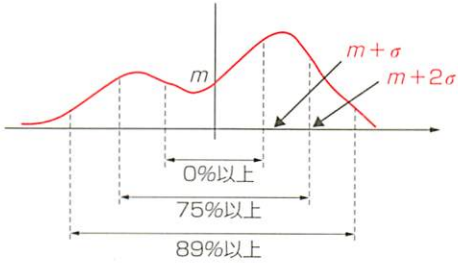
から，0.75 以上ということになります。

このチェビシェフの不等式は，確率変数 X の分布がどんな分布でも成り立つすばらしいものです。

しかし，そのために，不等式の成り立つ確率の条件はかなりあまくなっています。このことは，正規分布と比較するとよくわかります。

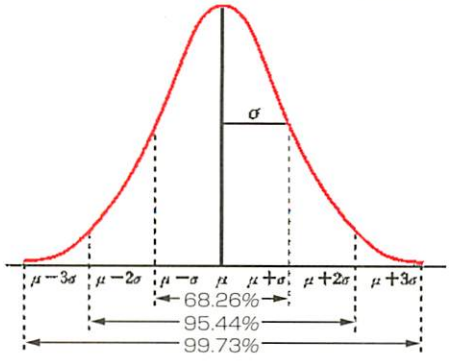
〔どんな分布でも成り立つ〕

……チェビシェフの不等式より

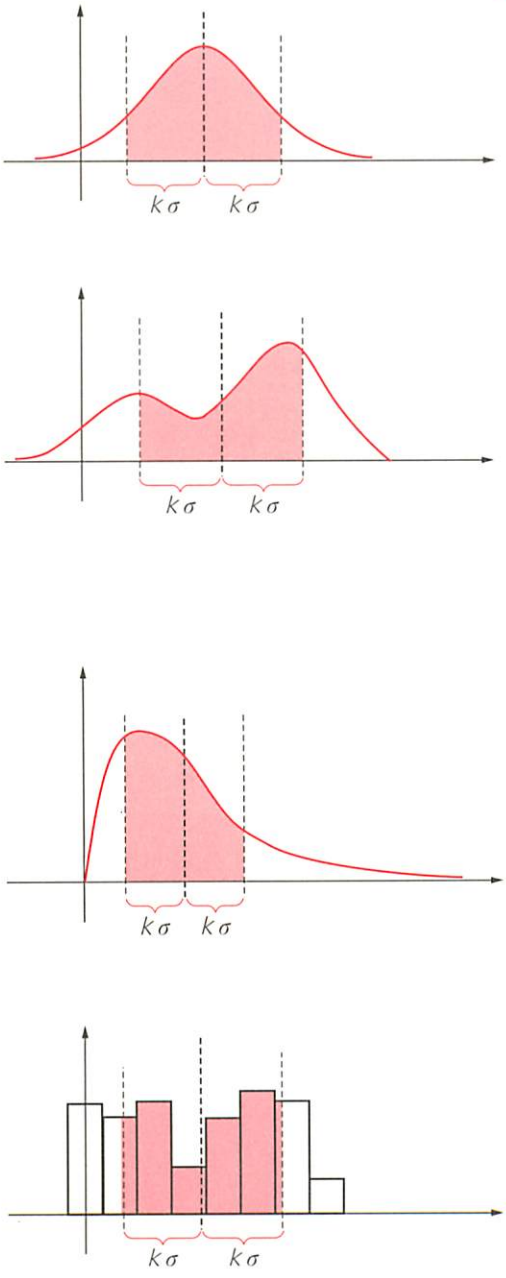


〔正規分布のとき成り立つ〕

……正規分布の性質より



チェビシェフの不等式はどんな確率分布でも成り立つ



確率変数 X が，
$m-k\sigma \leqq X \leqq m+k\sigma$
である確率（うす色部分の面積）は，

$1-\dfrac{1}{k^2}$ 以上。

ただし，m は平均，
σ は標準偏差

§75 変数なのに自由度

ここが要点！ 見かけ上，変数が n 個あるからといって，それらが独立に n 通りの値を自由にとれるとは限りません。そのため，変数が実際に自由に動ける度合いを表すものとして自由度が利用されます。

やさしい解説

母集団から復元抽出で抽出した大きさ n の標本の観測値を $\{x_1, x_2, \cdots, x_n\}$ とすると，$x_1, x_2, \cdots, x_n$ は互いに母集団の自由な値をとることができます。したがって，$\{x_1, x_2, \cdots, x_n\}$ の**自由度**は n だと考えられます。

次に，$y_i = x_i - \overline{x}$ $(i=1, 2, \cdots, n)$ とした n 個の変数 $\{y_1, y_2, \cdots, y_n\}$ を考えてみます。

ここで $\overline{x}$ は標本平均で，以下の式で与えられます。

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \cdots\cdots ①$$

この式①を変形すると，

$(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) = 0$

となり，

$y_1 + y_2 + \cdots + y_n = 0$

が成立します。

これは，n 個の変数 $\{y_1, y_2, \cdots, y_n\}$ についてはその間に1つの制約が生じ，互いに自由に n 個の値をとることができないことを意味しています。したがって，自由度は n より小さく（この場合 $n-1$）なります。

具体例でいえば，3個のデータがあり，そのうち1番目と2番目のデータが3と4で，平均が5であれば3番目のデータは8と決まってしまうということです。8以外の値はとりようがないのです。したがって，3個のデータの自由度は2になります。

この自由度は不偏分散（§61参照）の分母にも関係しています。標本 $\{x_1, x_2, \cdots, x_n\}$ の不偏分散は，

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

です。ここで，$\overline{x}$ は標本平均で，

$(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) = 0$

です。

本来，大きさ n の標本は n 個の情報を持ちますが，この条件がついた分，$(x_1 - \overline{x})$，$(x_2 - \overline{x})$，…，$(x_n - \overline{x})$ の動ける範囲は小さくなってしまいます（自由度は $n-1$）。それだけ，不偏分散 s^2 の分子の値は小さくなります。その小さくなった分子を n で割ると，分散は本来の値よりも小さく求められることになります。したがって，補正して $n-1$ で割ることになります。

見かけの自由より本当の自由

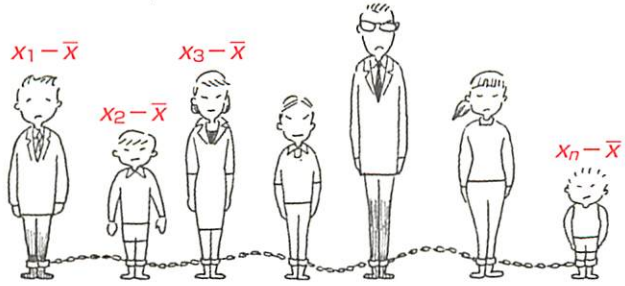
母集団

大きさ n の標本

$x_1, x_2, \cdots, x_n$

僕たち，自由に動けるぞ!!（自由度 n）

ところが



僕たち，勝手に動けないぞ!!（自由度 $n-1$）

関係 $(x_1-\overline{x})+(x_2-\overline{x})+\cdots+(x_n-\overline{x})=0$ で結ばれている。

ただし，$\overline{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$

§76 標準化によって同じスケール

ここが要点！ 確率変数を変換することによって，その分布の平均や標準偏差（分散）を特定の値にすることを標準化（または，基準化）といいます。この結果，同一の基準で判断したり，同一の表が使えるようになります。

やさしい解説

確率変数をある変換式で他の変数に変換することにより，もとの確率変数の平均や分散などを任意の値に変化させることができます。

ここに，確率変数Xがあり，その平均値はmで分散はσ^2であるとしましょう。このとき，この確率変数Xを次の変換式で確率変数Zに変換してみましょう。

$$Z=\frac{X-m}{\sigma} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots①$$

式中のσは標準偏差です。

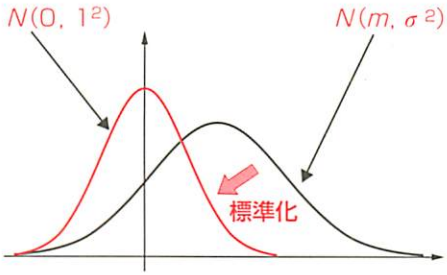
すると，確率変数Zの平均値は0で分散は1になってしまいます。つまり，「確率変数を変換することによって，分布の平均や標準偏差を特定の値0と1に変えて」しまいました。これが**標準化**（または，**基準化**）です。

注）通常，確率変数を変換することによって分布の平均を0，標準偏差を1にすることを標準化（または基準化）といいます。

たとえば，確率変数Xの分布が正規分布$N(m,\ \sigma^2)$であるとき，確率変数

$$Z=\frac{X-m}{\sigma}$$

の分布は平均値が0で，標準偏差が1の**標準正規分布**$N(0,\ 1^2)$になります。



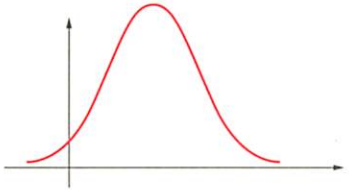
標準化することにより標準正規分布表が使えるようになり便利です。

注）最近ではコンピュータが普及したため，標準化しないでもいろいろな確率計算ができるようになりました。

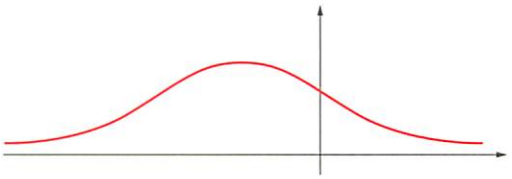
標準化されたデータについては，単位や尺度の違いをそろえることができ，比較が容易になるなどのメリットがたくさんあります。

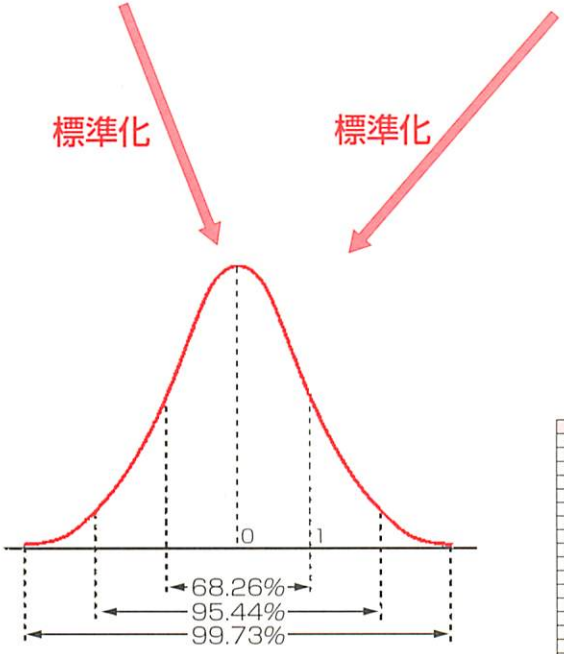
標準化で1つの世界に

平均値が180，分散20
の正規分布



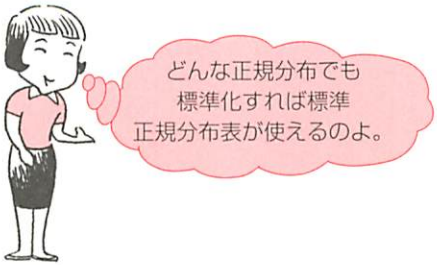
平均値が−45，分散40
の正規分布





標準正規分布表

x	0.00	0.01	0.02	0.03	0.04	0.05	
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4945	0.4
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4



注）平均値が０で，分散が１の正規分布を標準正規分布という。

§77 偏差値も標準化の一種

ここが要点！ 確率変数を変換して分布の平均を50，標準偏差を10としたときの確率変数の値を偏差値といいます。もとはアメリカ軍の着弾性能の指標として用いられていたものですが，日本では受験業界などで頻繁に使われています。

やさしい解説

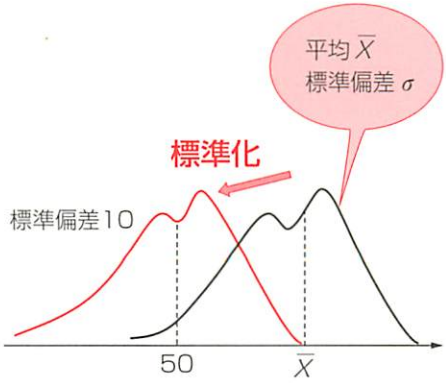
確率変数Xの平均値を$\overline{X}$，標準偏差をσとします。この$\overline{X}$とσを用いて，確率変数Xを次の変換式で確率変数Tに変換してみます。

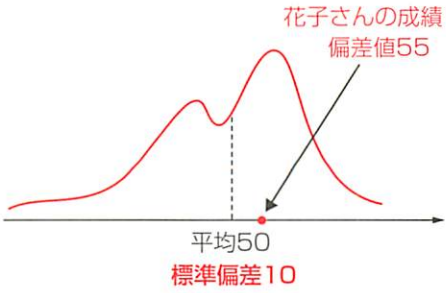
$$T=\frac{X-\overline{X}}{\sigma}\times 10+50 \cdots\cdots\cdots\cdots①$$

すると，確率変数Tの平均値は50で標準偏差は10になります。このとき，もとの確率変数Xの値に対して式①で求めたTの値を**偏差値**（**Tスコア**）といいます。つまり，「偏差値に変換することによって平均が50，標準偏差が10の確率分布という共通の土俵で議論ができる」ようになります。

たとえば，花子さんの英語の試験の得点が80だといわれても，花子さんの成績が他の人と比較していいのかどうかわかりません。みんなが100点近くとっているのであれば，花子さんを誉めるわけにはいきません。ところが，みんなが30点そこそこであれば，誉めても誉め足りません。このように，素点だけでは情報不足なのです。

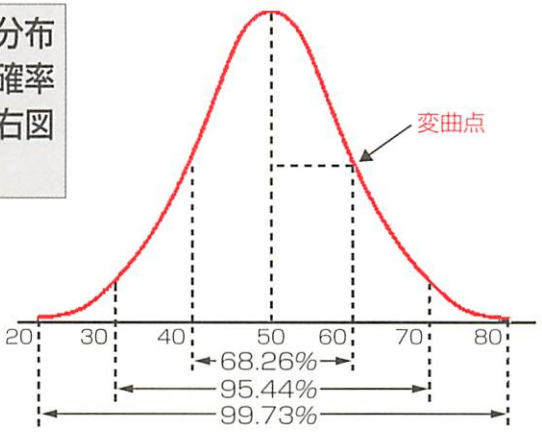
ところが，花子さんの英語の試験の偏差値は55点だといわれれば，「平均よりややいいが，抜群に優れているわけではない」と判定できます。このように偏差値がわかれば，全変量に対するその値の位置関係がわかってしまうのです。





偏差値はいろいろな情報を含んでいる

1 もとの確率分布が正規分布に近ければ，偏差値の確率分布もほぼ正規分布で右図のようになる



2 偏差値は 100 を超えたり，負の数になることがある

	データ	偏差値		データ	偏差値
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	100	51.85695		1	48.14305
	1	−3.85165		100	103.8516
平均＝	96.7	50		4.3	50
標準偏差＝	17.77104	10		17.77104	10

1が29個，100が1個の場合，100の偏差値は103.8…

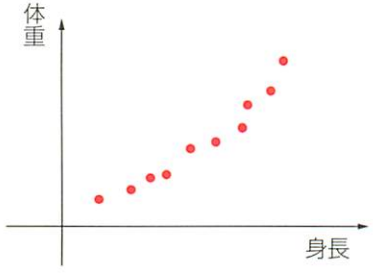
100が29個，1が1個の場合，1の偏差値は負の数 −3.8…

§78 相関は原因結果ではありません

ここが要点！ 「身長と体重」の間には強い正の相関があります。なぜならば，身長の3乗に比例して体重が決まるからです。したがって，身長が高ければ体重は重くなります。このようなことから，相関関係は因果関係を表していると考える人がいます。しかし，相関関係は一方の大小と他方の大小が関係しているのかを示したものにすぎず，原因結果の関係は表していません。

やさしい解説

統計学では，相関係数の値が正の数で1に近ければ**正の相関**があるといい，−1に近ければ**負の相関**，0に近ければ**無相関**と判断します。ここで疑問は，正の相関が強いときには変量のうち，一方が原因で他方が結果だと考えていいのかということです。



そこで，相関係数γの定義をもう一度調べてみましょう。

2つの変量$\{x_1, x_2, \cdots, x_n\}$，$\{y_1, y_2, \cdots, y_n\}$に対して**相関係数$\gamma$**は，

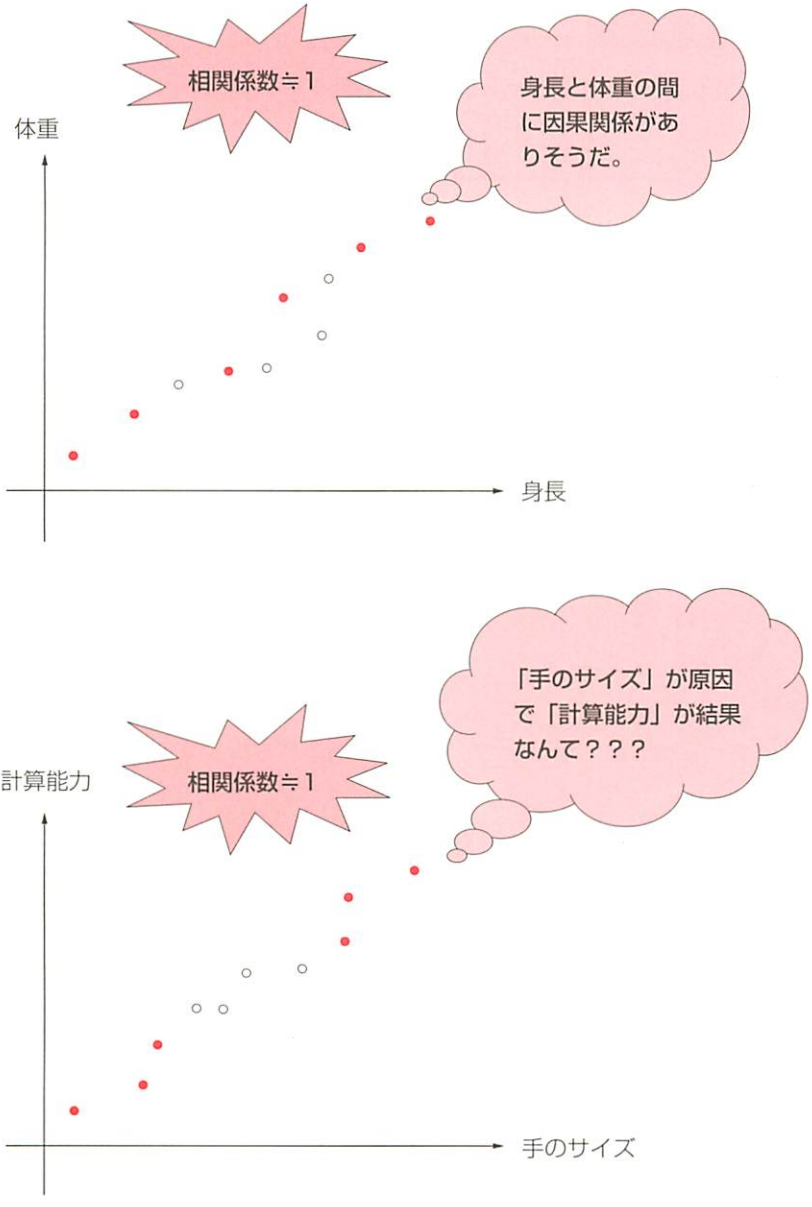
$$\frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\sqrt{(y_1-\bar{y})^2+\cdots+(y_n-\bar{y})^2}}$$

で定義されています。

この式を見ると，次のことがわかります。もし，一方が原因で他方が結果という関係があって，一方が増えれば他方も増えているのであれば，相関係数γの値は大きくなることがわかります。身長と体重の場合がまさしくそうです。

ところが，手のサイズと計算能力はどうでしょうか。この場合にも正の強い相関があります。しかし，これをもとに「手のサイズが計算能力の原因である」と主張する人はいないでしょう。計算能力を伸ばすには，コツコツ勉強するよりも指のストレッチ体操に励むことになります。年齢が両者に共通の要因となっていて，相関が高くなっているにすぎないのです。

相関は大小関係で因果関係ではない



§79 相関関係はベクトルと絡む

ここが要点！ 2つの変量の相互関係の大小を数値で表したものに相関係数というものがあります。この値は，−1以上＋1以下の値をとります。じつはこの相関係数，ベクトルで見るとベクトルの内積と関係があります。

やさしい解説

1つの変量$\{x_1, x_2, \cdots, x_n\}$に対して，分散$\nu_x$は次式で与えられます。

$$\nu_x = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}$$

ここで，$\overline{x}$は平均値です。

2つの変量$\{x_1, x_2, \cdots, x_n\}$，$\{y_1, y_2, \cdots, y_n\}$に対して，共分散$\nu_{xy}$は次式で与えられます。

$$\nu_{xy} = \frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n-1}$$

この共分散は，xとyがともに増加したり減少したりして同じ方向に動けば，ν_{xy}の値はプラスになり正の相関が強く出ます。xとyの動きが逆方向であれば，ν_{xy}の値はマイナスになり負の相関が強く出ます。ν_{xy}の値が0に近ければ相関は弱くなります。

このように，共分散は2つの変量の相関を知る上で便利ですが，変量によってはいくらでも大きな値や小さな値になります。そのため，たんに共分散の値を知っただけでは相関の強弱ははっきりしません。そこで，考えられたものが次式で定義される**相関係数**γというものです。

$$\gamma = \frac{\nu_{xy}}{\sqrt{\nu_x}\sqrt{\nu_y}} = \frac{(x_1-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\sqrt{(y_1-\overline{y})^2+\cdots+(y_n-\overline{y})^2}} \quad \cdots\cdots①$$

つまり共分散を標準偏差$\sqrt{\nu_x}\times\sqrt{\nu_y}$で割ったものです。この結果，

$$-1 \leqq \gamma \leqq 1$$

となります。ここで，

$\vec{p} = (x_1-\overline{x}, x_2-\overline{x}, \cdots x_n-\overline{x})$

$\vec{q} = (y_1-\overline{y}, y_2-\overline{y}, \cdots y_n-\overline{y})$

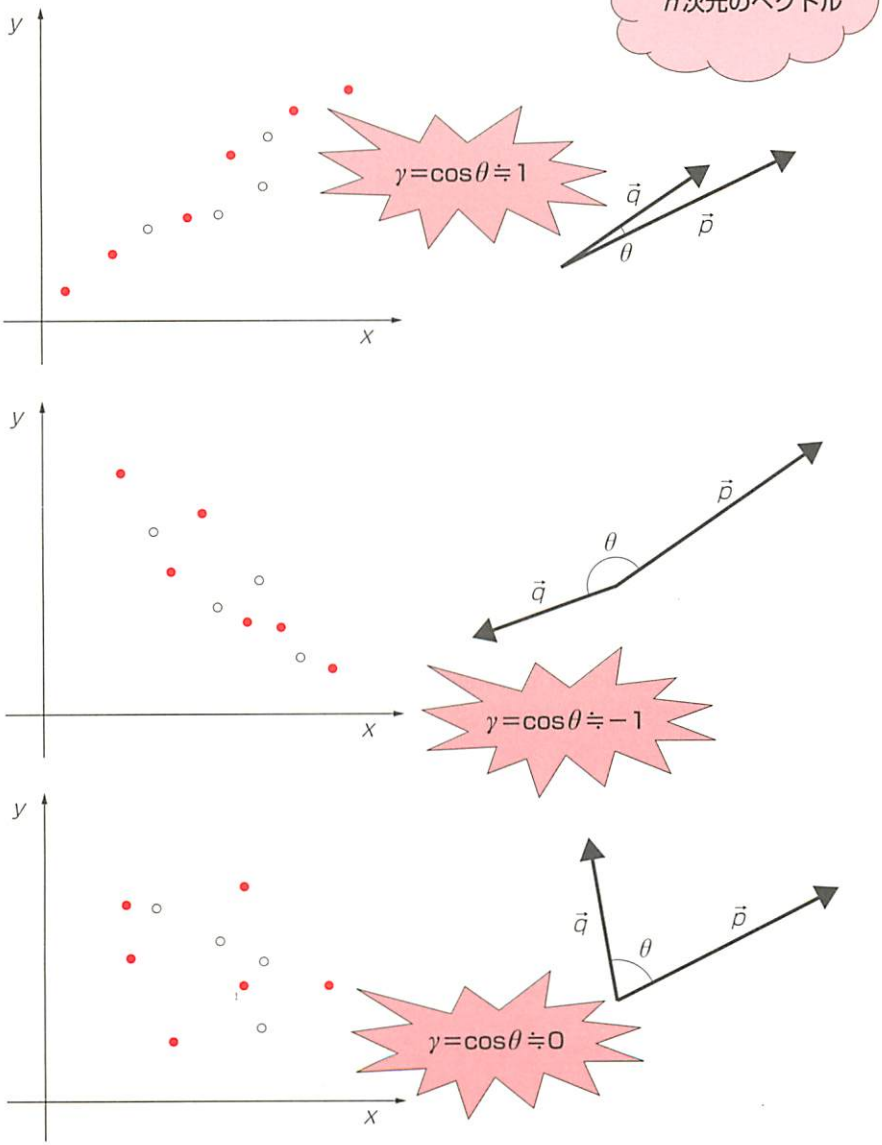
とすると，式①はベクトルの内積$\vec{p}\cdot\vec{q}$を用いて，次のように書けます。

$$\gamma = \frac{\vec{p}\cdot\vec{q}}{|\vec{p}||\vec{q}|}$$

なお，2つのベクトルのなす角をθとすると，$\vec{p}\cdot\vec{q} = |\vec{p}||\vec{q}|\cos\theta$より，相関係数$\gamma$は$\cos\theta$と書けます。

相関係数はベクトルと親密

2つの変量 $\{x_1, x_2, \cdots, x_n\}$ と $\{y_1, y_2, \cdots, y_n\}$ の相関図と相関係数 γ
それに2つのベクトル　$\vec{p}=(x_1-\bar{x},\ x_2-\bar{x},\ \cdots,\ x_n-\bar{x})$
$\vec{q}=(y_1-\bar{y},\ y_2-\bar{y},\ \cdots,\ y_n-\bar{y})$
の関係は下図のようになる。



§80 推定の妥当性には確率が

ここが要点！ 標本から得られた平均や比率などの統計量をもとに，母集団の平均や比率などの母数（パラメータ）を推測することを推定といいます。通常，推定というと，母数に幅を持たせて推定することになります。つまり，区間推定です。このとき，判断の正しさは確率で示されます。

やさしい解説

ほんの少人数の身長を測定しただけで，日本人全体の身長の平均を99％の確からしさで予測したり，選挙で投票用紙のほんの一部を開票しただけで当確を宣言したり，……など，あたかも「一を聞いて十を知る」ような神業は，すべて統計学の「区間推定」という考え方を用いています。少しお堅くいうと，**区間推定**とは未知の母集団から抽出された標本をもとに，母数（パラメータ，母集団の平均や比率など）がどのような範囲にあるのかを確率的に推測する方法なのです。これは，母集団と標本が確率関係によって結ばれているから可能になります。

たとえば，母平均mの区間推定は次の公式を使って行われます。

95％の信頼度の区間は，

$$\overline{X}-1.96\frac{s}{\sqrt{n}}\leqq m\leqq \overline{X}+1.96\frac{s}{\sqrt{n}} \quad \cdots\cdots①$$

99％の信頼度の区間は，

$$\overline{X}-2.58\frac{s}{\sqrt{n}}\leqq m\leqq \overline{X}+2.58\frac{s}{\sqrt{n}} \quad \cdots\cdots②$$

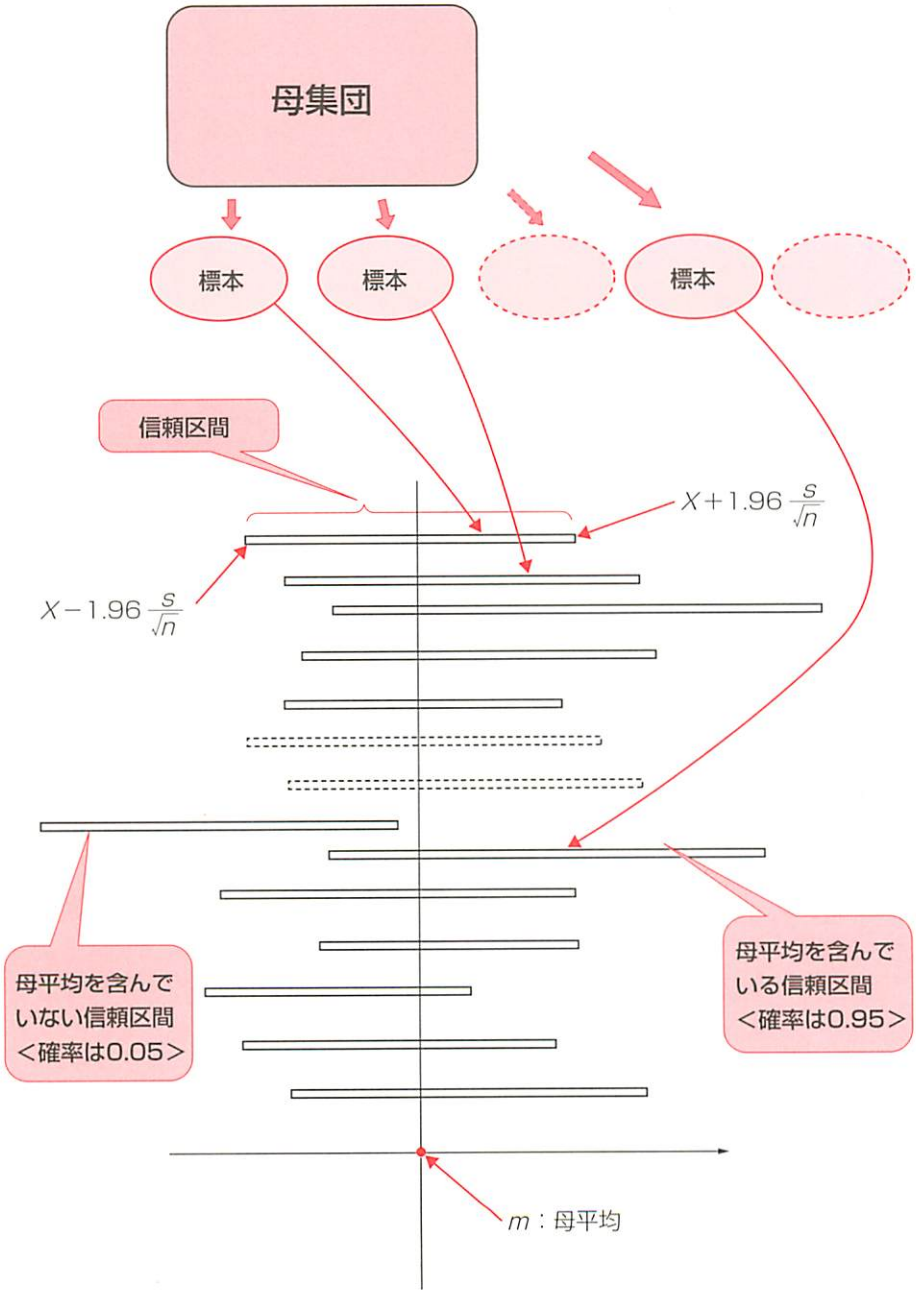
ただし，nは標本の大きさ，$\overline{X}$は標本平均，sは不偏分散から求めた標準偏差です。

この式が導かれる過程については，§82で紹介することにして，ここでは，式①，②を利用した区間推定の信頼度の意味について調べておきましょう。

抽出した標本から得た$\overline{X}$とsを式①に代入して区間を求めると，母平均mがその区間に存在している確率が0.95であることを式①は主張しているのです。

式②は99％の信頼度で母平均mが，その区間に存在していることを主張しているのです。ともに確率判断なのです。したがって，間違っている可能性もあります。式②の区間は式①の区間よりも幅が広いのですが，確信の度合いを高めようとしたために慎重になったのです。

95％の信頼区間に母平均が入っている確率は 0.95



§81 点推定と区間推定，どっちが偉い

ここが要点！ 母平均や母分散など母集団の特質を表す数値をパラメータ（母数）といいます。パラメータを標本から推定する方法に，点推定と区間推定の２種類があります。点推定はパラメータを１つの数値でズバリいい当てるものですが，確信の度合いを示すことができません。区間推定はパラメータが含まれている区間を推定するもので，ズバリとはいきません。しかし，確率95％であっているというように，確信の度合いを示すことができます。

やさしい解説

大きさnの標本$\{X_1, X_2, \cdots, X_n\}$をもとに，パラメータ$\theta$を推定するために用いられる関数や式を**推定量**といい，$\hat{}$（ハット）をつけて，

$\hat{\theta} = \theta(X_1, X_2, \cdots, X_n)$

と書き表します。推定量に標本値

$X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$

を代入すれば**推定値**

$\theta' = \theta(x_1, x_2, \cdots, x_n)$

を得ることになります。

(1) 点推定

点推定は，標本から母集団のパラメータを１つの数値でいい当てる推定です。

母平均に対する点推定量としては，次の相加平均が考えられます。

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \quad \cdots\cdots ①$$

この式①から得られる推定値が，どのくらい正しいかは何ともいえませんが，標本の数nがある程度大きければ，パラメータのよい近似値を得ることができます（右ページの図参照）。

(2) 区間点推定

区間点推定は，標本から母集団のパラメータを区間でいい当てる推定です。この場合は，推定がどのくらい正しいのかを確率で示すことができます。

たとえば，標本平均$\overline{X}$が，標本標準偏差sとすると，標本の大きさのnがある程度大きければ，母平均mは次のように区間推定されます。

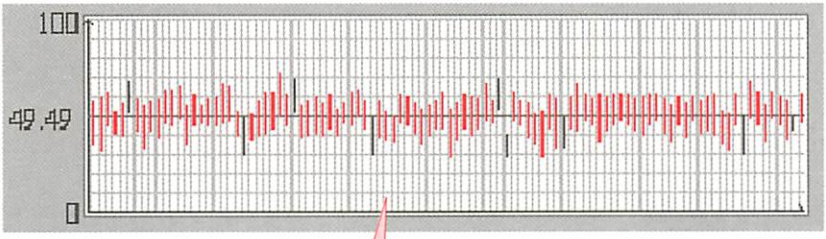
$$\overline{X} - 1.96\frac{s}{\sqrt{n}} \leqq m \leqq \overline{X} + 1.96\frac{s}{\sqrt{n}} \quad \cdots\cdots ②$$

ただし，このことが正しい確率は95％です。

点推定も簡単で素敵

1 区間推定の確率の意味

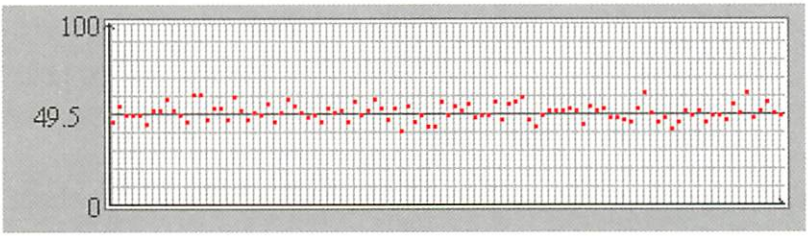
下のグラフは母集団（180,000 個の正規分布をなす 100 点満点の得点データ）から，大きさ 10 の標本を 100 回抽出し，そのたびごとに前ページ式②で求めた区間を縦にグラフで表示したもの。



母平均49.49を含む区間は95%ぐらいある。

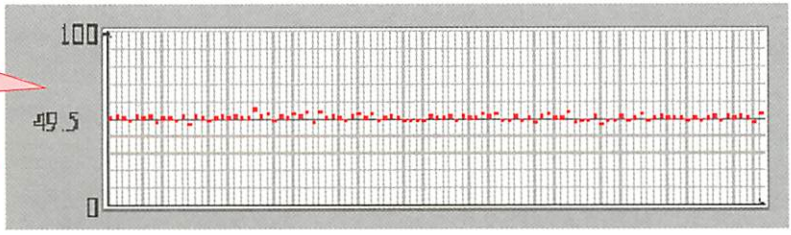
2 推定もなかなかいけるぞ

上記 1 と同じ母集団から大きさ 10 の標本を 100 回抽出し，そのたびごとに前ページ式①で求めた値をプロットしたものが下のグラフ。



上記 1 と同じ母集団から大きさ 100 の標本を 100 回抽出し，そのたびごとに前ページ式①で求めた値をプロットしたものが下のグラフ。

標本平均のほとんどが母平均49.5に近い値となる。



§82 資料が多いときの平均値の推定は

ここが要点！

標本の大きさを n，標本平均を $\overline{X}$，不偏分散から求めた標準偏差 s とすると，母集団の平均（母平均）m は次式で推定できます。ただし，$n>30$ とします。

95％の信頼区間　$\overline{X}-1.96\dfrac{s}{\sqrt{n}}\leqq m\leqq\overline{X}+1.96\dfrac{s}{\sqrt{n}}$ ……①

99％の信頼区間　$\overline{X}-2.58\dfrac{s}{\sqrt{n}}\leqq m\leqq\overline{X}+2.58\dfrac{s}{\sqrt{n}}$ ……②

やさしい解説

上記の式①，②を用いれば，誰でも標本調査で母平均を推定できます。たとえば，日本に住んでいる勤労者からランダムに3000人抽出したところ，その貯蓄額の平均は1600万円で，不偏分散から求めた標準偏差は250であるとします。すると，標本平均 $\overline{X}=1600$，標本標準偏差 $s=250$，標本の大きさ $n=3000$ を式①に代入すると，

95％の信頼区間　［1591，1609］

ということになります。つまり，勤労者の貯蓄は確率0.95の正しさで，1591万円以上1609万円以下であることがわかります。

ここでは，どのようにして式①，②が導かれたのかを簡単に示しておきましょう。

母集団の平均が m で，母集団の標準偏差を σ としてみます。

ここから，大きさ n の標本をとった場合，標本平均 $\overline{X}$ の分布は中心極限定理より，n が大きければ，平均が m で標準偏差が $\sigma/\sqrt{n}$ の正規分布に従います。すると，$\overline{X}$ を標準化した，

$$Z=\frac{X-m}{\dfrac{\sigma}{\sqrt{n}}}$$

は標準正規分布に従います。したがって，

$$-1.96\leqq\frac{\overline{X}-m}{\dfrac{\sigma}{\sqrt{n}}}\leqq 1.96$$

を満たす確率は0.95であることがわかります。この式を変形すると，

$$\overline{X}-1.96\frac{\sigma}{\sqrt{n}}\leqq m\leqq\overline{X}+1.96\frac{\sigma}{\sqrt{n}}$$

となります。$n>30$ なので，σ を s に置き換えて（§72参照），

$$\overline{X}-1.96\frac{s}{\sqrt{n}}\leqq m\leqq\overline{X}+1.96\frac{s}{\sqrt{n}}$$

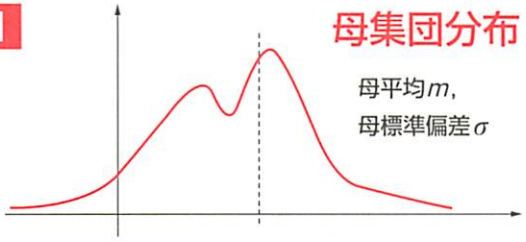
が導かれます。

式②についても同様です。

母平均の推定は公式に入れるだけ

……その原理は中心極限定理

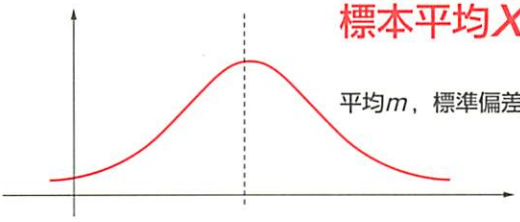
1 母集団分布



母平均m，
母標準偏差σ

中心極限定

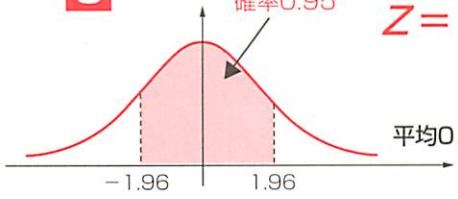
2 標本平均$\overline{X}$の分布



平均m，標準偏差$\frac{\sigma}{\sqrt{n}}$ 正規分布

標準化

3 $Z=\dfrac{X-m}{\frac{\sigma}{\sqrt{n}}}$ の分布



平均0，標準偏差1の標準正規分布

4

$$-1.96 \leqq \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \leqq 1.96$$

を満たす確率は0.95

5

$$\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X}+1.96\frac{\sigma}{\sqrt{n}}$$

を満たす確率は0.95。
ここで$n>30$なので，σを標本標準偏差sに置き換える。

注）式①，②は標本の大きさが 30 を超える大標本の場合です。30 以下の小標本の場合は，ステューデント分布を利用した推定になります（§ 72 参照）。

§83 資料が少ないときの平均値の推定は

ここが要点！

標本の大きさ n が 30 以下の場合，標本平均 $\overline{X}$，不偏分散から求めた標準偏差 s とすると，母集団の平均（母平均）m は次式で推定できます。

95 %の信頼区間　$\overline{X}-t(0.5)\dfrac{s}{\sqrt{n}}\leqq m\leqq\overline{X}+t(0.5)\dfrac{s}{\sqrt{n}}$ …①

99 %の信頼区間　$\overline{X}-t(0.1)\dfrac{s}{\sqrt{n}}\leqq m\leqq\overline{X}+t(0.1)\dfrac{s}{\sqrt{n}}$ …②

ただし，$t(0.5)$，$t(0.1)$ は t 分布の確率で，n の値によって異なります。

やさしい解説

上記の式①，②を用いれば，小標本の場合でも簡単に母平均を推定できます。たとえば，大きな会社の従業員 20 人をランダムに抽出し，血圧を調べたところ平均値は 116.35，標本標準偏差は 11.24 であることがわかったとします。

すると，標本平均 $\overline{X}=116.35$，標本標準偏差 $s=11.24$，標本の大きさ $n=20$ を式①に代入すると，95 %の信頼区間は [111.0，121.7] ということになります。つまり，この会社の従業員の平均血圧は確率 0.95 の正しさで 111.0 以上 121.7 以下であることがわかります。ただし，$n=20$ のときの $t(0.5)$ の値は t 分布表より求めた 2.093 を利用しました。

それでは，なぜ小標本の場合は大標本の場合と推定式が異なるのでしょうか。母集団の平均が m で，母集団の標準偏差を σ としてみます。ここから，大きさ n の標本をとった場合，標本平均 $\overline{X}$ の分布は中心極限定理より，n が大きければ，平均が m で標準偏差が $\sigma/\sqrt{n}$ の正規分布に従います。すると，$\overline{X}$ を標準化した，

$$Z=\frac{X-m}{\dfrac{\sigma}{\sqrt{n}}}$$

は標準正規分布に従います。ところが，この σ は通常，未知なので，これを標本標準偏差 s で代用します。

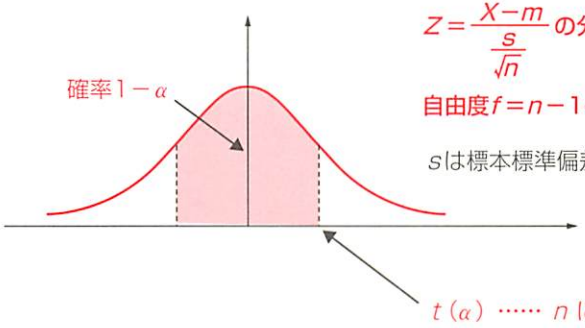
このとき，

$$Z=\frac{X-m}{\dfrac{s}{\sqrt{n}}}$$

は，n が小さいとき正規分布ではなく自由度 $n-1$ の t 分布に従ってしまいます。このことが，大標本の場合の推定式が小標本で使えない理由なのです。

小標本(大きさ 30 以下)の場合は t 分布を利用

1 標本平均を標本標準偏差で標準化したものは n が小さいときは t 分布に従う



$Z=\dfrac{X-m}{\dfrac{s}{\sqrt{n}}}$ の分布は n が小さいとき

自由度 $f=n-1$ の t 分布

s は標本標準偏差

2 t 分布表での t(α) の値は自由度 f の値によって異なる

f \ α	0.1	0.05	0.02	0.01	0.001
1	6.313749	12.70615	31.82096	63.6559	636.5776
2	2.919987	4.302656	6.964547	9.924988	31.59977
3	2.353363	3.182449	4.540707	5.840848	12.92443
4	2.131846	2.776451	3.746936	4.60408	8.610077
5	2.015049	2.570578	3.36493	4.032117	6.868504
6	1.943181	2.446914	3.142668	3.707428	5.958718
7	1.894578	2.364623	2.997949	3.499481	5.408074
8	1.859548	2.306006	2.896468	3.355381	5.041366
9	1.833114	2.262159	2.821434	3.249843	4.780886
10	1.812462	2.228139	2.763772	3.169262	4.586764
11	1.795884	2.200986	2.718079	3.105815	4.436879
12	1.782287	2.178813	2.68099	3.054538	4.317844
13	1.770932	2.160368	2.650304	3.012283	4.220929
14	1.761309	2.144789	2.624492	2.976849	4.140311
15	1.753051	2.131451	2.602483	2.946726	4.07279
16	1.745884	2.119905	2.583492	2.920788	4.014873
17	1.739606	2.109819	2.56694	2.898232	3.965106
18	1.734063	2.100924	2.552379	2.878442	3.921741
19	1.729131	2.093025	2.539482	2.860943	3.883324
20	1.724718	2.085962	2.527977	2.845336	3.849564
21	1.720744	2.079614	2.517645	2.831366	3.819296
22	1.717144	2.073875	2.508323	2.818761	3.792229
23	1.71387	2.068655	2.499874	2.807337	3.767636
24	1.710882	2.063898	2.492161	2.796951	3.745372
25	1.70814	2.059537	2.485103	2.787438	3.725145
26	1.705616	2.055531	2.478628	2.778725	3.706664
27	1.703288	2.051829	2.472661	2.770685	3.689493
28	1.70113	2.048409	2.467141	2.763263	3.673922
29	1.699127	2.045231	2.46202	2.756387	3.659516
30	1.69726	2.04227	2.457264	2.749985	3.645982
1000	1.646379	1.962339	2.33008	2.580746	3.300229

§84 支持率の推定は

ここが要点！

標本の大きさを n，標本比率を $\overline{X}$ とするとき，母集団の比率（母比率）p は次式で区間推定できます。

95％の信頼区間

$$\overline{X}-1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})}\leqq p\leqq \overline{X}+1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})} \quad \cdots\cdots①$$

99％の信頼区間

$$\overline{X}-2.58\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})}\leqq p\leqq \overline{X}+2.58\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})} \quad \cdots\cdots②$$

やさしい解説

上記の式①，②を用いれば，標本比率をもとに母比率を推定できます。たとえば，内閣の支持率を調べるために成人からランダムに1744人抽出したところ，530人が支持していることがわかったとします。そこで，標本支持率 $\overline{X}=0.304$，標本の大きさ $n=1744$ を式①に代入すると，母比率の95％の信頼区間は［0.282，0.326］ということになります。つまり，内閣の支持率は確率0.95の正しさで，0.282以上0.326以下であることがわかります。

ここでは，どのようにして式①，②が導かれたのかを簡単に示しておきましょう。

いま，0と1からなる母集団において1の比率が p であるとします。このとき母集団分布は二項分布で平均は p，標準偏差は $\sqrt{np(1-p)}$ となります。

したがって，ここから大きさ n の標本をとった場合，標本平均（この場合標本比率）$\overline{X}$ の分布は中心極限定理より，n が大きければ，平均が p で標準偏差が，$\sqrt{\dfrac{p(1-p)}{n}}$ の正規分布に従います。すると，$\overline{X}$ を標準化した，

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

は標準正規分布に従い，

$$-1.96\leqq\frac{\overline{X}-p}{\sqrt{\dfrac{p(1-p)}{n}}}\leqq 1.96$$

を満たす確率は0.95であることがわかります。この式を近似計算をしながら変形すると，

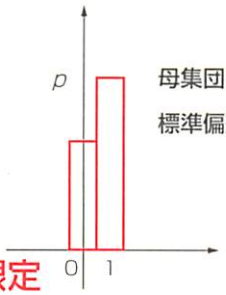
$$\overline{X}-1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})}\leqq p\leqq \overline{X}+1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})}$$

となり，式①が成立するのです。式②の場合も同様です。

母比率の推定は公式に入れるだけ

……その原理は中心極限定理

1



母集団は平均p，

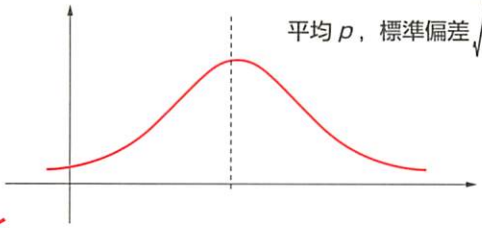
標準偏差$\sqrt{np(1-p)}$の二項分布

中心極限定

2

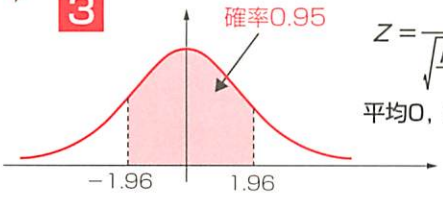
標本平均（この場合，標本比率）$\overline{X}$の分布は，

平均p，標準偏差$\sqrt{\dfrac{p(1-p)}{n}}$の正規分布



標準化

3



$Z=\dfrac{\overline{X}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ の分布は，

平均0，標準偏差1の標準正規分布

4

$$-1.96 \leqq \frac{\overline{X}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \leqq 1.96$$

近似計算をしながら
pについて解く。

式変形

5

$$\overline{X}-1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})} \leqq p \leqq \overline{X}+1.96\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})}$$

を満たす確率は0.95

§85 分散も推定の対象に

ここが要点！

標本の大きさを n，不偏分散を s^2 とすると母集団の分散（母分散）σ^2 は，次式で推定できます。

100(1 $-\alpha$)%の信頼区間は $\dfrac{(n-1)s^2}{\kappa_2} \leqq \sigma^2 \leqq \dfrac{(n-1)s^2}{\kappa_1}$ …①

ただし，κ_1，κ_2 は χ^2（カイ２乗）分布の確率で，n と α の値によって異なります。

やさしい解説

分散は２乗の計算を使って求めますので，平均や比率などに比べて推定するのが難しそうな気がします。しかし，分散については素敵な性質があります。つまり，大きさ n の標本から得られる不偏分散を s^2 とするとき，

$$z=\frac{(n-1)s^2}{\sigma^2}$$

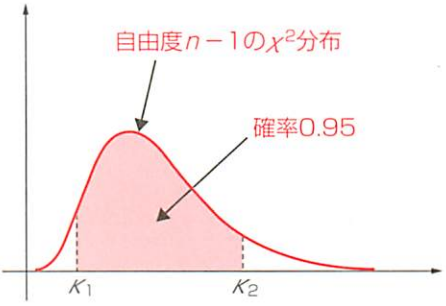
は χ^2（カイ２乗）分布に従うという性質です。ただし，σ^2 は問題にしている母分散です。したがって，n に対して不等式

$$\kappa_1 \leqq \frac{(n-1)s^2}{\sigma^2} \leqq \kappa_2 \quad \cdots\cdots\cdots\cdots ①$$

を満たす確率が 0.95 であれば，式①を変形して得られる

$$\frac{(n-1)s^2}{\kappa_2} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\kappa_1} \quad \cdots ②$$

が成立する確率も 0.95 ということになります。これが，母分散の推定の原理です。



表題の式の信頼度は表現が難しそうですが，$\alpha=0.5$ の場合が 95 %の信頼度ということになります。

それでは，実際に母分散を表題の公式を使って推定してみることにしましょう。たとえば，A社が製造している 30 分用のカセットテープの実際の録音時間の分散を求めるために 10 本のテープの録音時間の不偏分散を調べてみました。すると 0.361 でした。

$s^2=0.361$，$n=10$，χ^2（カイ２乗）分布表より $\alpha=0.5$ のときの κ_1，κ_2 の値を求めて式②に代入すると，母分散の 95 %の信頼度の区間として「0.17 以上 1.21 以下」を得ます。

母分散の推定は公式に入れるだけ

……その原理は中心極限定理

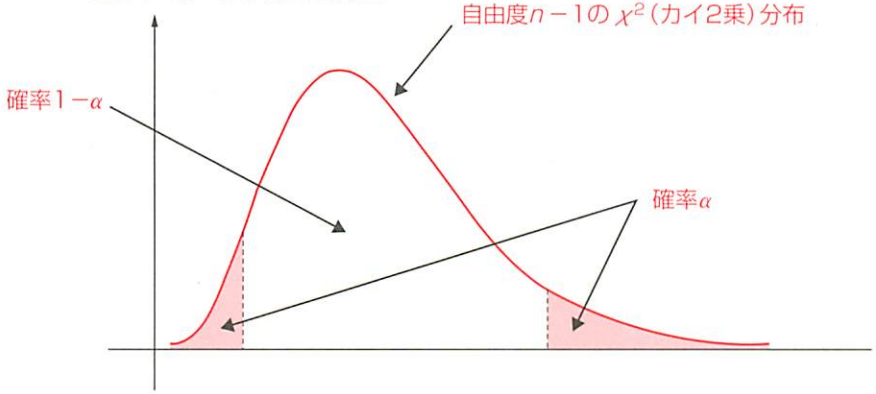
1 100(1−α)%の信頼区間は

100(1−α)%の信頼区間は，

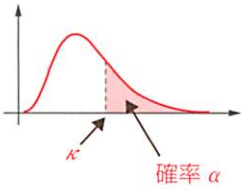
$$\frac{(n-1)s^2}{\kappa_2} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\kappa_1}$$

ただし，nは標本の大きさ，s^2は不偏分散，σ^2は母分散

ここでκ_1，κ_2は下図の通り。



2 χ^2分布表を利用すればκ_1，κ_2がわかる



f \ α	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

§86 危険回避が検定の考え方

ここが要点！ 検定というと難しそうに思えますが，これは「ある仮定のもとで起こりにくいことが起きれば，感心するよりも，この仮定そのものを棄てたほうが無難だよ」という極めて常識的な判断に基づいています。この考え方を利用して，自分の抱いている主張の正しさを説得するのが検定なのです。

やさしい解説

たとえばサイコロを考えてみましょう。ふつうは，どの面が出ることも同等であるように作られていると考えます。ところが，このサイコロを10回振って9回偶数の目が出たら私達はどう考えるのでしょうか。「偶然に10回中9回も偶数の目が出た」と考える人もいるかもしれませんが，ふつうは「このサイコロちょっとおかしいかな」と考えるはずです。つまり，「どの面が出ることも同等であるように作られている」という前提条件を疑います。なぜならば，正常なサイコロなら10回振って9回も偶数の目ということは，きわめて起こりにくいことだからです。どのくらい起こりにくいことなのかを調べると，次のようになります。

10回投げて偶数の目がr回出る確率は，試行の独立の定理から，

$${}_{10}\mathrm{C}_r\left(\frac{3}{6}\right)^r\left(\frac{3}{6}\right)^{10-r}={}_{10}\mathrm{C}_r\left(\frac{1}{2}\right)^{10}$$

となります。

この式を使って，偶数の目が出た回数の確率を求めると，次のようになります。

偶数の目が出た回数	確率
0	0.000977
1	0.009766
2	0.043945
3	0.117188
4	0.205078
5	0.246094
6	0.205078
7	0.117188
8	0.043945
9	0.009766
10	0.000977

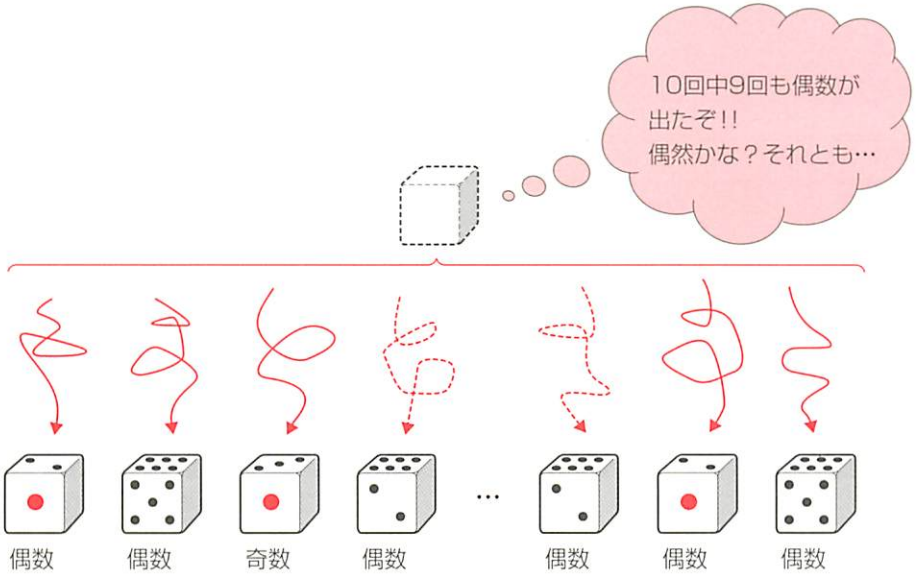
この表からわかるように，10回中9回偶数の目が出る確率は0.009766となります。つまり，1％にも満たないのです。10回中9回以上偶数の目が出る現象に着目しても，その確率は，

$$0.009766+0.000977=0.010743$$

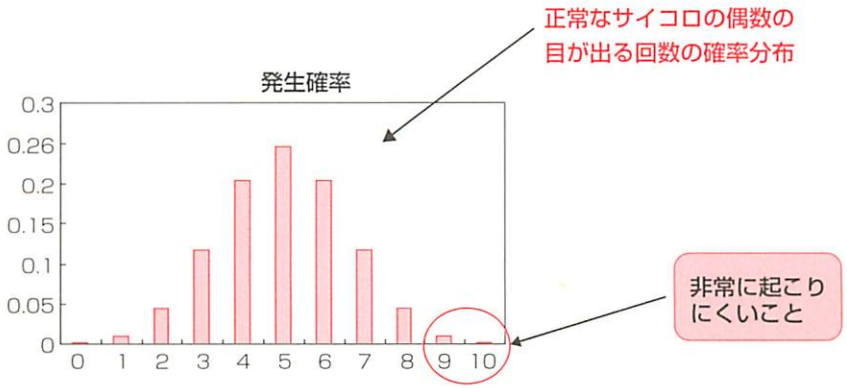
にすぎません。きわめて珍しいことが起きたとして感心するよりも，「このサイコロの目の出方は同等である」という前提そのものを棄てたほうが無難なようです。このとき，この前提が本当は正しいのに，これを棄ててしまう危険性は1％ぐらいで済みます。

起こりにくいことが起きたら仮定を疑え

1 正常なサイコロとしてみたが



2 起こりにくいことが起きてしまった



3 「正常なサイコロ」という仮定を棄てたほうが無難だ

§87 検定そのものは機械的

ここが要点！ 検定そのものは，あまり統計学に詳しくない人でも行えます。それは，検定の手続きが機械的であり，これを踏めば誰でもそれなりの結論を導き出せるからです。素敵な反面，恐ろしいことでもあります。

やさしい解説

検定は，次の手順で行われます。

(1) 対立仮説の樹立

母集団のある特性について，なんらかの予測H_1を得ます（この仮説を**対立仮説**という。これから正しいことを説得したい仮説）。

(2) 帰無仮説の設定

正しいと思われる仮説H_1を否定した仮説H_0を立てます（この仮説を**帰無仮説**という。棄ててしまいたい仮説という意味）。

(3) 棄却域の設定

標本調査や実験をして統計量Tを得たとき，その値がどんな値のときに，この帰無仮説H_0のもとでは起こりにくいことが起きたと判断するのかをあらかじめ決めておきます。起こりにくい，このような値の範囲を**棄却域**といいます。また，その起こりにくいことが起こる確率αはどのくらいかを求めておきます。

(4) 検定で用いる統計量の入手

実際に標本を抽出したり，実験をしたりして検定で用いる統計量を計算します。

(5) 仮説の採否の決定

あらかじめ決めてある判断基準（(3) 棄却域の設定）に照らして帰無仮説H_0を棄てるかどうかを決めます。つまり，統計量が棄却域に入れば有意水準αで帰無仮説を棄却（対立仮説を採択）します。統計量が棄却域に入らなければ帰無仮説を棄てないでおきます。

注）有意水準αとは危険率αのことで，帰無仮説が正しいにもかかわらず，これを捨ててしまう危険性が確率αであることを主張しています。

以上が検定の手順です。この論法は，高校数学で学んだ間接証明法の「背理法」の論理と似ています。ただし注意したいのは，検定はあくまでも確率的な判断であり，絶対に正しいとか絶対に誤りだとか主張しているわけではないということです。

注）「pならばq」を証明するのに，pであってqでないとすると矛盾が起きてしまう。よって，pならばqでしかありえないとする論法が背理法です。

検定はマニュアル通り

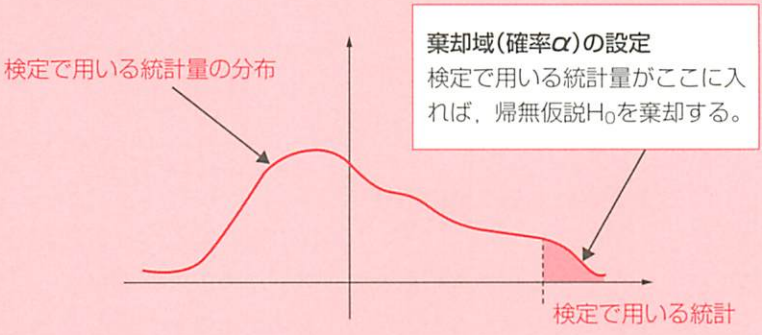
1 母集団のある特性についてなんらかの予測を得る。
…対立仮説H_1の設定

↓

2 対立仮説を否定した帰無仮説H_0を設定

↓

3 検定で用いる統計量と検定する際の判断基準(棄却域)を設定



↓

4 標本を抽出し，検定で用いる統計量を計算

↓

5 判断基準に照らして仮説の採否を決定
統計量が棄却域に入れば，有意水準αで帰無仮説H_0を棄却する。
(対立仮説H_1を採択)
統計量が棄却域に入らなければ，帰無仮説H_0を棄てない。

§88 検定に2種類の誤り

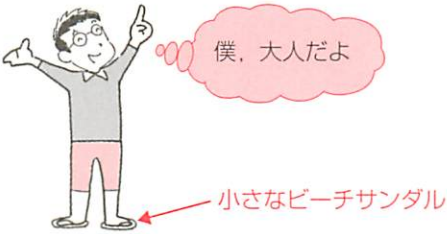
ここが要点！ 検定においては，帰無仮説が正しいにもかかわらず，これを棄ててしまう「第一種の誤り」と，帰無仮説が間違っているにもかかわらずこれを棄てない「第二種の誤り」があります。

やさしい解説

統計的仮説検定においては2種類の判断ミスがあります。1つは，仮説（帰無仮説）が正しいのに仮説を棄却してしまう誤りです。これを**第一種の誤り**といい，第一種の誤りを犯す確率を**有意水準**といいます。これに対して，仮説（帰無仮説）が誤りなのに，これを受容してしまう誤りを**第二種の誤り**といいます。

海辺に残されたビーチサンダルをもとに，その使用者が大人であるか否かを考えてみましょう。

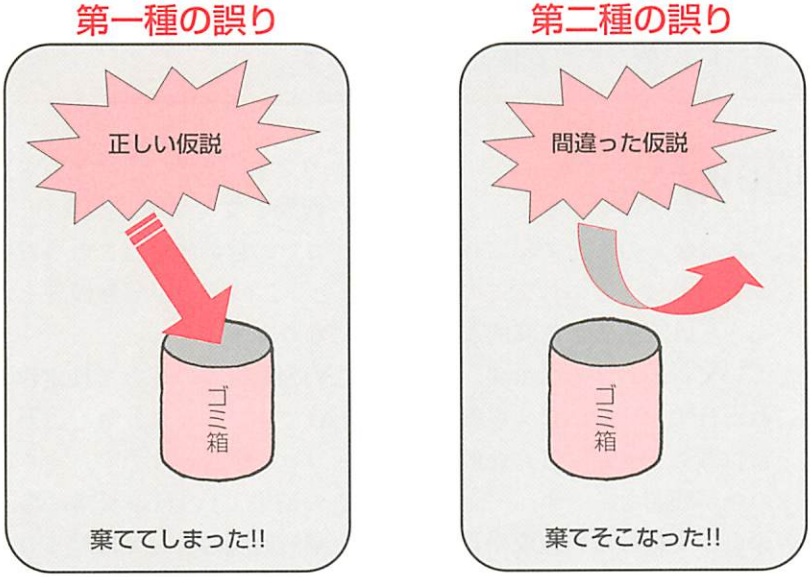
いま，ビーチサンダルがある大きさ以下なら大人でないと判断する検定の方式を採用したとします。この場合，小さなビーチサンダルを見て，大人ではないと判断したのですが，実際にはそれが小柄な大人のビーチサンダルであったということが起こり得ます。そのような誤った判断が第一種の誤りの例です。



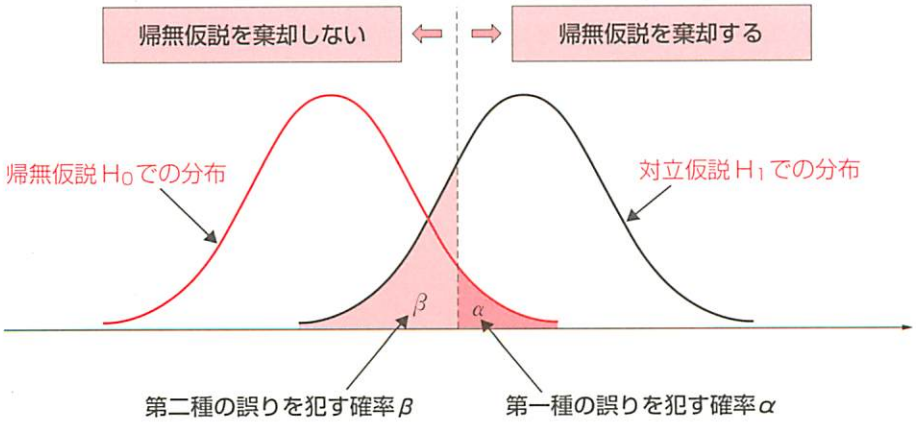
これに対して，ある程度の大きさのビーチサンダルを見て，これを大人のものであると判断したとき，それが実際には子供のビーチサンダルであることもあり得ます。このような判断の誤りは第二種の誤りといいます。

棄却域の設定には第二種の誤りも考慮

1 第一種と第二種の誤りは下図で一目瞭然



2 仮説検定の際には，第一種の誤りを犯す確率α（つまり有意水準）の値を一定の水準に保ちながら，第二種の誤りを犯す確率βの値がなるべく小さくなるように棄却域を定める



§89 いいたいことが対立仮説

ここが要点！ 検定においては，検定者が同意できない（したくない）説のことを“無に帰したい説”という意味で「帰無仮説」といいます。これに対して，検定者が主張したい説を「対立仮説」といいます。

やさしい解説

検定は「ある仮定のもとで起こりにくいことが起きれば，この仮定そのものを棄てる」という極めて常識的な判断にもとづいて行われるものです。しかし，これだけでは検定の考えを理解したことにはなりません。どんな仮定を棄てるのかが問題なのです。

検定の場合，支持したい仮説や正しいと認めたい仮説があるとき，この仮説そのものを検定にかけるわけではありません。面白いことに，「支持したい仮説とは反対の仮説を検定にかける」のです。そして，この仮説を**帰無仮説**といいます。そして，もとの支持したい仮説を**対立仮説**といいます。

たとえば，あるサイコロについて「これはインチキではないか」，つまり，「目の出方に偏りがあるのではないか」という考えに至ったとします。これが対立仮説です。

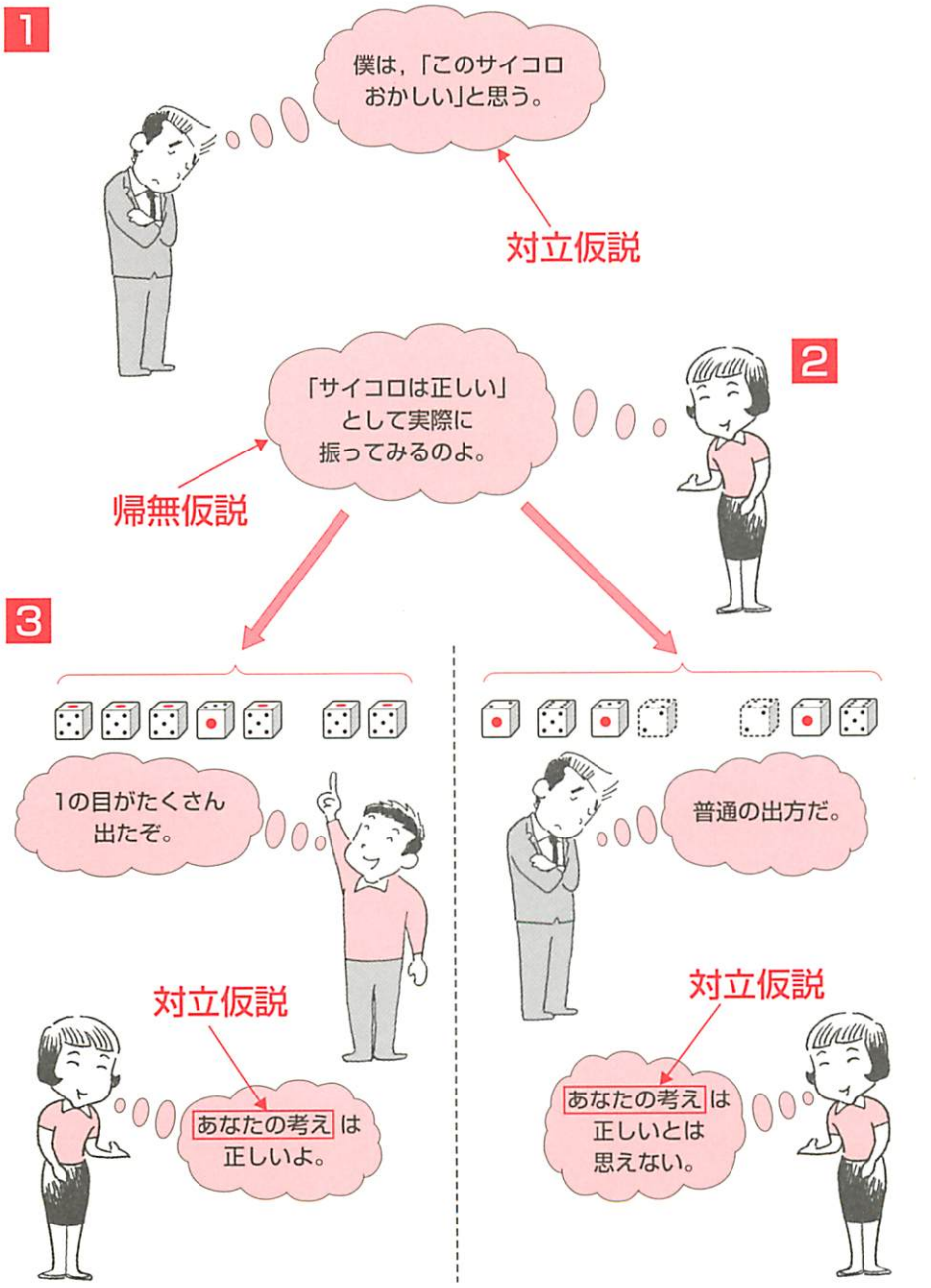
この対立仮説が正しいことを説得したいのです。そこで，対立仮説を否定した仮説を立てます。つまり，「このサイコロの目の出方はどれも同等である」と。この仮説が帰無仮説ということになります。

この帰無仮説のもとで検定作業を行うわけです。つまり，無心にかえってサイコロを振るわけです。

この結果，1の目が異常に多く出れば，帰無仮説のもとでは起こりにくいことが起きたので，帰無仮説を棄てます。……これぞ検定者のやりたかったことです。

しかし，どの目も異常と思えない程度に適当に散らばって出れば，この帰無仮説を棄てることはできません。ある仮説のもとで起こりやすいことが起きたとしても，その仮説を誤りだとして棄てることはできないのです。このとき，帰無仮説を採択するということがありますが，これは「帰無仮説を棄てるほどの理由が見当たらない」という意味です。

自分の信念「対立仮説」を否定したのが「帰無仮説」



§90 ノンパラメトリックとは

ここが要点！ 統計的仮説検定において，母集団分布として特定の分布を仮定しない検定，つまり母数（パラメータ）を使わない検定のことをノンパラメトリック検定といいます。パラメトリック検定は適用範囲が広く便利ですが，多数の検定方法があり，どれを使うか判断するのが難しい。

注）母数とは母集団分布を規定するいろいろな定数のことで，母平均，母分散，母比率，母メジアンなどがあります。

やさしい解説

通常の統計的仮説検定では，母集団分布として正規母集団などの特定の分布を仮定します。また，測定値は連続量であることを前提としています。このように母集団分布を特に定め，測定値として連続量を想定する検定を**パラメトリック検定**と呼びます。

これに対して，母集団分布を特定しなかったり，きわめてゆるい仮定しか設けない検定があります。また，測定値が離散的な順序や度数として表されているデータを対象とする検定があります。このような検定を**ノンパラメトリック検定**と呼びます。

ノンパラメトリック検定には，アンサリ・ブラドレイの検定，ウィルコクスンの検定，ジーゲル・テュキーの検定，順位相関に関する検定，順位和検定，スミルノフの検定，符号検定，マン・ホイトニーのU検定，ラン検定，アンサリ・ブラドレイの検定，……など，じつにいろいろな方法があります。

ノンパラメトリック検定の長所としては次のことがあげられます。

- 適用範囲が広い
- 母集団分布にかかわらず。また，完全な量でない測定値に対して適用できる
- 検定の実施が容易である
- 小標本に対しても適用できる

などなど。

また短所としては，次のことがあげられます。

- 特殊な数表を必要とする場合が多い
- あまりにも多くの方法があり，そのうちのどれを使えばよいかわからない

などなどです。

ノンパラメトリック検定は母数(パラメータ)を使用しない検定

1 順位着目して検定

下表は2つの母集団から抽出した2つの標本である。各データの順位をもとに，2つの母集団の中央値が等しいかどうかを検定する。

	A君	B君	C君	D君	E君	F君	G君	H君	I君	J君
身長	180	170	140	165	155	185	150	160	145	175
順位	2	4	10	5	7	1	8	6	9	3

測定値を順位に置き換えても検定できる。

	イ君	ロ君	ハ君	ニ君	ホ君	ヘ君	ト君	チ君	リ君	ヌ君
身長	155	180	175	165	160	170	150	185	155	145
順位	7	2	3	5	6	4	9	1	7	10

2 度数に着目して検定

下表は従業員30人の前期Xと後期Yの営業成績である。＋の度数と－の度数をもとに，前期と後期の営業成績に差がないということを検定する。

	X	Y	X−Y	X	Y	X−Y	X	Y	X−Y
1	94	99	−	53	50	＋	40	80	−
2	38	50	−	73	97	−	76	33	＋
3	72	54	＋	51	67	−	51	80	−
4	29	68	−	50	70	−	55	48	＋
5	84	85	−	53	64	−	78	81	−
6	45	15	＋	47	66	−	38	69	−
7	90	50	＋	91	80	＋	78	63	＋
8	15	60	−	52	81	−	82	36	＋
9	51	95	−	80	91	−	50	75	−
10	84	76	＋	90	80	＋	54	70	−

§91 統計資料にはウラがある

ここが要点！ 「この統計資料は何気なく作ってみたものです」なんていうことはまずありません。世の中の統計資料の多くは，何らかの目的をもって作られています。とくに，資料をもとにあることを説得しようとするときには，何らかの誇張や省略，それにある程度のウソはつきものです。十分，気をつけなければなりません。

やさしい解説

「今回売り出したやせるための薬は60％強の人がその効力に驚いている」と宣伝されると，そうか過半数の人が効き目を認めているのか，と思われがちです。しかしこのとき，何人の人が標本のサイズを気にしたでしょうか。

もし3000人が調査対象で，このうち2000人がこの薬の効果を認めたのであれば，この宣伝の内容は，まずは受け入れることができそうです。もちろん，調査方法や調査対象者など疑問が残りますが…。

しかし，もし3人が調査対象で，このうち2人がこの薬の効果を認めていたのであれば，60％強の人がその効力に驚いたことは事実としても，この宣伝は受け入れることはできません。サイコロを3回投げて，そのうち1の目が2回出ることは珍しいことではないのですから。

また，よく使われる統計量に平均値があります。これは，それなりに正しく理解されているものですが，じつは実態を反映していないことが少なくありません。

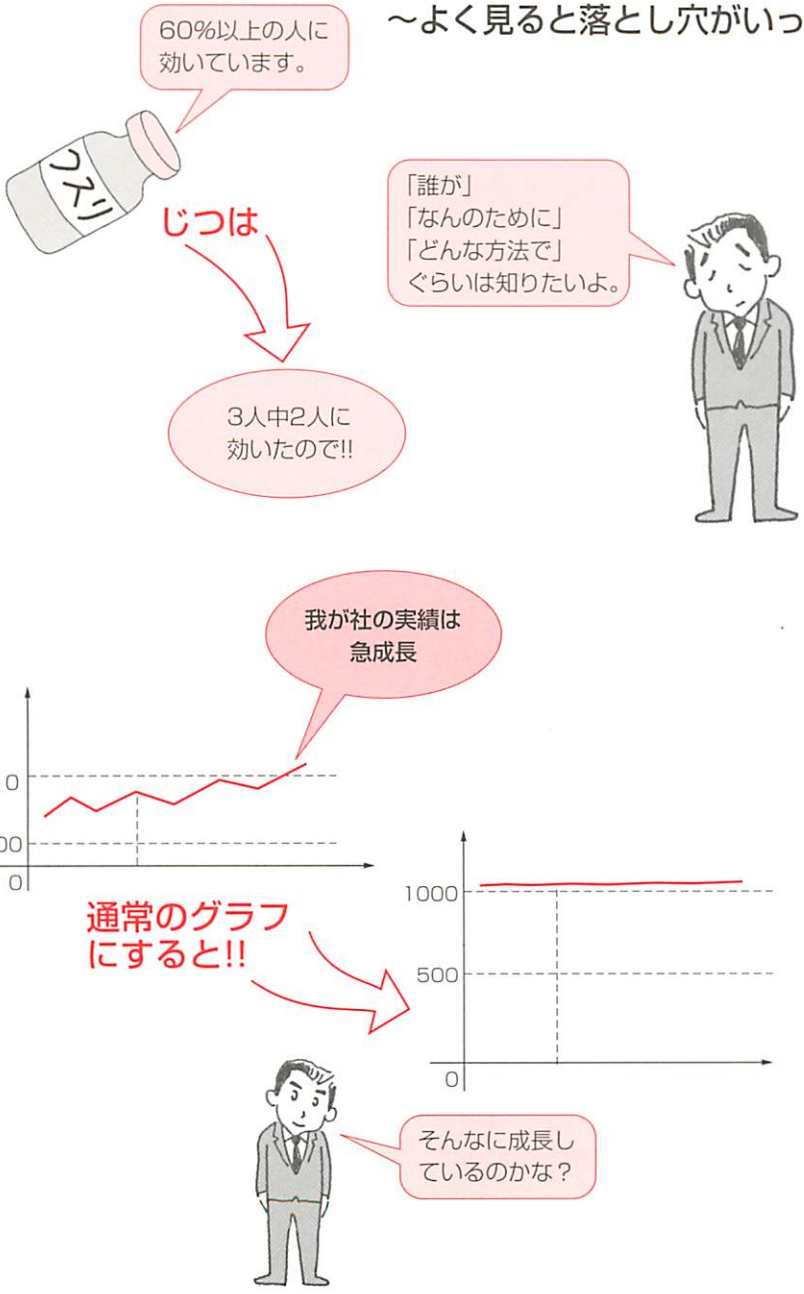
§97で紹介しているように，1世帯当たりの平均貯蓄額が1630万円といわれても，普通のサラリーマンにとっては納得できません。きっと，日本にはお金持ちが多いということなのでしょう。10世帯のうち9世帯が700万円でも1世帯が1億円であれば平均値は1630万円ですから。

ここでは，標本比率と平均値の2つについて少し触れただけですが，統計学にはまだまだたくさんの統計量があり，しかも，これらを複雑に絡めて判断することがあります。

統計資料を見たときには十分注意する必要があります。意外に直観にたよった常識的な判断が正しいことがあります。

統計資料はウラを知れ

～よく見ると落とし穴がいっぱい



§92 RDD法はどんな抽出法

ここが要点！ RDD法とはRandom Digit Dialing（乱数番号法）の略です。この方法は電話帳に掲載されていない番号を含め，すべての固定電話番号の中から乱数を用いて抽出するものです。したがって，すべての世帯の電話番号から無作為標本を作ることができます。

やさしい解説

電話を利用して世論調査を行う場合には，主に，次の2つの方法があります。

(1) TDD法
(2) RDD法

TDD法（Telephone Directory Dialing）は，電話帳に掲載されている世帯から無作為に選ぶ方法です。**RDD法**（Random Digit Dialing）は，存在し得るすべての電話番号から無作為に選ぶ方法です。

世論調査に電話聴取法を導入した背景には，固定電話がほぼ全世帯に普及したという事実があります。そして，加入者がほぼ全員，電話番号を電話帳に掲載していたころはTDD法が有効でした。

しかし，プライバシーの保護や迷惑電話防止などのために，自宅の電話番号を電話帳に掲載しない傾向が都市部を中心に顕著になってきました。

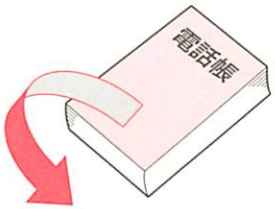
電話帳掲載率は公表されていないため正確な数値は不明ですが，各種の統計を総合すると，2003年時点でおよそ60～70％（全国平均）と判断されます。このため，電話帳を抽出枠として標本抽出をしていた従来のTDD方式では，電話帳に番号を掲載しない人々が調査対象からもれてしまいます。そこで，最近では標本抽出方法をRDD法に変更されるようになりました。

標本調査の理論によると，標本の大きさは2～3千もあれば十分です。しかし，RDD法による場合は，最初の段階でランダムに選ぶ電話番号は，その数倍は必要になります。使用していない電話番号を除去したり，会社関係を対象外にしたりなど，篩（ふるい）にかけて選別する必要があるからです。ただし，その際にも，無作為抽出を損なわないようにする必要があります。

TDD 法から RDD 法へ

1 TDD 法は電話帳，RDD 法は乱数利用

TDD法



03 4322 9942
043 222 2222

RDD法

乱数表

80	86	47	12	25	14	48	50	96	25	21	37	30	76	64	95	90	40	56	69
49	19	98	41	96	12	29	49	39	19	32	44	42	25	62	12	23	18	46	69
84	57	90	47	87	34	98	63	22	30	38	21	68	62	15	52	30	37	14	30
27	33	78	45	49	22	15	37	99	21	77	35	20	95	38	74	47	85	44	14
60	43	60	50	22	50	19	91	94	49	80	42	38	82	73	60	61	46	15	61
89	57	25	69	24	59	86	83	47	22	51	53	64	59	17	64	79	98	25	94
51	29	60	33	17	14	13	45	24	35	54	71	89	38	37	82	21	64	13	48
60	64	69	98	52	88	32	57	46	73	70	81	84	29	30	89	29	95	48	67
12	73	32	12	27	89	16							97	81	96	79	59	16	68
54	22	37	35	82	55	82							70	73	48	45	88	64	29
13	62	33	70	64	49	90							53	14	17	97	68	44	40
42	28	61	57	68	88	95							49	22	87	77	36	28	87
75	67	97	72	74	91	16	48	94	50	30	39	27	95	73	61	30	15	44	98
95	93	68	16	86	72	36	49	21	12	89	54	50	64	98	54	24	34	80	14
37	65	34	22	47	31	74	94	56	93	30	33	86	35	15	87	28	93	47	34
61	55	12	62	27	36	62	39	61	90	90	68	55	15	89	63	52	22	39	98
83	32	35	96	39	41	16	69	77	12	82	75	56	65	39	27	25	43	77	25
15	80	43	75	73	58	27	63	30	29	79	66	61	82	75	69	84	36	95	63
29	60	72	43	22	65	18	93	79	28	23	86	90	67	85	84	82	46	50	67
61	90	78	85	98	84	80	47	13	87	81	94	42	21	83	72	61	84	28	71

2 RDD 法は，必要とされる標本の何倍もの初期標本が必要

§93 このサイコロは正常かな？

ここが要点！ 理論的に予想される度数分布が，実際の調査から得られる度数分布と統計的に一致するかどうかを確かめる方法に「適合度の検定」があります。この検定では，χ^2（カイ２乗）分布が利用されます。

やさしい解説

サイコロを600回振って出た目の数を実際に調べて，結果が下表のようになったとします。

サイコロの目	観測度数A	理論度数B	A－B
1	110	100	10
2	95	100	−5
3	105	100	5
4	90	100	−10
5	80	100	−20
6	120	100	20
合　計	600	600	0

このサイコロが正常であるという仮定のもとでは，

$$\frac{(\text{観測度数}-\text{論理度数})^2}{\text{論理度数}}$$

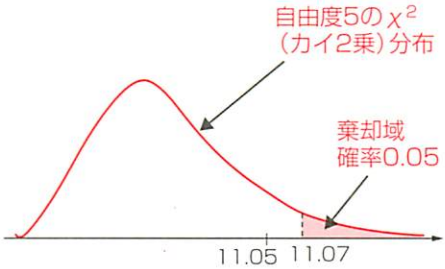
の総和がχ^2（カイ２乗）分布に従うことが知られています。

そこで，χ^2の値を計算してみると，

$$\chi^2=\frac{(110-100)^2}{100}+\frac{(95-100)^2}{100}+\frac{(105-100)^2}{100}+\frac{(90-100)^2}{100}+\frac{(80-100)^2}{100}+\frac{(120-100)^2}{100}$$

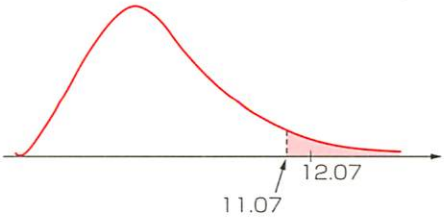
$$=10.5$$

この値は自由度５（＝６－１）のχ^2（カイ２乗）分布に従います。この分布のとき，上側５％の点は11.07であり，χ^2の値10.5は棄却域にありません。



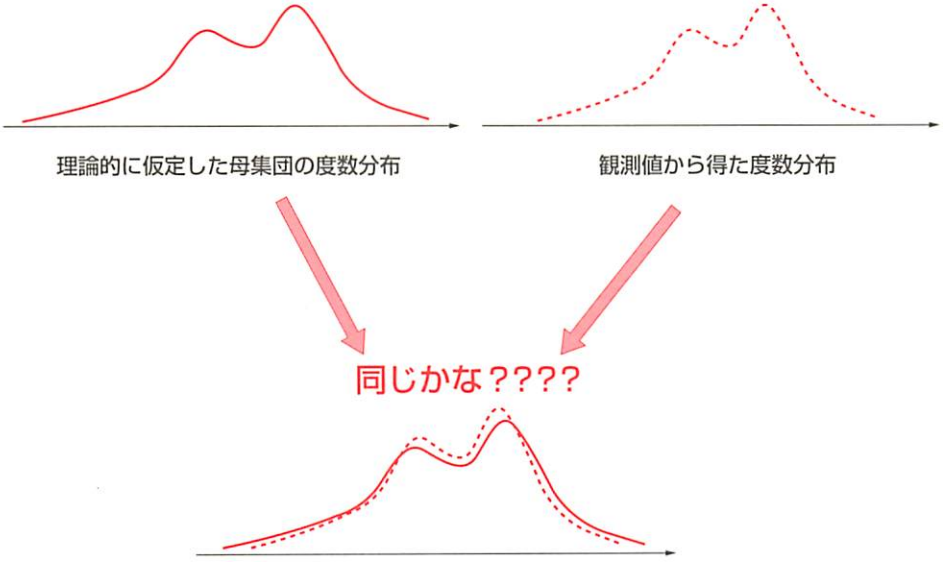
したがって，このサイコロは異常であるとはいえません。つまり，正常なサイコロでも先の表のようなことはよく起こり得ることなのです。

注）もし，χ^2の値が，たとえば12.0であれば，棄却域に入ってしまいますから，「サイコロは正常」という仮説は棄却されます。つまり異常だと判断されます。この判断の誤る確率は0.05です。



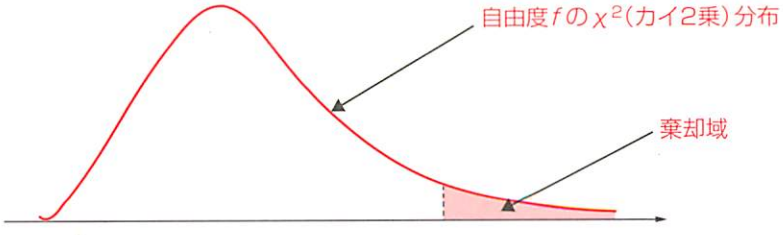
理論度数と観測度数の食い違いはχ^2(カイ2乗)検定で調べる

1 理論度数が観測度数と統計的に一致するかな？



2 1の疑問はχ^2(カイ２乗)検定で調べる

$\frac{(\text{観測度数}-\text{理論度数})^2}{\text{理論度数}}$ の総和 は，χ^2(カイ2乗)分布に従う。



§94 鯛焼き職人の技量のバラツキは違うか

ここが要点！ 工場の2つのラインで生産されている製品の寿命のバラツキは等しいか。女性ゴルファーと男性ゴルファーのスコアのバラツキは同等か。……などというように，2つの集団のデータの散らばり具合が等しいかどうかを問題にすることがあります。このようなときには，「等分散の検定」を利用します。

やさしい解説

下表は鯛焼き職人である太郎君と花子さんが作成した鯛焼きの重さを，おのおの50匹ずつ測定した結果です。熟練の2人でも，まったく同じ重さの鯛焼きを作ることは簡単ではないようです。

＜太郎君の鯛焼きの重さ＞

100.2	100.5	97.71	101.6	101.6	99.84	101.6	98.54	99.18	99.77
102	99.45	100.8	101	100.1	100.6	100.7	100.4	100.4	97.9
100.6	100	101	99.81	99.88	100.9	100.1	100.8	100.4	100.6
98.49	99.59	98.24	98.29	98.69	101.4	100.2	99.7	100	102.2
100.3	100	99.53	99.93	99.9	101.3	99.04	100.7	100.6	100.6

＜花子さんの鯛焼きの重さ＞

101.1	98.84	98.35	100.1	99.69	98.43	99.57	100.8	99.34	101.1
100.1	99.68	100.2	101.1	100.3	99.71	99.79	98.66	99.75	99.77
99.77	99.17	98.59	99.58	101.7	100.6	99.79	99.85	98.84	99.8
101	98.62	99.08	98.4	99.55	99.55	99.18	100.4	100.6	100.5
100.1	100.2	101	100	99.98	99.54	102.1	98.5	98.9	99.96

このデータから2人によって作成された鯛焼きの重さの分散が等しいかどうかを調べてみることにしましょう。

まずは表から太郎君の不偏分散 s_T^2 は約1.04，花子さんの不偏分散 s_H^2 は約0.76です。この値をもとに分散，つまり，バラツキが等しいかどうかを判定しようというわけです。この母分散の違いを問題にするには，**等分散の検定**という手法を利用します。

まずは，「2人の作成した鯛焼きの重さの分散は等しい」という仮説を設定します。このとき，不偏分散の比

$$F=\frac{{s_T}^2}{{s_H}^2} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots ①$$

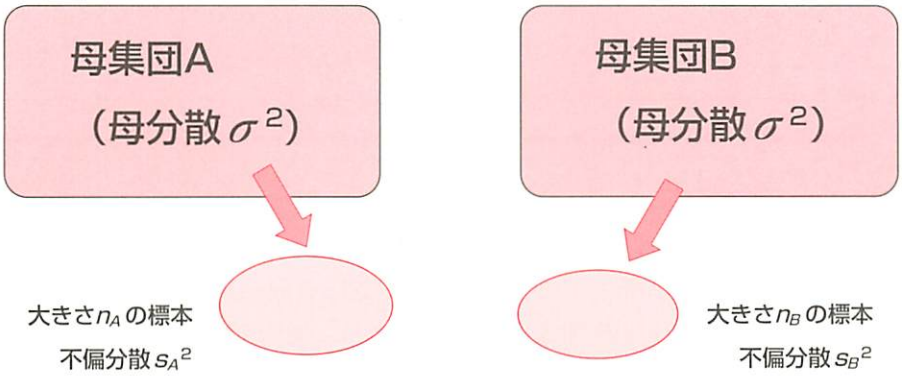
は自由度（n_T-1，n_H-1）の F 分布に従うことが知られています。ただし，n_T，n_H は，それぞれ太郎君と花子さんの標本の大きさ（測定したデータ数。ここでは，ともに50）です。

${s_T}^2=1.04$，${s_H}^2=0.76$ を式①に代入すると $F=1.37$ になります。

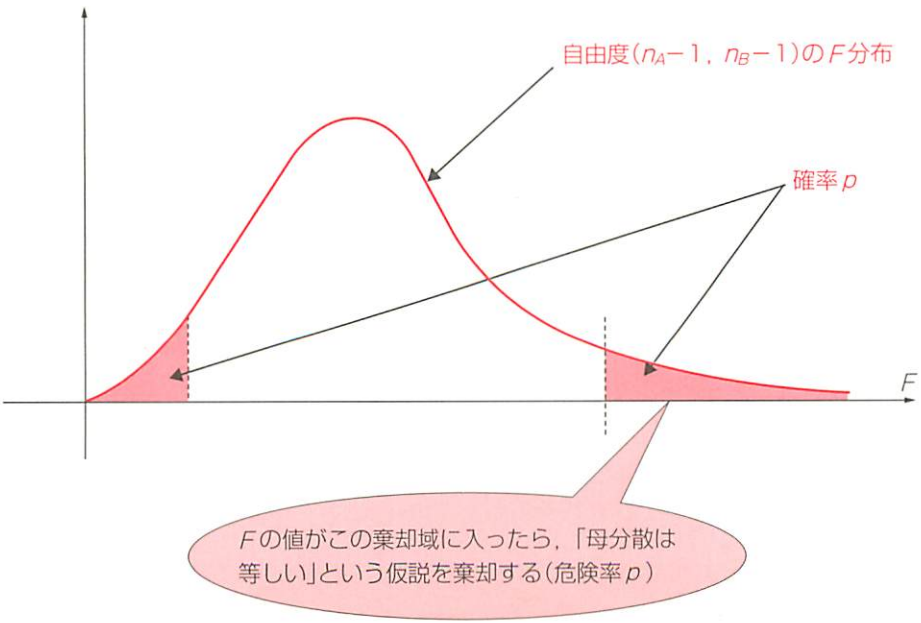
この値は両側5％の棄却域（右の値が1.21）に入り，「2人の作成した鯛焼きの重さの分散は等しい」という仮説は，棄却されます。したがって，2人には，仕事の均一さに差があるということになります。

母分散が等しいかどうかは「等分散の検定」

1 母分散の等しい２つの正規母集団から大きさ n_A，n_B の標本を抽出し，それぞれの不偏分散を s_A^2，s_B^2 とする



2 $F = s_A^2/s_B^2$ は近似的に自由度（$n_A - 1$，$n_B - 1$）の F 分布



§95 1000年に一度は世論調査の対象に

ここが要点！ あなたの世帯がどのくらいの確率で世論調査の対象になるのかは，世論調査の回数などの条件によって異なります。そこで1つの機関が5000世帯を対象に，年間10回の世論調査をしたとしてみましょう。すると，あなたの世帯がこの機関による世論調査の対象になる可能性は，1000年に1回あるかないかということになります。

やさしい解説

少し前の日本の世帯数は約5000万です。そこで，仮に1つの機関がおのおの年10回の世論調査を5000世帯を調査対象として無作為抽出するとします（通常は，5000世帯までは選びません）。

すると，ある世帯がそれらの機関の1回の世論調査で選ばれる確率pは，

$$p=\frac{5000}{50000000}=\frac{1}{10000}$$

となります。このことから，1000年に1回世論調査で選ばれる確率は，次のようになります。

つまり，この間に合計1〔機関〕×10〔回〕×1000〔年〕＝10000回の世論調査が独立に行われたと考えて，

$$=0.367898\cdots$$

$${}_{10000}C_1\left(\frac{1}{10000}\right)^1\left(1-\frac{1}{10000}\right)^{10000-1}$$

となります。

なお，この確率は1000年に一度だけ世論調査で選ばれる確率です。実際には2，3回と選ばれることだってあります。そこで，1000年にk回選ばれる確率を紹介すると，次のようになります。

kが1，2，3，4，5，6，7，8，9，10の

$${}_{10000}C_k\left(\frac{1}{10000}\right)^k\left(1-\frac{1}{10000}\right)^{10000-k}$$

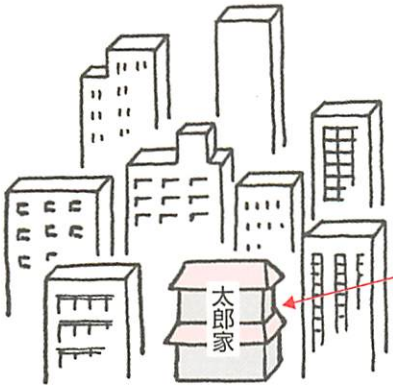
場合を計算すると，次のようになります。

	確　率
K= 1	0.367898
K= 2	0.183949
K= 3	0.06131
K= 4	0.015324
K= 5	0.003064
K= 6	0.00051
K= 7	7.29E-05
K= 8	9.11E-06
K= 9	1.01E-06
K=10	1.01E-07

理論上，kの値は10000まで考えられますが，上記の段階での確率の総和はほぼ0.63になっています。

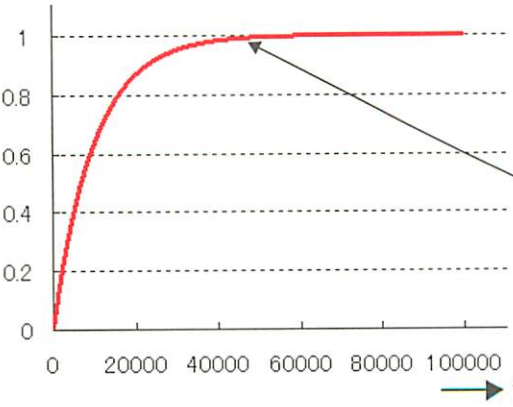
世論調査の対象になるのは大変なこと

1 1つの機関が年間10回世論調査を行うとするとき，この機関の世論調査に選ばれるのは1000年スケールだ



1000年に少なくても1回，太郎家が抽出される確率は0.63

2 世論調査をする機関（新聞社や放送局など）は1つではない。現実的には10機関がおのおの年間10回行うとすると，年間で100回，100年間で10000回の世論調査が行われる。そこで，以下に n 回世論調査が行われたとき，少なくとも1回，その対象となる確率をグラフ化した



少なくても1回対象になる確率
＝1－（1回も対象ならない確率）

$$=1-\left(1-\frac{1}{10000}\right)^{n}$$

§96 ポートフォリオで低リスク高収益率

ここが要点！ ポートフォリオの考え方は，一極集中を避けて，いろいろなものと組み合わせることによって危険を回避したり，収益率を高めようとするものです。

注）ポートフォリオ（Portfolio：紙ばさみ，折り鞄，投資配分）

やさしい解説

むかしから，資産の1/3は不動産，1/3は預貯金，1/3は株式で運用すれば，危険が少ないといわれてきました。このように，たった1つのものに限定しないで，絶えず複数のものに分散させて処理しようということは，危険回避という意味ではすごく大事なことです。これがポートフォリオの考え方です。

いま，2つの株の銘柄A，Bがあったとします。A，Bの収益率をr_A，r_Bとします。この2つの銘柄を$x:y$の割合で組んだポートフォリオをCとしてみます。ただし，$x+y=1$。

このとき，Cの収益率r_Cは，

$$r_C = r_A x + r_B y$$

と書けます。この期待値については，

$$E(r_C) = E(r_A)\,x + E(r_B)\,y \quad \cdots①$$

また，Cの収益率r_Cの分散$\sigma_C{}^2$は，

$$\sigma_C{}^2 = E[\{r_C - E(r_C)\}^2] = \cdots = \cdots$$
$$= \sigma_A{}^2 x^2 + 2\rho_{AB}\,\sigma_A\,\sigma_B xy + \sigma_B{}^2 y^2$$

と書けます。ただし，$E(\ \ )$は期待値$\sigma_A{}^2$，$\sigma_B{}^2$は銘柄A，Bの収益率の分散，ρ_{AB}は銘柄A，Bの収益率の相関係数です。

したがって，$\sigma_C{}^2$の標準偏差σ_Cは，

$\rho_{AB}=1$のとき，

$$\sigma_C = \sqrt{(\sigma_A x + \sigma_B y)^2} = \sigma_A x + \sigma_B y$$

$\rho_{AB}<1$のとき，

$$\sigma_C < \sigma_A x + \sigma_B y \quad \cdots\cdots②$$

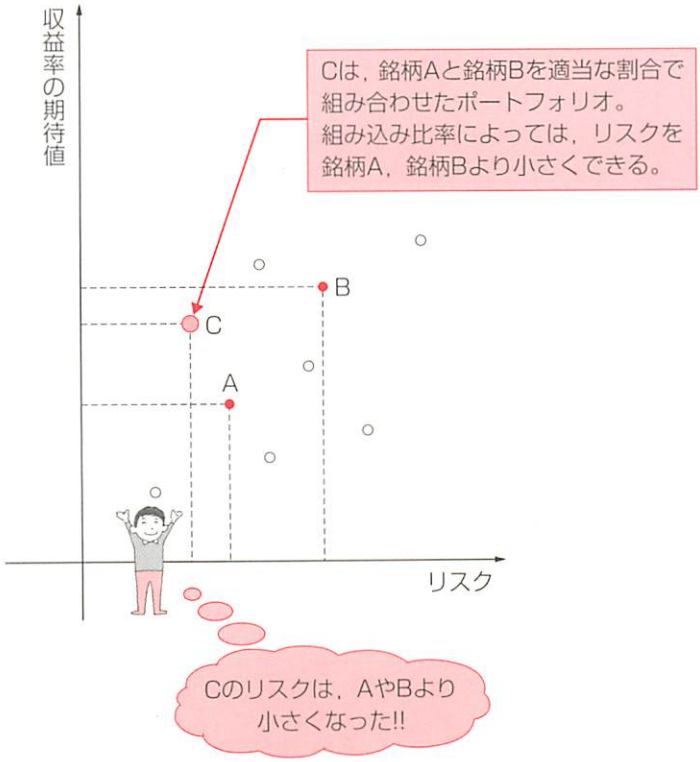
となります。

ここで収益率の標準偏差をリスクと考えると，式②は，ポートフォリオCのリスクは個々の銘柄A，Bのリスクの加重平均より少なくすることができることを主張しています。このとき，式①より，ポートフォリオCの収益率の期待値は銘柄A，Bの収益率の期待値の加重平均のままです。これがポートフォリオのアイデアなのです。

注）投資理論では，期待している収益率とほぼ変わらない結果が出るときはリスクは低いといい，大きく変わるときリスクは高いということにします。したがって，$\sigma_C{}^2$の標準偏差σ_Cをもってリスクと考えます。

株式のポートフォリオで低リスク高収益率

1 株を組み合わせるとリスクを低くできる



2 株，債権，競輪・競馬のリスクと投資収益率のグラフ

注）ポートフォリオの考え方は，マーコヴィッツがシカゴ大学の大学院生であったころにまとめたもので，後にノーベル経済学賞を受けることになった。

§97 貯蓄現在高の平均は1688万円

ここが要点！ 平成14年における1世帯当たり平均貯蓄額は1688万円となっています。いかがでしょう，多くの世帯で「うちには，そんなたくさんの貯金はないよ」とお思いでしょう。これが平均値のトリックです。平均値は，極端な値のデータによって大きく左右されてしまいます。実際には，貯蓄額が200万を割る世帯もたくさんあります。

やさしい解説

総務省統計局統計センターの「貯蓄動向調査」によると，平成14年における全世帯を対象とした1世帯当たり貯蓄現在高は1688万円となっています。

この数値を見て驚いている人は少なくないはずです。しかし，それほど心配する必要はありません。「平均値」は特殊なデータの存在によって，その値が大きく左右されてしまうからです。

たとえば，貯蓄額50万円の人9人と貯蓄額1000万円の人1人の合計10人で統計をとると，平均値は145万円になります。このとき，貯蓄額50万円の9人は「俺はそんなに貯金が少ないのか」と驚いてしまいます。もし，10人全員が145万円の貯蓄額であれば，平均値が145万円だといわれてもフンフンとうなずくだけでしょう。

このように，資料の分布のようすを見ないで，単に平均値だけを論じるのは正しい理解を妨げることになります。

こうした平均値の欠点を補うのが**中央値**（メジアン）や**最頻値**（モード）です。中央値とは，「資料を小さいものから大きなものへと順に並べたとき，中位に位置する資料の値」のことです。

今回の場合，「中央値」（中位数）は1022万円となっています。したがって，1022万円より多い貯蓄額の人が全体の半分，少ない人も半分ということになり多少実感に近づきます。

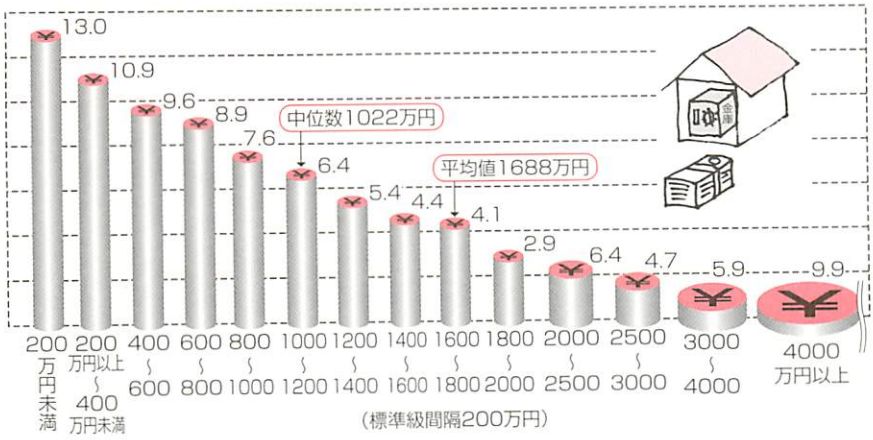
さらに，一番その値をとることが多い「最頻値」で見てみると，この場合200万円以下となっています。これでほっとする人も多いことでしょう。

なお蛇足（だそく）かもしれませんが，「1世帯当たり貯蓄現在高は1688万円」というのは全世帯の平均値です。しかし，勤労世帯に限定して平均値を算出すると，なんと300万円以上も下がった1356万円になってしまいます。不思議ですね。

平均値でピンとこないときはモードで納得

1 貯蓄現在高階級別世帯分布（全世帯）……平成 14 年

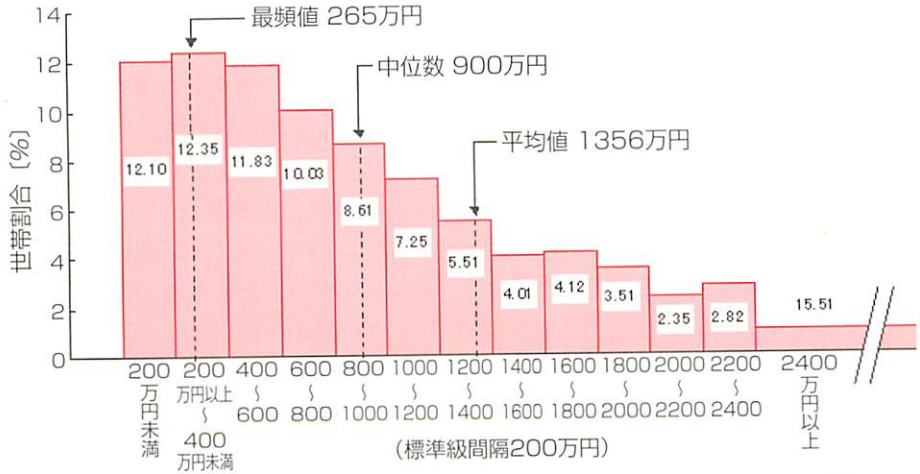
全世帯の約３分の２の世帯が平均貯蓄現在高未満



総務省統計局のホームページより　http://www.stat.go.jp/info/guide/asu/2004/26.htm

2 勤労者世帯のみの貯蓄高……平成 12 年

貯蓄現在高階級別世帯分布（勤労者世帯）



総務省統計局のホームページより　http://www.stat.go.jp/data/chochiku/2.htm

§98 パソコンは確率分布表を不要にした

ここが要点！ 統計学では，さまざまな確率分布表が必要とされています。しかし，これらの表の値は，パソコンにインストールされたソフト（Excelなど）を利用すればすぐに求めることができます。また，プリントアウトすれば簡単に数表を作成できます。

やさしい解説

統計学の本には，必ずといっていいほど正規分布表やt分布表，χ^2（カイ2乗）分布表などの確率分布表が掲載されています。統計解析にこれらの数表が欠かせないからです。

しかし，パソコンが身近にある現在では，それらの必要性は乏しくなってきました。なぜならば，パソコンにインストールされたソフト（Excelなど）を利用すれば，確率分布表の必要な値をすぐに求めることができるからです。

いちいち数表で調べる必要がありません。また，プリントアウトすれば，各書籍に掲載されているような確率分布表が簡単に作成できます。

たとえば，多くのパソコンにインストールされているExcelを例にして調べてみましょう。

平均が0で分散が1である標準正規分布の確率変数Xがz以下の値をとる確率はExcelの組み込み関数

NORMSDIST(z)

を利用すれば求まります。

また，自由度nのχ^2（カイ2乗）分布の確率変数Xがz以上の値をとる確率はExcelの組み込み関数

CHIDIST(z，n)

を利用すれば求まります。

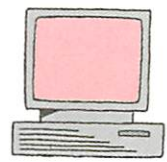
さらに，自由度nのt分布の確率変数Xがz以上の値をとる確率はExcelの組み込み関数

TDIST(z)

を利用すれば求まります。

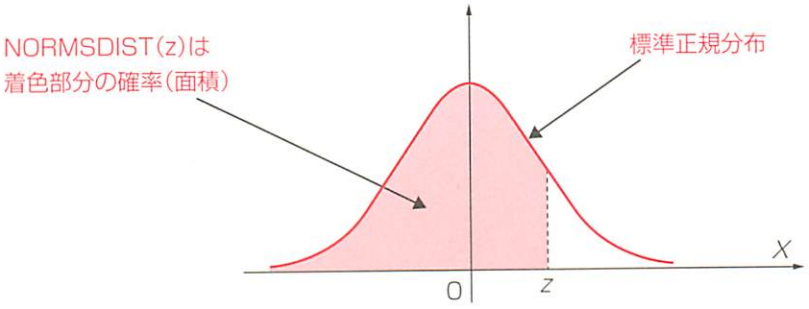
その他，いろいろな確率分布の確率を求める関数が用意されています。それに，このような組み込み関数を利用すると，$x \geqq z$ である確率pを与えて，zの値を知ることも簡単にできます。

このように，パソコンを使うと確率分布のいろいろな値を自由自在に操ることができるのです。



パソコンは確率分布表要らず

1 たとえばExcelに組み込まれた関数 NORMSDIST(z) を利用すれば，標準正規分布の $X \leqq z$ である確率(下図の着色部分)が求まる



2 下表はExcelの関数 NORMSDIST(z) を利用して作成した標準正規分布表

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

§99 円周率とネイピアの数は非常に大事

ここが要点！ よく使われる正規分布（ガウス分布）の式が $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$ で表されることからもわかるように，円周率 π とネイピアの数 e は統計学では非常に重要な数なのです。

注）π と e は数学全般で重要な数です。それゆえ，統計でも重要だというべきかもしれません。

やさしい解説

円周率 π は小学生でも知っているように，円周の長さと直径との比の値です。この π という文字はギリシャ文字の「周り」を表す単語の頭文字だといわれています。

「π は円の面積や円周の長さ，球の体積を求める」ことにも使われますが，用途はこれに限りません。じつにさまざまな分野で利用されています。

また，円周率 π の値 3.14159…は，紀元前から多くの人がその正しい値を求めようとして苦労してきました。精度を1桁あげるのに一生を費やした人もいたそうです。

現在ではコンピュータを利用して，必要とされる桁数の円周率を求めることができるようになりました。

ネイピアの数 e は高校の微積の教科書で初めて顔を出す数で，「h を限りなく0に近づけたときに $(1+h)^{\frac{1}{h}}$ が近づいていく値のことで $e=2.71828\cdots\cdots$」です。

この π と e は，いずれも有理数ではありません。つまり，

$$\frac{\text{整数}}{\text{整数}}$$

とは書けない数なのです。実数で有理数でない数を**無理数**といいます。したがって，π と e は無理数です。しかも，単なる無理数ではありません。π と e は，超越数と呼ばれる無理数なのです。

超越数とは，「代数方程式の解になることができない数」のことです。なお，代数方程式とは，係数が整数（有理数）の次のような n 次方程式のことです。

$$a_nx^n+a_{n-1}x^{n-1}+\cdots+a_2x^2+a_1x+a_0=0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots①$$

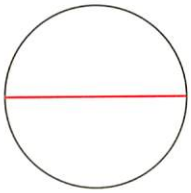
この式①において，整数 a_n，a_{n-1}，$\cdots a_2$，a_1，a_0 をどう選んでも π と e は方程式①の解にはなり得ないのです。

また，円周率 π とネイピアの数 e は他人ではありません。密接な関係で結ばれています。それが次の**オイラーの公式**です。

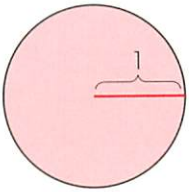
$$e^{i\pi}+1=0 \quad (i \text{ は虚数単位})$$

πとeは数学では非常に大事

1 πの図形的意味

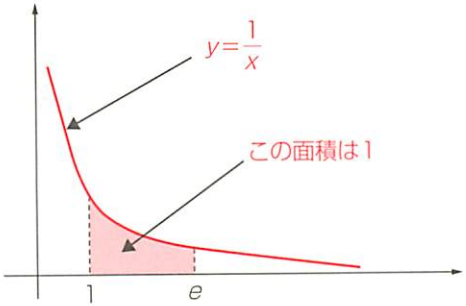


$\pi = \dfrac{\text{円周の長さ}}{\text{直径}}$



π＝半径1の円の面積

2 eの図形的意味



紀元前3世紀…アルキメデスは，円に内接・外接する正96角形による計算によって $3.14084 < \pi < 3.142858$ から $\pi = 3.14$ という近似値を求めた。

1614年……ネイピア（Napier John：1550～1617，スコットランドの貴族）が対数の考えを発見した際にeを見つけた。

1844年……フランスのリウヴィル（Joseph Liouville：1809～1882）は，初めて超越数の存在とそれが無限にあることを証明。

1873年……エルミート（Charles Hermite：1822～1901）は，eが超越数であることを示した。

1877年……カントール（Georg Ferdinand Ludwig Philipp Cantor：1845～1918）は集合論を用い，超越数は代数的数（代数方程式の解となる数）よりはるかに多いことを示した。

1882年……リンデマン（Carl Louis Ferdinand von Lindemann：1852～1939）はπが超越数であることを示した。

§100 統計の歴史

ここが要点！ 統計学には古い歴史があります。ピラミッド建設や，さらには紀元前2000～3000年の中国における人口調査にまでさかのぼることができます。現代の統計学は確率論の上に構築されたもので，標本をもとに全体を推測する「推測統計学」が主流です。

やさしい解説

統計の考え方は大昔からありますが，いわゆる「統計学」(Statistics)という言葉と概念が出来上がったのは17世紀ごろです。

「Statistics」という語は，ラテン語のStans(国家)に由来することからわかるように，統計学は国勢(国の状態)調査研究する上で発達してきました。その後，ゴールトン(1822～1911)やピアソン(1857～1936)らによって「記述統計学」が生み出されました。

これは，データを要約して数学的に記述することが中心でした。つまり，できるだけ多く収集したデータを整理・集約し，もとの集団についての構造や特性を図表や数量などで記述する統計学です。

この統計学の基本は，平均などの代表値や標準偏差のような散布度，それに回帰や相関などを記述することです。

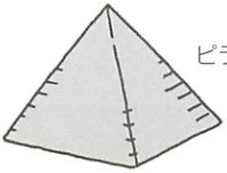
この記述統計学の流れとは別に，その昔から別の発想が芽生えていました。ゴールトンやピアソンをさかのぼる18世紀にはベイズによって，標本から確率分布を求める考えがなされていたのです。また，ガウス(1777～1855)によって，誤差論が研究される過程において，統計学には欠かせない正規分布も見いだされました。

このような背景を踏まえて，20世紀にはゴセットやフィッシャー(1890～1962)らによって，新しい統計学がまとめられました。それは，「母集団から抽出した標本をもとに母集団の特性を確率的に推測しよう」というものです。これを**推測統計学**あるいは簡単に**推計学**と呼んでいます。現在では統計学といえばこの「推測統計学」を指します。

現在の統計学は，高度な確率論を取り込んで数理統計学と呼ばれる分野に発展しています。

統計学の歴史のウラには確率論が

紀元前3000年



ピラミッドを作るための調査

中国の人口調査

18世紀

ベイズ

統計的推測を試みる

ガウス

正規分布

19世紀

ゴールトン

記述統計学

データを要約し
数学的に記述

ピアソン

20世紀

ゴセット

推測統計学

標本と母集団
仮説検定

フィッシャー

数理統計学

数学と確率と
統計の融合

21世紀

〈著者略歴〉

涌井　良幸
昭和25年7月13日生れ
昭和49年　東京教育大学
理学部数学科卒業
現　在　千葉県立千葉東高等学校
数学科教諭

涌井　貞美
昭和27年12月14日生れ
昭和53年　東京大学理学部
大学院修士課程終了
現　在　神奈川県立大原高等学校
数学科教諭

ピタリとわかる確率・統計解析入門

2005年2月25日　発　行　　NDC 417

著　者　涌井　良幸
涌井　貞美
発行者　小川　雄一
発行所　株式会社　誠文堂新光社
〒113-0033 東京都文京区本郷3-3-11
(編集)　電話 03-5800-5763
(販売)　電話 03-5800-5780
http://www.seibundo-net.co.jp/

印　刷　広研印刷(株)
製　本　(株)岡嶋製本工業

検印省略　　万一，落丁・乱丁の場合はお取り替えします．

ISBN4-416-80503-9