基礎統計シケプリ(福地 源一郎)

基礎統計（福地 源一郎：木５）のシケプリです。

◎テストについての情報は下記の通り；

1) テスト問題は講義中に配られるプリントの内容に沿った問題である。
よって、テスト勉強をするには教科書ではなく、配布プリント or シケプリを用いた方が効率的。「それでも不安」「まだまだ勉強したりない」という方は教科書もどーぞ。

2) 平方根計算機能付き電卓が必須。←買えということですね

3) 教科書のみ持ち込み可。
講義の配布プリントやシケプリは持ち込み不可。
持ち込んで OK なのは教科書だけ。←ここ注意！
あと教科書にメモ用紙等を張り付けるのも×
ただし、教科書には何を書き込んでも構わないそうです。このシケプリは配布プリントの内容をまとめたものなので、シケプリに記載されている範囲を確認したうえで教科書と照らし合わせ、テストに出そうな定理や計算式などをあらかじめ教科書にマークしておくことを推奨。

持ち込みおｋってことから考察するには、おそらく記述よりも主として計算・データ処理問題が多数出題される予感。あらかじめ問題をこなしておきましょう。

ちなみに福地氏は今年から駒場に来たっぽいので彼のテストの過去問はありません。いちおう講義で出された例題は載せておきますが、問題の絶対量が少ないので問題をたくさん解きたい人は廣松氏の方の過去問がホムペにあるようなので、そっちの問題をどうぞ。試験範囲は異なっていると思うけどそこはまぁ・・・

○第一章　統計学の基礎

◎統計とは・・・

「ある目的をもって一定の条件（時間、空間、標識）で定められた集団を対象に、調べ集めたデータを集計、加工して得られた数値」のこと。

《特徴》

1) 架空のものではなく、存在が明確に規定された具体的な集団を対象とする。

2) 集団を構成する各個体の特定の性質（＝標識）を数値としてとらえ、集団的に把握する。

※ 2) → 統計的規則性

個々にはばらばらになっていて特徴的な傾向や規則性が見えない現象でも、集団として見た場合、さまざまな傾向や規則性が浮かび上がる。これを統計的規則性という。

◎全数調査と標本調査

・全数調査・・・集団の性質を調べるために、集団を構成するすべての個体について調査する。

・標本調査・・・すべての個体を調査するのではなく、一部のみを調査する方法。母集団から無作為に、あるいは一定の抽出法により、標本を選出し、調査する。

	全数調査	標本調査 （抽出調査、サンプル調査）
調査法	集団を構成する個体すべてを調査	集団の一部のみを調査
例	国勢調査 事業所・企業統計調査など	家計調査 など
長所	・結果の信頼性が大きい	・費用が小さい ・調査に要する時間が少ない
短所	・費用が大きい ・調査に要する時間が長い	・誤差が伴う ・偏った標本で集計すると偏りのある数値が得られる

※国勢調査：５年に一度、総務省統計局により実施。大規模調査と簡易調査がある。人口の把握などが目的。

※家計調査：毎月、世帯などを対象として、家計の収入・支出、貯蓄・負債などを調査。層化三段抽出法で標本を抽出。調査結果は景気動向の把握などの基礎資料として利用。

○第二章　１次元のデータ

◎度数分布表とヒストグラム

統計分析の第一歩として、データがどのように分布しているかを知ることが大事である。度数分布表とヒストグラムはそのための便利な道具である。

度数分布表とは、階級ひと区分あたりのデータの個数（＝度数）を階級順に並べたもの。ヒストグラムとは、柱状のグラフのこと。ヒストグラムは底辺を階級区間に一致させ、柱の面積が度数と比例するように書く。次のように作成の手順を踏む：

1. データ数が比較的少ないときは、大きさの順に並べると見やすい。
2. 階級区分を設けて、各階級に属するデータを数える。
3. 度数分布表・ヒストグラムを作成する。

例:)

データ

162	188	200	158	213	215	232	195	179	197
195	185	184	185	170	183	177	177	205	177
183	201	166	165	185	183	177	177	205	177
194	186	205	215	200	161	189	145	187	195
179	173	196	191	153					

↓

度数分布表

階級	度数	相対度数
140-150	1	0.025
150-160	2	0.050
160-170	4	0.100
170-180	7	0.175
180-190	10	0.250
190-200	7	0.175
200-210	5	0.125
210-220	3	0.075
220-230	0	0.000
230-240	1	0.025

階級はデータのとる値、度数は指定された階級区間におけるデータの個数、相対度数は度数全体に対する指定された階級区間の度数の占める割合を表す。

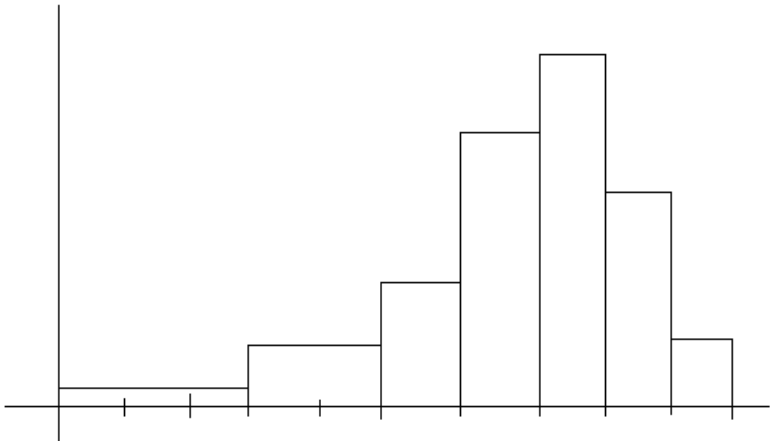
度数分布表を柱状グラフに起こしたものが、ヒストグラムである。この際、階級区分数、すなわち階級区間の個数は 10 前後になるよう、あるいは目的に沿うよう、階級幅を決めるとよい。階級区分数が大きすぎたり、小さすぎたりするとグラフの特徴が見えなくなってしまう。

また、社会・経済現象では、特に分布範囲の一部分が高いことがある。このような場合、度数の集中している部分を細かい幅で分類し、度数が少なくなるにしたがて大きい幅の階級を用いるという手法がとられる。この場合、度数をヒストグラムの面積に比例させるため、

$$\text{階級幅調整済度数} = \text{度数} \times \frac{\text{最小の階級幅}}{\text{階級幅}}$$

を各階級について計算し、これが各階級の柱の高さとする。

こんな感じ。



目盛りが不均一なのは気のせいです

◎代表値・・・分布を代表する値のこと。(算術)平均（＝ミーン)、メディアン、モードなどがある。

●(算術)平均：

$x_1, x_2 \cdots x_n$ についての観測値(算術)平均 $\bar{x}$ は次のように定義される:

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

《特徴》

・異常値(外れ値)による影響が大きい。

・重心の値を示す。

●メディアン：

データ $x_1, x_2, \cdots, x_n$ を値の小さいものから大きさの順に並べ替えたものを $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ で表すと、メディアンは

$x = 2m+1$ のとき $x_{(m+1)}$ 、$x = 2m$ のとき $\dfrac{x_{(m)}+x_{(m+1)}}{2}$

で定義される。

例 1) データが 1,1,2,4,16 であるときのメディアンは 2 である。

2) データが 1,1,2,4,5,16 であるときのメディアンは 2 と 4 の平均をとって 3 である。

《特徴》

・異常値による影響が小さい。

●モード：

度数分布表で、度数が最大である階級の階級値がモード。

《特徴》

・峰が二つある場合、有効ではない。

●平均、メディアン、モードの関係

峰が一つであり、分布が完全に左右対称の場合、この 3 つは完全に一致。

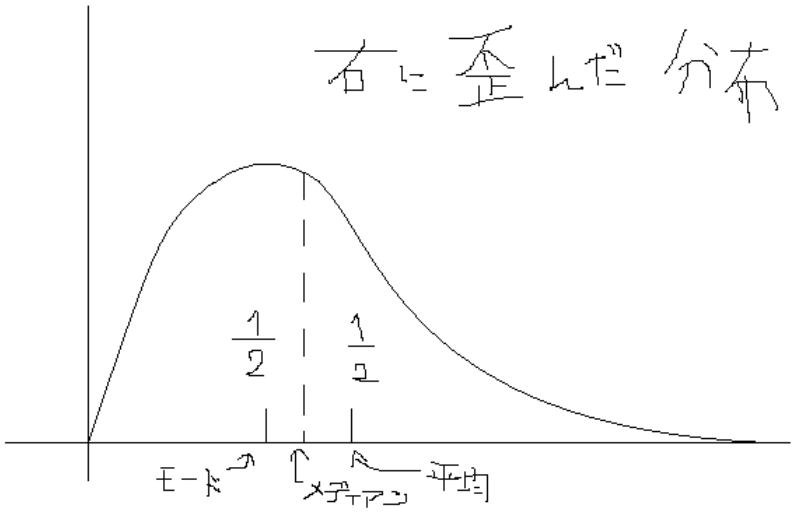
右に歪んだ分布（峰が左に寄っている分布）では一般的に平均、メディアン、モードの順に値が小さくなる。

Mean>Median>Mode　で、辞書と同じ順で並ぶと覚える！

下図を参照のこと。

ここで注意点は、「右(左)に歪んだ分布」は左(右)に峰がある、ということ。

ちなみに左に歪んだ分布だと逆になる。



歪んでいるのは字の方だとかそういう突っ込みはやめてね

◎分散と標準偏差

代表値だけではその分布の様子を決定することはできない。そこで分散と標準偏差という散らばりの尺度を示す値を用いる。

分散は S^2 で表され、

$$S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}$$

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

で定義される。(1/n を忘れないように！)

これは、$x_1, x_2, \cdots, x_n$ の観測値およそ一個あたりが $\bar{x}$ とどれくらい離れているかを計算するものであるが、二乗しているため、単位を揃えるために標準偏差、すなわち

$$S = \sqrt{S^2}$$

を用いることが多い。

実際に標準偏差を計算するときは、次の形が便利:

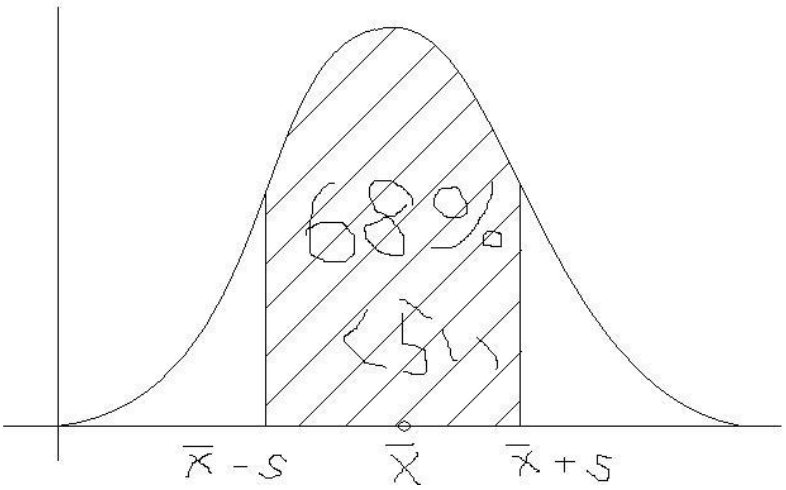
$$S^2 = \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - (\bar{x})^2$$

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

下図のヒストグラムのような形状の分布をベル型の分布という。

人間の特性値（身長・体重など）、大規模な試験の点数分布、測定誤差の分布などはベル型に近いことが知られている。

ベル型の分布では、平均の位置から１標準偏差離れた点にヒストグラムの変曲点が位置している。また、平均から±１標準偏差の区間 ＝ $[\bar{x}-S,\ \bar{x}+S]$ に度数の約 68%が含まれている。



◎一次変換・標準化

データ $x_1, x_2, \cdots, x_n$ について、

$$z_i = ax_i + b$$

のように一次変換する。このときの平均、分散、標準偏差の値は、

$$\bar{z} = a\bar{x} + b$$

$$S_z^2 = a^2 S_x^2$$

$$S_z = |a| S_x$$

となる。データの一次変換は、為替レートなどに利用される。

また、特にデータ $x_1, x_2, \cdots, x_n$ について、

$$Z_i = \frac{x_i - \bar{x}}{S_x} \quad (i = 1,2,\ldots,n)$$

のように一次変換することにより、

$$\bar{z} = 0$$

$$S_z^2 = S_z = 1$$

と変換できる。このような変換を特に標準化という。

標準化された値は、各観測値が標本平均から正負どちらの方向に、標準偏差の何倍離れているかを測っているのである。標準化されたデータをみることによって、個々の観測値の相対的な位置を知ることができる。

○第三章　2 次元のデータ

◎2 次元データ

観測対象の個体に対して、2 変数 x , y を観測して n 組のデータを得る場合、そのデータを 2 次元データという。

2 次元データを　$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ で表す。

◎相関関係

2 つの変数の間に直線関係に近い傾向が見られるとき、その 2 つの変数に「相関関係がある」という。

特に、一方の変数が増加するときに他方の変数も増加する傾向を「正の相関関係がある」、減少する傾向を「負の相関関係がある」という。

また、直線的な傾向の度合いは「強い」「弱い」と表現する。

●相関係数

データが $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ で与えられたとき、x と y の相関係数 r_{xy} は次のように定義される:

$$r_{xy} = \frac{\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{n}}{\sqrt{\frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}}\sqrt{\frac{(y_1-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n}}}$$

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

$$= \frac{C_{xy}}{S_x S_y}$$

ここで、$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$ は共分散と呼ばれる。

《相関係数の特徴》

・常に $-1 \leq r_{xy} \leq 1$ が成り立つ。

・x と y の間に正の相関があるとき、$r_{xy} > 0$

負の相関があるとき、$r_{xy} < 0$

・すべてのデータが直線状にあるとき、$r_{xy} = 1$ あるいは $r_{xy} = -1$ である。このとき、それぞれ「x と y が正の完全相関である」「負の完全相関である」という。

●直線および平面のあてはめ

変数 x が変数 y をある程度決定する関係があるとき、x を独立変数（説明変数）、y を従属変数（被説明変数）という。このとき x と y の関係を近似する一次式 $y = a + bx$ を求め、それを用いて分析を行う方法が回帰分析である。

近似直線を求める方法に、最小二乗法がある。これは、点(x_i, y_i) から直線 $y = a + bx$ への垂直距離の二乗和を L とすると、

$$L = \{y_1 - (a + bx_1)\}^2 + \cdots + \{y_n - (a + bx_n)\}^2 = \sum_{i=1}^{n}\{y_i - (a + bx_i)\}^2$$

を最小にする a , b を求める方法である。これによって、a , b の値は次のように得られる：

$$a = \bar{y} - b\bar{x}$$

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

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{C_{xy}}{S_x^2}$$

上式によって得られた a , b による一次式 $y = a + bx$ は回帰直線と呼ばれる。特に b は偏回帰係数と呼ばれる。

○第四章　確率

さいころを投げたりするような、同じ条件のもとで繰り返し行うことができる実験や観測を試行といい、試行の結果として起きる事柄を事象という。

1 個のさいころを投げる試行において、「1 か 2 の目が出る」という事象を A とすると、

A={1,2} のように表せる。このとき、1,2 は A の要素と呼ばれる。また、A の要素の個数を #A で表す。

Ω={1,2,3,4,5,6} は「1 から 6 の目が出る」という事象で、標本空間と呼ばれる。Ω は必ず起こる事象であり、A は Ω の部分集合である。Ω 内の要素:ω=1,2,3,4,5,6 は標本点と呼ばれる。Ω の一つひとつの要素から成る集合{1},{2},{3},{4},{5},{6} はそれぞれこれ以上細かく分けることができない。これらの事象を根元事象という。

また、「何も起こらない」とする事象を空集合 ϕ で表す。

集合には演算も使用される。演算については次のように定義される:

積事象 $A \cap B$:A かつ B が起こる事象

和事象 $A \cup B$:A または B が起こる事象

補事象 A^c : 事象 A が起こらないという事象

$A \cap B$ が空集合のときに、「A と B は互いに排反である」という。

さて、確率とは、事象の起こる「確からしさ」を数値で表したものであり、次のように求められる:

1. 起こりうる結果が有限個で、それぞれ可能性が同様に確からしいとき（例えばサイコロを投げる試行)、事象 A に属する要素の個数を標本点の個数で割ったものを事象 A の確率とする。つまり、その事象 A の起こる確率を P(A) とすると、

$$P(A)=\frac{\#A}{\#\Omega}$$

2. 等可能性の原理による方法が適用できないとき（例えば将棋の駒や画びょうを投げる試行)、事象 B が起こる回数を試行の回数で割った比率を事象 B の確率の近似値とする。

以上のどちらの方法で確率の値を決定しても、確率 P(・)は次の 3 つの性質をもつ:

(C1) 任意の事象 A に対して $0 \leq P(A) \leq 1$

(C2) 標本空間 Ω に対して $P(\Omega) = 1$, 空事象 ϕ に対して $P(\phi) = 0$

(C3) A と B が互いに排反ならば $P(A \cup B) = P(A) + P(B)$

以上 3 つの条件を確率の公理という。現代数学では、確率の公理を満たせば、P(・)は確率であると定義する。

確率の公理から、次のような性質を導くことができる:

・$P(A^c) = 1 - P(A)$

・$A \subset B$ ならば、$P(A) \leq P(B)$

・任意の事象 A,B に対して $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

・$A_1, A_2, \ldots, A_n$ が互いに排反ならば、$P(A_1 \cup \ldots \cup A_n) = P(A_1 + \cdots + A_n)$

◎条件付き確率と独立性

●条件付き確率

赤玉 3 個と白玉 2 個が入っている袋から、1 個ずつ 2 回玉を取り出す試行を考える。

この試行において、1 回目に赤玉が出るという事象を A 、2 回目に赤玉が出るという事象を B とする。ただし、取り出した玉はもとに戻さないものとする。

このとき、1 回目に赤玉が出たときに 2 回目に赤玉が出る確率は $\frac{2}{4}$ である。このような確率を、A が起こったときに事象 B の起こる条件付き確率という。これを $P(B|A)$ で表す。

（注：本当は/と逆向きの棒線です。打てなかったので縦棒にしてあります）

上の例では$P(B|A)$が簡単に求まったが、等価性原理が使えないような場合では違う方法で計算することが必要となる。

いま、標本空間Ωは 5×4=20 個の要素から成る。事象 A は 3×4=12 個、事象 A∩B は 3×2=6 個の要素から成る。A が起こったときに B の起こる条件付き確率は、A を新たに標本空間と考え、その標本空間の中で B も起こる確率だから、

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

となる。これは等価性原理の使えないときでも成り立つので、上式を条件付き確率の定義とする。また、これを変形すると、

$$P(A \cap B) = P(A)P(B|A)$$

となる。これを乗法公式と呼ぶ。また、P(B)>0 のとき、

$$P(A \cap B) = P(B)P(A|B)$$

もまた成り立つ。

●独立性

2 つの事象 A、B について $P(A \cap B) = P(A)P(B)$が成り立つとき、A と B は互いに独立であるという。この条件は、P(A)>0 のとき、

$$P(B|A) = P(B)$$

と同等である。上式は、事象 A が起こるという条件を加えても、事象 B の確率は変わらないことを意味している。

3 つの事象 A、B、C について $P(A \cap B) = P(A)P(B)$、$P(A \cap C) = P(A)P(C)$、$P(B \cap C) = P(B)P(C)$、$P(A \cap B \cap C) = P(A)P(B)P(C)$ のすべてが成り立つとき、A、B、C はすべて独立であるという。4 つ以上の場合も同様に定義する。

○第五章　確率変数

具体的な現象を記述したり、分析するためには、確率的に動く量(変数)、また、その値の出方の様子を表すことができるとよい。それを表すものが、それぞれ確率変数と確率分布である。

（例：サイコロの出る目は確率変数、サイコロの目の出る確率のグラフは確率分布）

◎期待値

不連続な値をとる確率変数を離散型確率変数という。サイコロの目、コインの表裏などがその例である。X は離散型確率変数とし、とり得る値が $x_1, x_2, \dots, x_k$ であるとし、また、$X = x_i$のときの確率を$p_i = P(x_i)$とすると、このとき、

$$E(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_k p(x_k) = \sum_{i=1}^{k} x_i p(x_i)$$

を X の期待値あるいは平均という。

●確率変数の関数の確率分布と期待値

X が離散的確率変数であれば、その関数g(X)も離散的確率変数である。$Y = g(X)$とし、かつ Y は l 通りの異なる値 $y_1, y_2 \dots y_l$をとると仮定すると、

$$E[g(X)] = \sum_{j=1}^{l} y_i p(y_i)$$

である。すなわち、一般に g(X)の期待値を求めるには、g(X)の確率分布を求めてから上の期待値の定義式を用いればよい。しかし、わざわざ確率分布を計算しなくとも、実際には次の形の式で計算する方がより簡単である場合が多い：

$$E[g(X)] = \sum_{i=1}^{k} g(x_i) p_i$$

《期待値の性質》・・・c は定数であるとする。

(a) $E(c) = c$

(b) $E(X+c) = E(X)+c$

(c) $E(cX) = cE(X)$

(d) 2 つの関数 g、h に対して、$E[g(X)+h(X)] = E[g(X)] + E[h(X)]$

証明は簡単なので割愛します。

◎分散と標準偏差

確率変数がどのくらいばらつくのかを測る特性値として、確率変数の分散がよく使われる。X の分散は次のように定義される：

$$\mathrm{Var(X)} = \mathrm{V(X)} = \mathrm{E[X - E(X)^2]} = \sigma^2$$

ここで、$\mathrm{g(X)} = \{\mathrm{X - E(X)}\}^2 = (\mathrm{X} - \mu)^2$とおいて確率変数の式に代入すると：

$$\mathrm{V(X)} = (\mathrm{x_1} - \mu)^2 + \cdots + (\mathrm{x_k} - \mu)^2 = \sum_{\mathrm{i}=1}^{\mathrm{k}} (\mathrm{x_i} - \mu)^2$$

さらに、期待値の性質(d)から、次のようにも書ける：

$$\mathrm{V(X)} = \mathrm{E(X^2)} - \mu^2$$

また、単位をそろえるために分散の平方根をとる。この値を X の標準偏差といい、SD(X)で表す。つまり、

$$\mathrm{SD(X)} = \sqrt{\mathrm{Var(X)}} = \sigma = \sigma(\mathrm{x})$$

である。分散と標準偏差に関しては第二章も参考のこと。

《分散の性質》・・・a と b は定数であるとする。

(a) $\mathrm{V(a)} = 0$

(b) $\mathrm{V(aX + b)} = \mathrm{a^2 V(X)}$

証明は割愛します。

分散と標準偏差を表す記号については、どれがテストにでてもいいようにすべて覚えておいてください。

※確率変数の変換にはちょっと注意（特に分散と標準偏差)。

$$\mathrm{E(aX + b)} = \mathrm{aE(X)} + \mathrm{b}$$

$$\mathrm{V(aX + b)} = \mathrm{a^2 Var(X)}$$

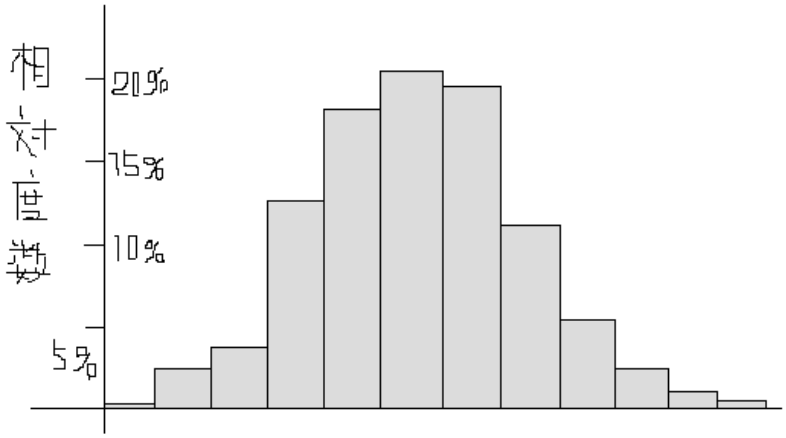
$$\sigma(\mathrm{aX + b}) = |\mathrm{a}|\sigma(\mathrm{X})$$

◎連続確率変数

連続した値をとる確率変数を連続型確率変数という。身長、体重、気温、為替レートな

どがその例である。これを、縦軸を相対度数とした相対ヒストグラムに起こすと、ヒストグラムの柱の上部がなめらかな曲線で近似できるように見える。

下図が相対ヒストグラムの例である：



試行の数がどんどん大きくなるとき、ある区間の相対度数は確率変数がその区間に入る確率におおよそ近くなると考えられる。そこで、確率の理論ではこのような曲線、すなわち関数を使って連続型の確率変数の確率分布を表すのである。

連続型確率変数 X に対して近似曲線となる関数 $f(x) \geq 0$ を定めるとき、任意の区間[a, b]に対して $a \leq X \leq b$ となる確率は、f(x)グラフ、直線 x=a、x=b と x 軸で囲まれる部分の面積に等しくなる。このような f(x)を確率密度関数(pdf)という。言い換えれば、確率変数 X の確率密度関数とは、すべての $a < b$ に対して

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

が成り立つような関数のことである。

ちなみにこのときの期待値は次のように定義される：

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

期待値と分散の性質は、すべての連続確率密度関数に対しても成り立つ。

○第六章　確率分布

◎正規分布

統計学の中でもっとも重要な連続型確率分布は正規分布である。確率密度関数 f(x)が次の

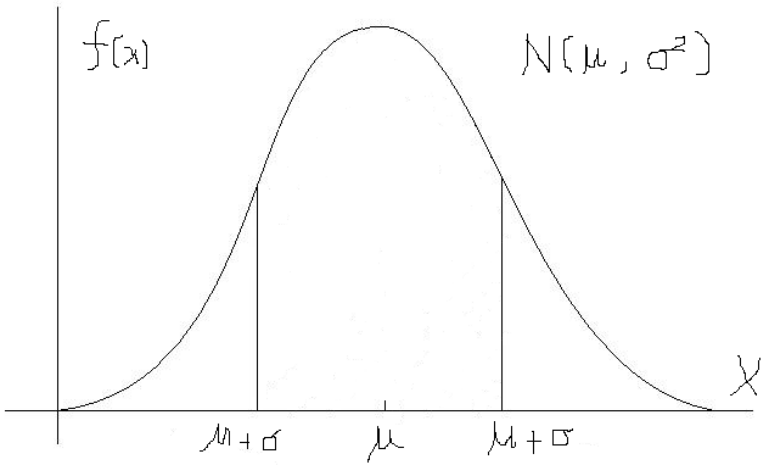
式で与えられるような確率分布を正規分布といい、$N(\mu,\sigma^2)$で表す：

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

また、$X\sim N(\mu,\sigma^2)$ である（＝「確率変数 X の確率分布が正規分布 $N(\mu,\sigma^2)$ である」、「X が正規分布 $N(\mu,\sigma^2)$ に従っている」）とき、

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

であることがわかっている。このことから、正規分布 $N(\mu,\sigma^2)$ は、平均μ ,分散がσ^2の正規分布と呼ばれる。特に平均 0、分散 1 の正規分布 $N(0,1)$ は標準正規分布と呼ばれる。



《正規分布 $N(\mu,\sigma^2)$ の確率密度関数 f(x)の性質》

(a) f(x)のグラフはμ を中心とするベル型である。

(b) 2 つの点 $(\mu-\sigma, f(\mu-\sigma))$ と $(\mu+\sigma, f(\mu+\sigma))$ が変曲点になっている

(c) 区間$[\mu-3\sigma, \mu+3\sigma]$にほとんどの確率が存在する。（約 0.997%）

正規分布が統計学でもっとも頻繁に用いられるのには次の 2 つの理由がある：

(1) 近似的に正規分布に従うと考えられる確率変数が多くあること。

(2) 確率分布が正規分布のときには、統計分析の理論的な結果が簡潔な形で得られること。

また、正規分布に従う確率変数を一次変換したものも正規分布に従うことがわかっている。よって、$X\sim N(\mu,\sigma^2)$であるとき、次が成り立つ：

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

特に、$Z=(X-\mu)/\sigma$ とおいたときに、

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

である。

このとき、Z は確率変数 X を標準化した変数と呼ばれる。X が正規分布に従うとき、X を標準化した変数は標準正規分布に従うのである。標準正規分布の上側確率と上側パーセント点は数表に与えられることから、標準化によってあらゆる正規分布を扱うことが可能になる。

また、一般には複数の独立な確率変数の和の確率分布各確率変数の確率分布とは異なる形状をしている。例えばサイコロを 1 個投げたときの確率分布は 2 個投げたときのそれと全く異なっている。しかし、独立に正規分布に従う確率変数の和は正規分布であることがわかっている。このとき、次が成り立つ：

n 個の確率変数 $X_1, X_2, \dots, X_n$ は互いに独立であり、$X_i \sim N(\mu_i, \sigma_i^2)$ であるとすると、

$$(X_1 + \cdots + X_n) \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$$

○第七章　多次元の確率分布

◎同時確率分布

同時に 2 つの離散型確率変数 X, Y を考える。X のとりうる値を $x_1 x_2, \dots, x_k$ 、Y のとりうる値を $y_1, y_2 \dots y_l$ とする。このとき、$p_{X,Y}(x_i, y_i) = P(X = x_i, Y = y_i)$ は X と Y を同時に考えた時の確率分布を定める。この 2 変数の関数 $p_{X,Y}(\cdot,\cdot)$ を X と Y の同時確率関数あるいは同時確率分布という。また、X と Y の同時確率分布がわかっていれば、X の確率分布を求めることは簡単である（Y を無視すればよい）。これを X の周辺確率分布とも呼ぶ。これは Y も同様である。

同時確率分布は下のように、表にまとめると便利である。

同時確率分布		Y		X の周辺確率分布
		40	50	
X.	20	0.3	0.1	0.4
	30	0.1	0.5	0.6
Y の周辺確率分布		0.4	0.6	

X と Y が確率変数であるとき、X + Y も確率変数である。また、2 つの確率変数 X,Y に対して次が成り立つ。確率が 3 つ以上の場合も同様である：

$$E(X + Y) = E(X) + E(Y)$$

◎共分散と相関係数

●共分散

以下、$E(X) = \mu_X$ ，$E(Y) = \mu_Y$ とすると、2 つの確率変数 X と Y の共分散は次のように定義される：

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\mu_X)(x_i-\mu_Y)\,P(x_i,y_i)$$

Cov(X,Y)は X と Y の直線的な関係を測る特性値である。Cov(X,Y) > 0 ならば、X と Y は同じ大小の方向に動く確率が大きい。つまり、X が大きい値をとるときに Y も大きい値をとり、逆に、X が小さい値をとるときには Y も小さい値をとる。Cov(X,Y) < 0 ならば反対方向に動く確率が大きい。

共分散を実際に計算するときには次の公式が成り立つ：

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

また、X+Y の分散については、次の式が成り立つ：

$$V(X+Y) = V(X) + 2\mathrm{Cov}(X,Y) + V(Y)$$

●相関係数

共分散は X と Y の直線的結びつきを測るが、単位に依存するため、直線的関係が強いかどうかを判断するときには不便である。そこで、強弱を判断するために相関係数を以下のように定義する：

$$\rho_{X,Y} = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\mathrm{Cov}\,(X,Y)}{\sigma(X)\sigma(Y)}$$

これは、第三章で定義された相関係数と同じ性質をもつ。

◎確率変数の独立性

2 つの離散型変数 X,Y のとる値がそれぞれ $x_1,\dots,x_k,y_1,\dots,y_l$ であるとする。このとき、すべての $i=1,\dots,k$ と $j=1,\dots,l$ に対して、$P(x_i,y_j) = P(x_i)P(y_i)$ が成り立つとき、X と Y は互いに独立であるという。X と Y が独立ならば、次が成り立つ：

$$E(XY) = E(X)E(Y)$$

$$\mathrm{Cov}(X,Y) = 0$$

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

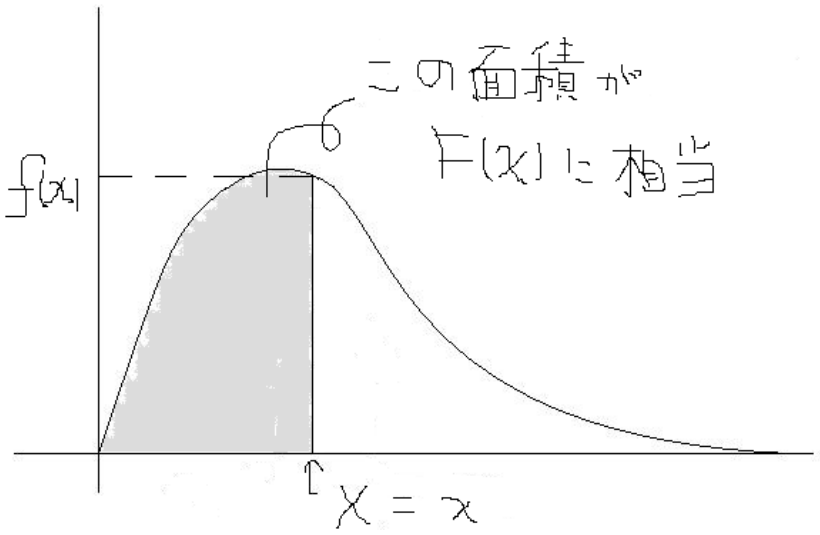
これは、確率変数が 3 つ以上の場合でも同様である。

○第八章　大数の法則と中心極限定理

◎累積分布関数

確率変数 X と実数 x に対して、x 以下の確率 $F(x) = P(X \leq x)$ を X の累積分布関数と呼ぶ。確率密度関数 f(x)が $X = x$ である場合の確率の値を示すのに対し、累積分布関数 F(x)は $X = x$ 以下である場合の確率の値を示す。その定義から、f(x)を確率密度関数として次の式が成り立つ：

$$F(x) = \int_{-\infty}^{x} f(x)dx$$



《累積分布関数の性質》

(a) 広義単調増加　　$x_1 < x_2$ のとき $F(x_1) \leq F(x_2)$

(b) $x \to \infty$ のとき $F(x) \to 1$

　　$x \to -\infty$ のとき $F(x) \to 0$

(c) 右連続　　各点 x で、$\varepsilon \downarrow 0$ のとき $F(x+\varepsilon) \to F(x)$

◎大数の法則と中心極限定理

●大数の法則

$X_1, X_2, \ldots, X_n$ は互いに独立で同一の確率分布に従うとする。このとき、任意の $\varepsilon > 0$ に対して、次が成り立つ：

$$P(|\overline{X} - \mu| > \varepsilon) \to 1 \ (n \to \infty)$$

つまり、試行の回数が十分大きければ、期待値およびその周辺が発生する確率が限りなく１に近づく、ということ。施行の回数が大きいほど、実際の確率が理論上の確率に近づくということを保証している。

●中心極限定理

$X_1, X_2, \ldots, X_n$ は互いに独立で同一の確率分布に従い、平均は $E(X) = \mu$,分散は $V(X) = \sigma^2$ であるとする。このとき、次が成り立つ：

$$P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) \to \Phi(x) \ \ (n \to \infty)$$

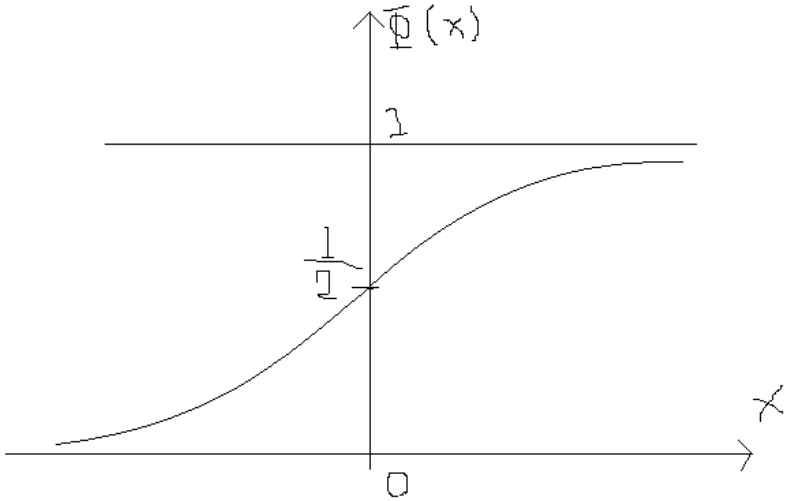
ただし、$\Phi(x)$は標準正規分布の累積分布関数であり、

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

で与えられる。

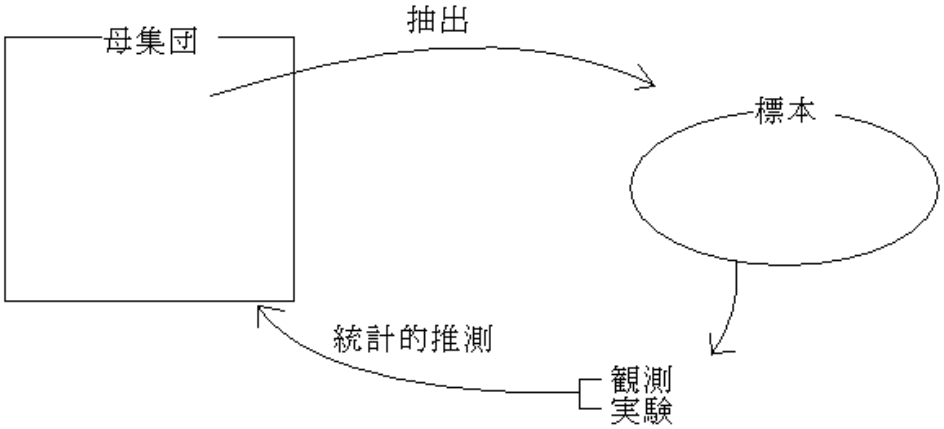
つまり、試行の回数が大きければ、和 $X_1 + X_2 + \cdots + X_n$ の確率分布の形状は正規分布とみなしてよい、ということ。さらに一次変換をすることで標準正規分布に近似することができる。

ちなみに標準正規分布 $N(0,1)$ の累積分布関数のグラフは下図のようになる。



○第九章　標本分布

実際にある集団について知りたいとき、全体からではなく部分から推量することがある。それが標本抽出による調査である。これは、集団全体（＝母集団）からランダムに標本を選び、それらの標本に対しての観測や実験により得られたデータから集団全体を推測する方法である。



ここで留意すべきことは、直接に調査するのは標本であるが、知ろうとするのは母集団の分布についてであるということである。

◎標本分布

標本から計算される量の確率分布を標本分布と呼ぶ。その中で母集団の特性を表すものとして重要なのが母平均μと母分散σ^2である。特に母集団が正規分布ならば、この二つで母集団分布の特性を完全に表現できる。母平均μと母分散σ^2を調べるためには、それぞれ標本平均と標本分散が指標となる。

●標本平均

標本 $X_1, X_2, \ldots, X_n$ から計算された平均を標本平均と呼び、標本平均 $\overline{X}$ は次式で与えられる：

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

標本平均 $\overline{X}$ は期待値が母平均 μ に一致する：

$$E(\overline{X}) = \mu$$

$\overline{X} = \mu$であるわけではないことに注意。

●標本分散

標本分散 s^2 は次のように定義される：

$$s^2 = \frac{1}{n-1}\{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\}$$

標本分散 s^2 は期待値が母分散σ^2に一致する：

$$E(s^2) = \sigma^2$$

この値は n の大きさに関わらず言っていだから、この s^2 を特に不偏分散と呼ぶ。ここで注意すべきは、標本分散 s^2 は n-1 で割ったものであるということである。これは標本分散の期待値が母分散に一致するようにするためである。n で割ったものも標本分散ではあるが、n の値に左右されるので不偏分散にはなりえない。

○第十章　正規分布からの標本

◎正規分布に関連する確率分布

ここでは、応用上重要な 2 つの確率分布、χ^2(カイ二乗)分布と t 分布について説明する。

●χ^2分布

χ^2分布は次のように定義される：

$Z_1, Z_2, \ldots, Z_k$ は独立で、それぞれが標準正規分布 N(0,1)に従う確率変数である。いま、

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

とすると、確率変数Y の確率分布を自由度 k のχ^2分布と呼び、$\chi^2(k)$ で表す。

●t 分布

t 分布は次のように定義される：

Z が標準正規分布 N(0,1)に従い、Y が自由度 k のχ^2分布$\chi^2(k)$ に従うとする。
さらに Z と Y は互いに独立であるとし、確率変数 t を

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

と定義する。このとき t の確率分布を自由度 k の t 分布と呼び、t(k)で表す。

《t 分布の特徴》

(1) t 分布の確率密度関数は 0 を中心に左右対称である。

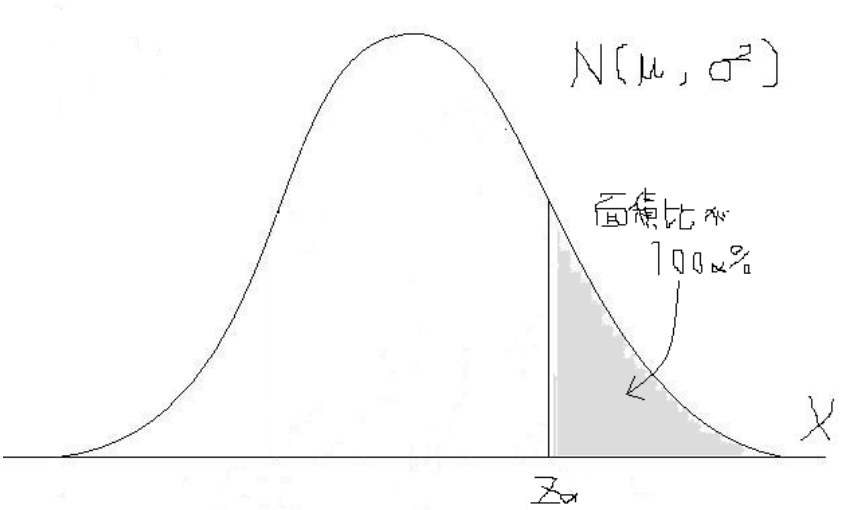
(2) 自由度 k の t 分布の確率密度関数は標準正規分布 N(0,1)の確率密度関数よりもすそが厚い

(3) 自由度 k が小さいほど t 分布のすそ野が長い。また $k \to \infty$ のとき、t 分布の確率密度関数は N(0,1)の確率密度関数に近づく。

◎分散が既知のときの標本平均の標本分布

標本分散 σ が既知のとき、標本平均 μ を推定することができる。$X_1, X_2, \dots, X_n$はそれぞれ独立かつ $N(\mu, \sigma)$ に従うとし、$\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ とすると、$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ は標準正規分布 N(0,1)に従う。

ここで、標準正規分布で、その点より上側の確率が 100α %となる点を上側パーセント点といい、Z_α で表す。ちなみに $Z_{0.025} = 1.96$, $Z_{0.05} = 1.64$ は覚えておいて損はないそうです。



このとき、次の式が成り立つ：

$$P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 1-\alpha$$

$$\leftrightarrow P\left(\overline{X} - Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

すなわち、母平均 μ が区間 $\left[\overline{X} - Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \quad \overline{X} + Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$ に存在する確率が$1-\alpha$ である。この区間は、μ の信頼係数$1-\alpha$ の信頼区間と呼ばれる。

◎標本分散の標本分布

前述のとおり、標本分散は次のように定義される：

$$s^2 = \frac{1}{n-1}\{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\}$$

ここで、上式を変形すると、

$$\frac{(n-1)s^2}{\sigma^2} = \left(\frac{X_1-\overline{X}}{\sigma}\right)^2 + \left(\frac{X_2-\overline{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n-\overline{X}}{\sigma}\right)^2$$

となる。この左辺を Y とおけば、上式はχ^2分布の定義に他ならない。すなわち、$\frac{(n-1)s^2}{\sigma^2}$は自由度 n - 1 の$\chi^2$ 分布に従う。χ^2分布は教科書 P282・283 に与えられている。

◎分散が未知のときの標本平均の標本分布

現実には、母分散σ^2 の値も未知なことが多い。この場合 t 統計量を用いる。t は次の式で定義される：

$$t = \frac{\overline{X}-\mu}{\sqrt{s^2/n}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{(n-1)}}} = \frac{Z}{\sqrt{\frac{Y}{(n-1)}}}$$

$\overline{X}$ とs^2 は互いに独立であることから、t は自由度 n−1 の t 分布に従うことがわかる。

ここで、自由度 k の上側確率 100α %のパーセント点を$t_\alpha(n-1)$ と書く。

このとき、次の式が成り立つ：

$$P\left(-t_{\alpha/2} \le t \le t_{\alpha/2}\right) = 1 - \alpha$$

$$\leftrightarrow P\left(\overline{X} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

すなわち、母平均 μ が区間$\left[\overline{X} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \overline{X} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$に存在する確率が$1-\alpha$ である。

（ただし、上式ではすべて$t_{\alpha/2} = t_{\alpha/2}(n-1)$ とする）

このように、t 分布を持ちいれば、母分散がわからない場合でも母平均を推定することができる。