# 第9回：正規母集団からの標本分布論

尚　晋
大学院経済学研究科　助教

2025年6月10日

**今日のポイント**

1. 正規母集団からの標本分布論と各種統計量



## 1　正規母集団からの標本分布論

標本分布の理論は，母集団が正規分布である場合に理論的にも応用的にも扱いやすい．実際，多くの統計的手法では正規分布の仮定が前提となる．

ここで扱う $t$分布，$F$分布などの基本的な知識は，今やどの分野でもデータを扱う者にとって常識である．本章では定義を中心に紹介するが，実際の応用はあとの「推定」章・「仮説検定」章で扱う．

**正規母集団**とは，母集団分布が正規分布であるような母集団である．正規分布は最も基本的かつ扱いやすい確率分布であり，その前提は統計学において理論・応用の両面で重要である．

統計的推論では，正規母集団から抽出された標本 $X_1, X_2, \ldots, X_n$ に基づく統計量の分布(標本分布)を求める必要がある．この標本分布を扱う理論を **正規標本論**と呼ぶ．

### 1.1　観測，測定，測定誤差

1 本の鉛筆の長さを測るという測定一広い意味では「観測」一を考えてみよう．これは，一見単純だが，「観測」「測定」あるいは「誤差」という統計学の本質的要素を含んでおり，基本的な説明例として非常に適している．

歴史的にも，統計学の数理的基礎は，測定誤差の理論を徹底的に追求して，正規分布に行きついたカール・フリードリッヒ・ガウス(C.F.Gauss, 1777–1855)によって作られたものである．したがって，この測定例は正規標本論の導入として非常に有用である．

1本の鉛筆の長さを $n$ 回測定値がお互いに影響を与えないように，かつ，鉛筆そのものや測定条件自体が同一に保たれるように測定することを考え，その測定値 $X_1, X_2, \ldots, X_n$を標

本とする．これをもたらすものは，「測定」そのものであるから，これを母集団としよう(無限に行いうるので，無限母集団である).

各測定値 $X_i$ は以下のように表される：

$$X_i = \mu + \varepsilon_i \quad (i = 1, 2, \ldots, n) \tag{1}$$

ここで，

- $\mu$：鉛筆の真の長さは定数だが，真の値は分からない
- $\varepsilon_i$：$i$回目の測定誤差(確率変数)：誤差は何が出るかわからないランダムな変数，全て同一の確率分布を持っており，しかも，ばらばらに(つまり，独立に)出る．

ガウスの誤差理論は，測定誤差 $\varepsilon_i$ は，平均0，分散$\sigma^2$の正規分布に従うと仮定される：

$$\varepsilon_i \sim N(0, \sigma^2) \tag{2}$$

すなわち，誤差は確率変数であり，正も負もあるが，平均は 0 であり，さらに，これが重要であるが，精度のよい測定では誤差の分散(ばらつき)$\sigma^2$は小さく，悪い測定では$\sigma^2$は大きい.

0を中心(平均)として正規分布に従う誤差に，真の値$\mu$だけ加えれば，つまり，確率分布を$\mu$だけその位置を平行移動すれば，測定値 $X_1, X_2, \ldots, X_n$の確率分布がわかる.

即ち，各測定値 $X_i$ は平均$\mu$，分散$\sigma^2$の正規分布に従う：

$$X_i \sim N(\mu, \sigma^2) \tag{3}$$

ただし，$X_1, X_2, \ldots, X_n$ は互いに独立である.

**誤差理論の意義**：このように，測定値は「真の値＋誤差」としてモデル化され，誤差が正規分布に従うというガウスの仮定から，正規標本論が導かれる.

## 1.2 正規分布の性質

**(再掲+新しい内容)** 確率変数$X$が正規分布に従うとき，その密度関数は，$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$で，その平均(期待値)，分散は$E(X) = \mu, V(X) = \sigma^2$である．すなわち，正規分布は平均$\mu$と分散$\sigma^2$の二つの母数で決まる．密度関数$f(x)$は$\mu$に関して左右対称であり，メディアン，モードは平均と一致する.

(i) $X$ が正規分布 $N(\mu, \sigma^2)$に従っているとき，その線形変換 $Y = aX+b$は $N(a\mu+b, a^2\sigma^2)$に従う。

(ii) 独立な二つ以上の正規確率変数の和および差は正規確率変数である。即ち、$X, Y$が独立で，それぞれ$N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$に従っているとき，$X + Y$は$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$に従い、$X - Y$は$N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$に従う.

(iii) 標準化変数 $Z = (X - \mu)/\sigma$ は正規分布 $N(0, 1)$に従う。これを標準正規分布(standard normal distribution)という。いかなる正規分布の確率計算も標準正規分布に帰着する。標準正規分布については，確率密度関数は$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$と定義し、その累積分布関数を

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{4}$$

で定義され、累積分布関数で求めた確率は**下側確率**という。従って、$1 - \Phi(z)$を相補累積分布関数、または**上側確率**という。なお,$\Phi(-z) = 1 - \Phi(z)$の関係があるので，一般の統計学教科書の付表「正規分布表」は$P(z > 0)$(**上側確率**)の表だけが与えられている。これより，おもな区間の確率がよく知られている。

$$P(-k \leq Z \leq k) = P(Z \leq k) - P(Z < -k) = \Phi(k) - \Phi(-k) \tag{5}$$

$k = 1, 2, ...$としてみよう．

$$P(-1 \leq Z \leq 1) = 0.6827 \quad (1/3\text{の確率で}[-1,1]\text{の区間外に落ちる})$$
$$P(-2 \leq Z \leq 2) = 0.9545 \quad (1/20\text{の確率で}[-2,2]\text{の区間外に落ちる})$$
$$P(-3 \leq Z \leq 3) = 0.9973 \quad (3/1000\text{の確率で}[-3,3]\text{の区間外に落ちる})$$
$$P(-4 \leq Z \leq 4) = 0.9999 \quad (1/10000\text{の確率で}[-4,4]\text{の区間外に落ちる})$$

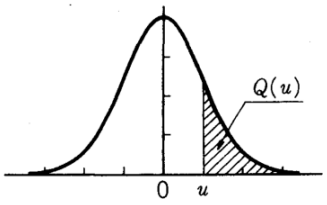
なお，$-3 \leq Z \leq 3$は，もとの $X$ でいえば，$\mu - 3\sigma \leq X \leq \mu + 3\sigma$に相当する．常識的にいえばこれで事実上すべて(全体の確率$= 1$) である．「事実上のすべて」の意味で，区間$[\mu - 3\sigma, \mu + 3\sigma]$を，3シグマ範囲という。

(iv) $Z \sim N(0,1)$，$P(Z > z_\alpha) = \alpha$となる時、即ちその点より上側の確率が$100\alpha\%$となる点、$z_\alpha$のことを$100\alpha\%$**パーセント点** percentage point という．

(v) 標準化変数$Z$が$z = 1.645, 1.96, 2.326$を超える確率は付表1より、$P(Z > z) = 0.05, 0.025, 0.01$、即ち、$5\%, 2.5\%, 1\%$．これらの$z(1.645, 1.96, 2.326)$を**標準正規分布の**$5\%, 2.5\%, 1\%$**臨界値**という。例えば、右端$2.5\%$臨界値$z_{0.025} = 1.96$、左端$2.5\%$臨界値$-z_{0.025} = -1.96$．$|Z|$の実現値が$1.96$を超える確率は$P(|Z| > 1.96) = P(Z > 1.96) + P(Z < -1.96) = 5\%$．$P(|Z| > z_\alpha)$を**両側確率**という．

**付表 1　正規分布表(上側確率)**

$$Q(u) = 1 - \Phi(u) = \int_u^\infty \phi(u)du$$



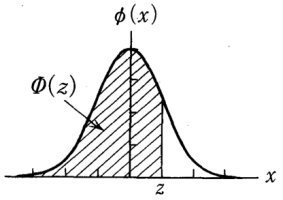




**図 10. 2**　標準正規分布の累積分布関数 $\Phi(x)$

## 1.3　母分散が既知場合の標本平均の標本分布

標本平均$\bar{X}$は，確率変数$X_1, X_2, \ldots, X_n$の和を$n$で割ったものである．$X_1, X_2, \ldots, X_n$は母集団分布と同一の分布に従う独立な確率変数であるから，各々独立で$N(\mu, \sigma^2)$に従う．

したがって，$\bar{X}$の分布は「独立な確率変数の和」と「統計量と標本分布」のところで述べたように，正規分布であり，平均は母集団分布と同一で$\mu$,分散は$\sigma^2/n$である．即ち、$\bar{X}$の分布は$N(\mu, \frac{\sigma^2}{n})$であり，$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$，それを標準化した

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{6}$$

は標準正規分布$N(0,1)$に従う．

- このことから$\sigma^2$がわかっていれば$\bar{X}$の標本分布は，標準正規分布$N(0,1)$を見ることに帰着する．
- $\bar{X}$の標準偏差は$\sigma^2/\sqrt{n}$である、このことから：
  - (i) $n$が増加するに従い，$\bar{X}$は母平均$\mu$のより正確な推定値となる．
  - (ii) 推定の誤差は$1/\sqrt{n}$のオーダーでしか減少しない．すなわち，母平均$\mu$を標本平均$\bar{X}$で推定する場合，推定の正確さを2倍にするためには$n$を4倍すればよい．なお，**「推定」**とは，今は，近似的な値とする，という位の意味である．
- 標本平均$\bar{X}$が正規分布に従うことも重要．即ち，測定値としての標本の各$X_i$が正規分布$N(\mu, \sigma^2/n)$に従うなら，$\bar{X}$は正規分布$N(\mu, \sigma^2/n)$に従っている．よって、$\bar{X}$も一つの測定値で、しかも、分散の小さい，単独の$X_i$よりもすぐれた測定値である．**標本の平均$\bar{X}$をとることの意味は，測定の正確さを単独の場合よりも増すことにある．**

例：鉛筆の長さの測定の例で，母平均$\mu = 18.0$ と仮定してあったとして，10回測定し，$X = 18.06$ を得たとすると，この鉛筆の真の長さが $\mu = 18.0$ という仮定は維持できるか．ただし，$\sigma = 0.02$ であることがわかっているとする．

そこで標準化すると$Z = \frac{18.06-18.0}{0.02/\sqrt{10}} = 9.49$となり，$z_{0.00005} \approx 3.89$ であるから，ましてこのような大きな値は，標準正規分布表を見るまでもなく，ほとんど絶無の可能性しかない．したがって，$\mu = 18.0$という仮定は正しくない，と結論づけるべきです．

## 1.4　標本分散の標本分布

標本分散(不偏分散の方をとる)は母分散 $\sigma^2$ に対応するものである。標本(不偏)分散は定義によって，標本 $X_1, X_2, \ldots, X_n$ から，

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{7}$$

によって計算される．正規母集団を仮定しなくても，「標本分布」節の「標本分散」のところで見たように$E(s^2) = \sigma^2$であることはわかっているが，さらに正規母集団を仮定すれば，標本分散 $s^2$ の標本分布を求めることができ，母分散 $\sigma^2$ についての統計的推測を行うことが可能になる．

測定の例で，ガウスの誤差理論によれば、各測定誤差の分散の大きさが $\sigma^2$ であるから，

$$V(e_i) = \sigma^2 \quad (i = 1, 2, \ldots, n) \tag{8}$$

となり，さらに誤差理論によって、$E(e_i) = 0$ であるから，

$$E(e_i^2) = \sigma^2 \quad (i = 1, 2, \ldots, n) \tag{9}$$

となる($V(e_i) = E(e_i^2) - [E(e_i)]^2 = \sigma^2$から)．測定誤差の二乗 $e_i^2$ が表れているが，$n$ 回の測定なら，

$$e_1^2 + e_2^2 + \cdots + e_n^2 \tag{10}$$

が**標本分散 $s^2$ の標本分布を求める**のに基礎的な量になることがわかる．

一般に，$e_1, e_2, \ldots, e_n$ は独立で，標準正規分布 $N(0, 1)$ に従っていると仮定する．

**$\chi^2$分布：**

$Z_1, Z_2, \ldots, Z_k$ を独立な標準正規分布 $N(0, 1)$に従う確率変数とする．いま，

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \tag{11}$$

とすると，確率変数 $\chi^2$ が従う確率分布を，**自由度 $k$ の$\chi^2$分布**という．ここの自由度は，独立な標準正規確率変数の二乗をいくつ加えたかを表す．また，自由度 $k$ のカイ二乗分布を $\chi^2(k)$ で記す($\chi^2$ は「カイ二乗」と読む)．

カイ二乗分布[1](Chi-square distribution)は，正規標本論で標本分散を扱うときに必ず関係してくる確率分布である．自由度 $k$ のカイ二乗分布の上側確率が $\alpha$(図 10.4 の斜線の部分)となる値(点)を $\chi^2_\alpha(k)$ と書き，**上側確率 $100\alpha\%$ のパーセント点**という．付表3の$\chi^2$分布表(パーセント点) に $\alpha$ と $\chi^2_\alpha(k)$ の関係が与えられている．たとえば，$k = 5$ の場合，$\chi^2_{0.05}(5) = 11.070$，すなわち$P(\chi^2 > 11.070) = 0.05$である．

[1] $\chi^2$分布の確率密度関数は$for\, x \geq 0, f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}; for\, x < 0, f(x) = 0.$ただし$\Gamma(\cdot)$はガンマ関数，$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx, \quad \alpha > 0$

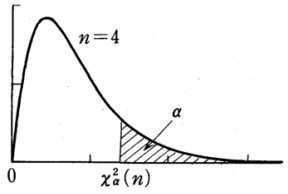
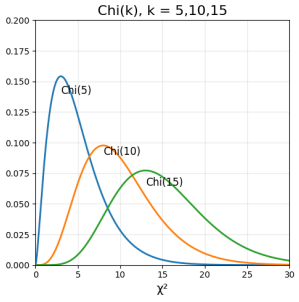



**図 10.4**　自由度 $k$ の $\chi^2$ 分布 $\chi^2(k)$ のパーセント点
$\chi^2$ 分布は，密度関数 $f(x)$ が $x>0$ の部分でのみ正の値をとるもので，自由度 $k$ が大きくなるに従い $f(x)$ は右方向へ移動する．$\chi_\alpha^2(k)$ は $P(\chi^2>\chi_\alpha^2(k))=\alpha$ で定義される．なお，$\chi^2(k)$ はガンマ分布 $Ga(k/2,1/2)$ のことである．




**標本分散に関する$\chi^2$統計量：**

カイ二乗分布を用いると，正規母集団からの標本 $X_1, X_2, \ldots, X_n$ に基づく標本分散 $s^2$ の標本分布は，次のようにまとめられる．標本分散を$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$とするとき，標本分散の統計量，カイ二乗統計量

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{12}$$

は自由度 $n-1$ のカイ二乗分布 $\chi^2(n-1)$ に従う．$(\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1))$

例：ある正規母集団は母平均が$\mu = 50$,母分散が$\sigma^2 = 25$であるとする．これから大きさ$n = 10$の標本をとったとき，標本分散$s^2$が50を越える確率はどれほどか.

不等式$s^2 > 50$を，$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$に変形してみると，自由度は$n - 1 = 9$となり,

$$P(s^2 > 50) = P(\chi^2 > \frac{9 \cdot 50}{25}) = P(\chi^2 > 18) = 0.038 \tag{13}$$

すなわち，母分散の2倍を越える標本分散は，出やすいわけではない．このように，標本分散の大きさを見るときには，その値の動きやすさ(標本分散の変動)に注意すべきである.

## 1.5　母分散が未知場合の標本平均の標本分布

1.3節では母分散$\sigma$が既知の場合の標本平均$\bar{X}$の標本分布を扱ったが、現実には母分散が既知であることはほとんどないため、この仮定は非現実的である。母分散が未知の場合、正確な標本分布は求められない。したがって、母分散$\sigma^2$の代わりに標本分散$s^2$を用いる方法が必要となる。ただし、この場合、標本平均 $\bar{X}$ の標本分布は $\mathcal{N}(\mu, \sigma^2/n)$ に従うので，標準化統計量$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$は母標準偏差 $\sigma$ を含むため、$\sigma$ が未知ならこの $Z$ は計算できない。

そこで，母分散 $\sigma^2$ を標本分散 $s^2$ で代用したより現実的な統計量，スチューデントの$t$統計量 Student's t-statistic

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \tag{14}$$

を定義する．この統計量は標準正規分布には従わない.

この確率分布はどうなるか考えてみると，分子と分母の形から，

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \Big/ \sqrt{\frac{s^2}{\sigma^2}} \\ &= \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}} \end{aligned} \tag{15}$$

の形に変形できる．分子$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$は標準正規分布$\mathcal{N}(0,1)$ に，分母$\sqrt{(n-1)s^2/\sigma^2}$は自由度$n-1$の$\chi^2$分布$\chi^2(n-1)$に従う．これらの比になっている.

従って、

$$t = \frac{Z}{\sqrt{\chi^2/(n-1)}} \tag{16}$$

正規分布の密度関数の計算から，$\bar{X}$と$s^2$は互いに独立であるので，$\mathcal{N}(0,1)$ と $\chi^2(n-1)$ の組み合わせから，$t$の密度関数が求まる．

正規標本論では，このような分布を，$t$分布と呼び，以下のように定義する．

**$t$分布：**

次の条件を満たす確率変数 $t$ は自由度 $k$ の $t$ 分布 $t(k)$ に従う：

- $Z \sim \mathcal{N}(0,1)$(標準正規分布)
- $Y \sim \chi^2(k)$(自由度 $k$ のカイ二乗分布)
- $Z$ と $Y$ は独立

このとき，確率変数 $t$

$$t = \frac{Z}{\sqrt{Y/k}} \tag{17}$$

と定義される．$t$が従う確率分布を自由度$k$の$t$分布(あるいは，スチューデントの$t$分布)という．自由度$k$の$t$分布を$t(k)$で表す．

**$t$統計量：**

先に式14で定義した統計量

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \tag{18}$$

は自由度 $n-1$ の t 分布に従う．また，$\bar{X}$の標準偏差$s/\sqrt{n}$ を，標本平均の**標準誤差(standard error)**という．

“スチューデント” (Student) は，この分布を導入したゴセット(William Gosset, 1876–1937)が使った仮名である．

- $t$分布の密度関数 $f(x)$ は $x = 0$ に対して対称[2](つまり偶関数)であり，標準正規分布とよく似ているが，自由度 $k$ が小さい場合は裾がやや広い．
- 自由度は1 の $t$分布をコーシー分布という．
- 自由度が大きい場合(例えば$k = 30$)を超えると標準正規分布とほとんど区別がつかなくなる．特に，$k \to \infty$ のとき，$t$分布は標準正規分布 $\mathcal{N}(0,1)$ に一致する．
  これは，カイ二乗分布の平均が自由度 $k$ に等しいことと，大数の法則によって $Y/k \to 1$ となることに起因する．また，**大標本**$n \to \infty$のときには，母集団と標本の区別がほぼなくなり，$s^2 \fallingdotseq \sigma^2$ となるため．この意味で，$t$分布は**小標本**の厳密な標本分布ということができる．
- 自由度 $k$ の $t$ 分布において，上側確率 $100\alpha$のパーセント点を $t_\alpha(k)$ と書き，付表2によりその値を求めることができる．自由度 $k$ が大きくなるにつれて，$t_\alpha(k)$ は同じ $\alpha$ に対して小さくなる．

[2] $t$分布の確率密度関数は$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$ である

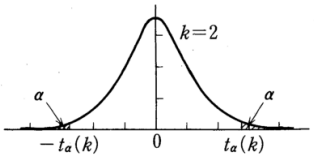
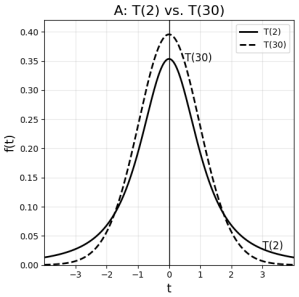



**図 10.5**　自由度 $k$ の $t$ 分布 $t(k)$ とパーセント点
母平均の仮説検定，推定論などにおいて，実際に用いられるのは，標準正規分布より，この $t$ 分布である．二つは酷似するが，大標本($n\to\infty$)のときは，一致する．$t_\alpha(k)$ は $\mathrm{P}(t>t_\alpha(k))=\alpha$ で定義される．なお，両側(一の側)も入れて $\alpha$ とする流儀もある．




## 1.6　2標本問題

身長の分布を考えた場合，一定地域で同一学年の男子の身長と女子の身長ではその母集団分布に図10.6のように大きな差があるかもしれない．このような場合，これまでのように一つの正規母集団から標本を抽出したと仮定して分析することは適当でない．男子の身長と女子の身長など明らかに異なる2種の標本による2母集団の比較を扱う問題を**2標本問題(two-sample problem)**という．
2標本問題では，二つの母集団から別々に標本を抽出したと考える．ここでは，大きさ $m$ の第一の標本 $X_1, X_2, \ldots, X_m$ を母集団分布 $N(\mu_1, \sigma_1^2)$ から，大きさ $n$ の第二の標本 $Y_1, Y_2, \ldots, Y_n$ を母集団分布 $N(\mu_2, \sigma_2^2)$ から独立に抽出した場合の標本分布について説明する．

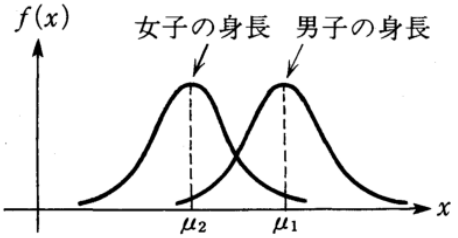



**図 10.6**　2母集団分布の比較
統計的推定，検定ではもっともよく扱われる問題の一つである．

### 1.6.1　標本平均の差の標本分布

2標本問題では，二つの母平均の差 $\mu_1-\mu_2$ を分析することがしばしば重要となる．例えば男女別の賃金を比較する問題で，同一条件下で格差がなければ $\mu_1-\mu_2=0$ となるはずであるし，男女格差があれば0とはならないであろう．

このような母平均の差 $\mu_1-\mu_2$ を分析するには，対応する標本平均

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{m}X_i,\quad \bar{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_j$$

の差 $\bar{X}-\bar{Y}$ を調べればよい．

実際、これらの標本分布はもちろん双方の母分散にも依存するので、ここでは下記の3通りに分けて考える．

**(a) 母分散 $\sigma_1^2, \sigma_2^2$ が既知のとき**　この場合，

$$\bar{X}\sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right),\quad \bar{Y}\sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$$

より，$\bar{X}-\bar{Y}$ は独立な正規分布の差なので，

$$\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right) \tag{19}$$

となる．したがって，標準化して

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \tag{20}$$

は標準正規分布 $N(0,1)$ に従う．

**例：** 男子の身長 $\mu_1 = 172.3$, $\sigma_1^2 = 30$，女子の身長 $\mu_2 = 160.2$, $\sigma_2^2 = 25$，男子10人，女子15人を抽出したとき，$\bar{X} - \bar{Y} = 12.1, \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = \frac{30}{10} + \frac{25}{15} = 4.667$より：

$$\bar{X} - \bar{Y} \sim N(12.1, 4.667)$$

**(b) 母分散は未知であるが等しいとき**　母分散は未知であるが，事前に等しいことがわかっており，$\sigma_1^2 = \sigma_2^2 = \sigma^2$ であると仮定できるときは，標本分散を母分散 $\sigma^2$ の代わりに使って $X - Y$ の分散を求める．

$X-Y$ の分布は(正規分布の性質より)正規分布 $N(\mu_1-\mu_2, (\frac{1}{m}+\frac{1}{n})\sigma^2)$ であるが，$\sigma^2$ は未知なので，母分散が共通の二つの標本を合併したものから，次の **合併分散(pooled variance)** で推定する：

$$s^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{X}) + \sum_{j=1}^{n}(Y_j - \bar{Y})}{m+n-2} = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2} \tag{21}$$

ここで，$s_X^2, s_Y^2$ は各標本の標本分散である．このように標本分散を定義すると，次の2点が知られている：

(i) $(m+n-2)\frac{s^2}{\sigma^2}$ は，自由度 $m+n-2$ の $\chi^2(m+n-2)$ 分布に従う．(「$\chi^2$統計量」より)
(ii) $s^2$ と $\bar{X} - \bar{Y}$ は独立である．

ここで，次のように標準化すると：

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{m} + \frac{1}{n})\sigma^2}} \tag{22}$$

$Z$ の分布は標準正規分布 $N(0,1)$ に従う．

**2標本 $t$ 統計量(two-sample $t$ statistic)**は

$$t = \frac{Z}{\sqrt{s^2/\sigma^2}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{m} + \frac{1}{n}}} \tag{23}$$

で定義され、自由度 $m+n-2$ の $t(m+n-2)$ 分布に従う．この統計量は，次々章で述べる **2標本 $t$ 検定** に主に用いられる．

**(c) 母分散が未知であり等しいとは限らないとき**　この場合は正確な分布は求めにくいが，近似的に分布を求める**ウェルチ(Welch)の近似法** が用いられる．

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \tag{24}$$

このとき，$t$ の自由度 $\nu$ は以下で近似される：

$$\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}} \tag{25}$$

この自由度に最も近い整数$\nu^*$を用いて，$t(\nu^*)$ 分布に従うと近似される．

### 1.6.2　標本分散の比の標本分布

標本平均の差 $X - Y$ の分布を求める場合，二つの母集団分布の分散 $\sigma_1^2, \sigma_2^2$ が等しいかどうかによってその分布を求める方法が異なっていた．

そのために，二つの標本分散 $s_1^2, s_2^2$ の相対的な比 $s_1^2/s_2^2$ を調べることが手がかりとなる．なぜなら，$s_1^2/s_2^2 \approx 1$ のとき，母分散についても $\sigma_1^2/\sigma_2^2 \approx 1$ であると推測されるからである．

二つの分散は独立で，それぞれが $\chi^2$ 分布に従うため，その比の確率分布として $F$**分布**が導入される．

**$F$分布の定義：**

確率変数 $U, V$ が次の条件を満たすとする：

(a) $U$ は自由度 $k_1$ の $\chi^2$ 分布 $\chi^2(k_1)$ に従う，
(b) $V$ は自由度 $k_2$ の $\chi^2$ 分布 $\chi^2(k_2)$ に従う，
(c) $U$ と $V$ は独立である．

ここで，$U$と$V$をそれぞれの自由度で割って調整した後にとった比，すなわちフィッシャーの分散比を

$$F = \frac{U/k_1}{V/k_2} \tag{26}$$

と定義すると，$F$が従う確率分布を**自由度** $(k_1, k_2)$ **の$F$分布** といい，$F(k_1, k_2)$ と表す．

標本分散 $s_1^2, s_2^2$ について，以下が知られている：

(i) $(m-1)s_1^2/\sigma_1^2 \sim \chi^2(m-1)$,
(ii) $(n-1)s_2^2/\sigma_2^2 \sim \chi^2(n-1)$,
(iii) $s_1^2$ と $s_2^2$ は独立である．

**標本分散の比ー$F$統計量：**

したがって，$F$分布の定義から$F$統計量：

$$F = \frac{(m-1)s_1^2/\sigma_1^2}{(n-1)s_2^2/\sigma_2^2} \cdot \frac{n-1}{m-1} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{27}$$

は，自由度 $(m-1, n-1)$ の $F$分布 $F(m-1, n-1)$ に従う．

ここに特に重要なのは，二つの母分散が等しいときであり、即ち$\sigma_1^2 = \sigma_2^2 = \sigma^2$とおけば，$F$分布文字通り標本の分散比 variance ratio

$$F = \frac{s_1^2}{s_2^2} \tag{28}$$

の標本分布となる .F 分布はしばしばこの形で用いられる．

**t分布と$F$分布の関係**　$t$ が自由度 $k$ の $t$ 分布 $t(k)$ に従うとき，$t^2$ は $F(1, k)$ に従う．

**パーセント点と$F$分布の性質**　自由度 $(k_1, k_2)$ の $F$分布 $F(k_1, k_2)$ において，上側確率が $\alpha$ となる値を **上側確率** $100\alpha\%$ **のパーセント点** といい，$F_\alpha(k_1, k_2)$ と表す(付表4参照)．

定義から $F \sim F(k_1, k_2)$ のとき，$\frac{1}{F} \sim F(k_2, k_1)$が成り立つので，

$$F_{1-\alpha}(k_1, k_2) = \frac{1}{F_\alpha(k_2, k_1)} \tag{29}$$

が得られる．

例えば、$F \sim F(3, 5)$ のとき，$1/F \sim F(5, 3)$ に従い，$F_{0.05}(5, 3) = 9.013$ であるから，

$$P(1/F > 9.013) = 0.05 \Rightarrow P(F < 0.110) = 0.05$$

すなわち，$F_{0.95}(3, 5) = 0.110$となる．

**標本分散の比の応用例**　仮定：母分散が同一の正規母集団から，$m = 10, n = 15$ の標本を抽出したとする．この仮定のもとでは，$F = \frac{s_1^2}{s_2^2}$であるから，二つの標本分散について，$s_1^2$ が $s_2^2$ の2倍以上となる確率は，$P(F > 2) \approx 0.1183$となる「scipy.stats.f.sf(Fvalue, df1, df2)で計算できる」．したがって，標本において $s_1^2$ が $s_2^2$ に対して2倍以上異なっていても，母集団において母分散が等しいと考えられる確率は，約10% 程度存在することになる．