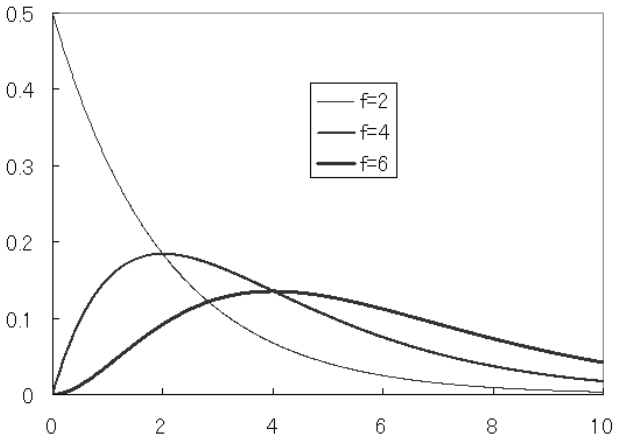
# 第８回　カイ二乗分布，Ｆ分布とその応用

## Ａ．カイ二乗分布

１．カイ二乗分布（$\chi^2$分布）

母分散に関する検定と推定を行うときはカイ二乗分布（$\chi^2$分布）を用いる．

カイ二乗分布とは正規分布する母集団の母分散が既知のときの標本分散に関する分布である．カイ二乗分布は自由度（=標本数−1）だけで決まり，σ は関与しない．カイ二乗分布は正規分布やｔ分布と異なり，**左右対称ではない**．




カイ二乗分布：

$x_1, x_2, \cdots, x_n$が互いに独立に$N(\mu, \sigma^2)$に従うとき

$\chi^2 = \dfrac{S}{\sigma^2}$は自由度$n-1$の$\chi^2$分布に従う

２．カイ二乗分布による母分散の推定

①　母分散の点推定

点推定の場合，標本分散をそのまま母分散の点推定値とする．

例：トマト 12 個を計ると右のような値を得た（単位：g）．母分散，母標準偏差を点推定せよ．

| | A | B | C |
|---|---|---|---|
| 1 | | | |
| 2 | | トマトの重さ(g) | |
| 3 | | 100.0 | |
| 4 | | 110.0 | |
| 5 | | 105.0 | |
| 6 | | 95.0 | |
| 7 | | 98.0 | |
| 8 | | 118.0 | |
| 9 | | 103.0 | |
| 10 | | 92.0 | |
| 11 | | 100.0 | |
| 12 | | 94.0 | |
| 13 | | 110.0 | |
| 14 | | 105.0 | |
| 15 | | | |
| 16 | 平均 | 102.5 | =AVERAGE(B3:B14) |
| 17 | 分散 | 57.90909 | =VAR(B3:B14) |
| 18 | 標準偏差 | 7.609802 | =STDEV(B3:B14) |

母分散$\sigma^2 = V = 57.91g^2$

母標準偏差$\sigma = SD = 7.6g$

②　母分散の区間推定

母分散の p%信頼区間はエクセルの関数を使うと以下の式で計算できる．

$$\frac{DEVSQ(B1:B10)}{CHIINV(\frac{100-p}{200}, n-1)} \le \sigma^2 \le \frac{DEVSQ(B1:B10)}{CHIINV(\frac{100+p}{200}, n-1)}$$

DEVSQ は平方和（$S = V \times (n-1)$），　$V$ は標本分散，　$n$ は標本数である．

例：トマト 12 個を計ると右下のような値を得た（単位：g）． 95%信頼区間をつけて母分散，母標準偏差を区間推定せよ．

95%信頼区間をつけた区間推定は，エクセルでは，以下のように計算できる．

$$\frac{DEVSQ(B3:B14)}{CHIINV(\frac{100-95}{200},12-1)} \le \sigma^2 \le \frac{DEVSQ(B3:B14)}{CHIINV(\frac{100+95}{200},12-1)}$$

母分散の区間推定

$$29.06 \le \sigma^2 \le 166.94(g^2)$$

母標準偏差の区間推定

$$5.39 \le \sigma \le 12.92(g)$$

このようにカイ二乗分布は左右対称ではないので，点推定値（母分散 57.91g², 母標準偏差 7.61g）が区間推定のまん中には来ない．

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | トマトの重さ(g) | | |
| 3 | | 100.0 | | |
| 4 | | 110.0 | | |
| 5 | | 105.0 | | |
| 6 | | 95.0 | | |
| 7 | | 98.0 | | |
| 8 | | 118.0 | | |
| 9 | | 103.0 | | |
| 10 | | 92.0 | | |
| 11 | | 100.0 | | |
| 12 | | 94.0 | | |
| 13 | | 110.0 | | |
| 14 | | 105.0 | | |
| 15 | | | | |
| 16 | 平均 | 102.5 | =AVERAGE(B3:B14) | |
| 17 | 分散 | 57.90909 | =VAR(B3:B14) | |
| 18 | 標準偏差 | 7.609802 | =STDEV(B3:B14) | |
| 19 | 平方和 | 637 | =DEVSQ(B3:B14) | |
| 20 | | | | |
| 21 | 分散95%上 | 166.94 | =DEVSQ(B3:B14)/CHIINV(0.975,11) | |
| 22 | 分散95%下 | 29.06019 | =DEVSQ(B3:B14)/CHIINV(0.025,11) | |
| 23 | 標準偏差95%上 | 12.92053 | =SQRT(B21) | |
| 24 | 標準偏差95%下 | 5.39075 | =SQRT(B22) | |

練習：K牧場の牛の乳脂肪率について 12 頭を無作為に調査した結果，

7.02, 7.03, 6.82, 7.08, 7.13, 6.92, 6.87, 7.02, 6.97, 7.08, 7.19, 7.15（%）であった．

K 牧場の牛の乳脂肪率の母標準偏差を 95%信頼区間をつけて区間推定せよ．

３．カイ二乗分布による検定

例：K牧場の牛の乳脂肪率の標準偏差は 0.07%であった．新しい飼育法を導入したが，乳脂肪率のばらつきが変化したかを知りたい．12 頭を無作為に調査した結果，

7.02, 7.03, 6.82, 7.08, 7.13, 6.92, 6.87, 7.02, 6.97, 7.08, 7.19, 7.15 であった．検定せよ．

① 帰無仮説と対立仮説の設定

帰無仮説：$H_0$： $\sigma^2 = \sigma_0^2 = 0.07^2$　乳脂肪率の母分散は変わらない．

対立仮説：$H_1$： $\sigma^2 \neq \sigma_0^2$　乳脂肪率の母分散は変わった．

ここで，$\sigma^2$ は新しい飼育法を導入したあとの母集団の分散，$\sigma_0^2$ は導入前の母集団の分散（すなわち $0.07^2$）である．

ここでは有意水準を 1%に設定するとしよう．

② p-値の計算

帰無仮説が成り立つとしたら，今回の標本が得られる確率である p-値はエクセルでは以下の式で計算する．$\sigma_0^2$ は帰無仮説で扱う母分散であり，標本の分散ではないことに注意すること．以下の例ではデータはエクセルワークシートの B1 から B10 にあるとする．

１） ${\sigma_0}^2 < V$ の場合（ばらつきが大きくなった場合）

$$\text{p-値} = 2 \times CHIDIST(\frac{S}{\sigma_0^2}, n-1) = 2 \times CHIDIST(\frac{DEVSQ(B1:B10)}{\sigma_0^2}, n-1)$$

２） ${\sigma_0}^2 > V$ の場合（ばらつきが小さくなった場合）

$$\text{p-値} = 2 \times (1 - CHIDIST(\frac{S}{\sigma_0^2}, n-1)) = 2 \times (1 - CHIDIST(\frac{DEVSQ(B1:B10)}{\sigma_0^2}, n-1)$$

１），２）とも片側検定の場合は，「2×」を削除する．

③ 検定結果

今回の例では標本分散 $V = 0.012733 > 0.07^2$

$$\text{p-値} = 2 \times CHIDIST(\frac{0.140067}{0.07^2}, 12-1) = 0.005263$$

したがって，1%の有意水準で帰無仮説は棄却され，乳脂肪率のばらつきが変化したと結論できる．

なお 99%信頼区間をつけて，母分散を区間推定すると，

母分散の区間推定　$0.005235 \le \sigma^2 \le 0.053806$

母標準偏差の区間推定　$0.07235 \le \sigma \le 0.23196(\%)$

練習：A農場ではジャガイモの新品種を導入した．10 カ所の圃場で栽培したところ，以下のような収量だった．収量のばらつきが以前に栽培していた品種の標準偏差 25g/m² と異なるかを検定せよ．

| 番号 | 収量(g/m²) |
|---|---|
| 1 | 445 |
| 2 | 489 |
| 3 | 490 |
| 4 | 476 |
| 5 | 469 |
| 6 | 493 |
| 7 | 471 |
| 8 | 480 |
| 9 | 491 |
| 10 | 457 |

## B．カイ二乗検定の応用

カイ二乗検定はメンデル遺伝の分離比や，計数（比率）データの標本（群）の差の検定にも利用できる．イエスーノー，生－死など二者択一的なデータであるため範疇(category)データとも呼ばれる．今回はメンデル遺伝の分離比の検定についてだけ取り上げる．その他のデータの場合はプリントの最後に参考として記しておく．

以下の式を用いて，メンデル遺伝の分離比や，計数（比率）データの標本（群）の差の検定にカイ二乗検定を利用する．

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{観測値} - \text{期待値})^2}{\text{期待値}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \text{式（1）}$$

p−値＝$CHIDIST(\chi^2, f)$であり，自由度 $f$ はメンデル遺伝の場合，考慮する遺伝子座の数を $n$ とすると $f = 2^n - 1$ である．なお，分離比から適合しないかを検定するので，上側検定（片側検定の１種）になる．（下側検定は分離比に当てはまりすぎるということを意味する）

メンデル遺伝の分離についてカイ二乗検定を利用する具体的な方法を以下，説明する．

### １．１遺伝子座の場合

自由度が１の場合（メンデル遺伝の分離比では１つの遺伝子座しか考えないとき）は，$\chi^2$の値がやや高めに算出されるため以下のように補正する．絶対値|　　|はエクセルではABS( )である．

$$\chi^2_{adj} = \sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i} = \frac{(|\text{観測値}_A - \text{理論値}_A| - 0.5)^2}{\text{理論値}_A} + \frac{(|\text{観測値}_B - \text{理論値}_B| - 0.5)^2}{\text{理論値}_B}$$

p−値＝$CHIDIST(\chi^2_{adj}, 1)$　　式（2）

例：$F_1$のエンドウの交配から赤花 80，白花 30 を得た．3：1 に分離するかを検定せよ．

自由度が 1 なので，補正した式（2）を用いる．

帰無仮説は「分離比は 3：1 である」．一方，対立仮説は「分離比は 3：1 でない」

期待値は 3：1 に分離した場合にどうなるかであるから，赤花 82.5，白花 27.5 になる．したがって，

$$\chi^2_{adj} = \frac{(|80-82.5|-0.5)^2}{82.5} + \frac{(|30-27.5|-0.5)^2}{27.5} = \frac{(ABS(-2.5)-0.5)^2}{82.5} + \frac{(ABS(2.5)-0.5)^2}{27.5} = 0.194$$

検定：p−値$= CHIDIST(\chi^2_{adj},1) = CHIDIST(0.194,1) = 0.6597$

したがって，有意水準 5%で帰無仮説は棄却できず，分離比は 3：1 でないという有意な証拠はない．つまり分離比は 3：1 であると考えてよい．

練習：$F_1$のエンドウの交配から赤花 105，白花 15 を得た．3:1 に分離するかを検定せよ．

2．遺伝子座が 2 つ以上の場合

例：「花色赤色・草丈が高い×花色白色・草丈が低い」を交配した$F_1$はすべて花色赤色・草丈が高いとなった．$F_1$同士を交配した結果，下の表のような分離比を得た．これは 9:3:3:1 の分離比かどうかを検定する．

| 遺伝子型 | 表現型 | 観測値 | 分離比 | 期待値 |
|---|---|---|---|---|
| 赤一高一 | 赤色･草丈高い | 65 | 9 | 90 |
| 赤一低低 | 赤色･草丈低い | 50 | 3 | 30 |
| 白白高一 | 白色･草丈高い | 30 | 3 | 30 |
| 白白低低 | 白色･草丈低い | 15 | 1 | 10 |
| | | | | |
| | 合計 | 160 | 16 | |

帰無仮説：　分離比は 9:3:3:1 である，対立仮説：　分離比は 9:3:3:1 ではない

補正の必要のないときはエクセルで簡単に p−値を計算できる．

p−値$= CHITEST$(観測値，理論値)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 遺伝子型 | 表現型 | 観測値 | 分離比 | 期待値 | |
| 2 | 赤一高一 | 赤色･草丈高い | 65 | 9 | 90 | |
| 3 | 赤一低低 | 赤色･草丈低い | 50 | 3 | 30 | |
| 4 | 白白高一 | 白色･草丈高い | 30 | 3 | 30 | |
| 5 | 白白低低 | 白色･草丈低い | 15 | 1 | 10 | |
| 6 | | | | | | |
| 7 | | 合計 | 160 | 16 | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | CHITEST= | 4.493E-05 | =CHITEST(C2:C5,E2:E5) | | |

p-値は 0.00004493 となる．（E-05 とは$\times 10^{-5}$のこと）

p-値は 0.01 より小さいので，有意水準 1%で帰無仮説は棄却され，分離比は 9:3:3:1 とはいえないと結論できる．

練習：次のデータでは 9:3:3:1 に分離しているか？

| 遺伝子型 | 表現型 | 観測値 | 分離比 | 期待値 |
|---|---|---|---|---|
| 赤一高一 | 赤色・草丈 | 80 | 9 | 90 |
| 赤一低低 | 赤色・草丈 | 35 | 3 | 30 |
| 白白高一 | 白色・草丈 | 30 | 3 | 30 |
| 白白低低 | 白色・草丈 | 15 | 1 | 10 |
| | | | | |
| | 合計 | 160 | 16 | |

## C．F 分布

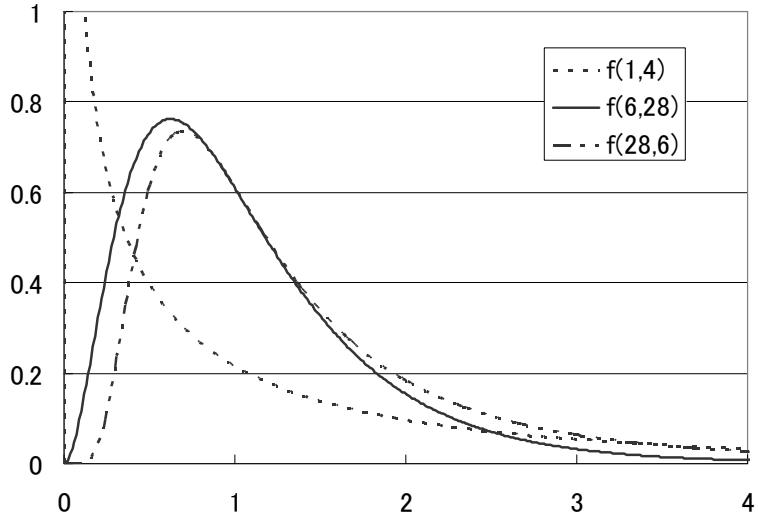
１．F 分布

２つの母集団から得られた標本の分散比はF分布で検定する．分散分析という手法はこのF分布によって行うので，F検定は重要な手法である．カイ二乗分布と違い，２つの標本を比較するので，２つの標本それぞれの自由度からF分布が決まる．

F 分布：$x_{11}, x_{12}, \cdots, x_{1n_1}$が互いに独立に$N(\mu_1, {\sigma_1}^2)$に従い，さらにそれらと独立に

$x_{21}, x_{22}, \cdots, x_{2n_2}$が互いに独立に$N(\mu_2, {\sigma_2}^2)$に従うとき，

$F = \dfrac{V_1/\sigma_1^2}{V_2/\sigma_2^2}$は自由度$(n_1 - 1, n_2 - 1)$の F 分布に従う．




F分布もカイ二乗分布と同じく，**左右対称ではない**.

２．Ｆ検定の手順

Ｆ分布に基づいて，２つの母集団の分散の比を推定することもできるが，あまり使わないので，ここでは述べない．

Ｆ検定は分散分析という手法の骨格をなす方法である．まずここでは基本的なＦ検定だけを練習しよう．

例：Ａ，Ｂ２種類の飼料を与えて一定期間飼育したハムスターの体重増加量を測定した結果，右のようなデータを得た（単位　g）．飼料による体重増加量のばらつきに差があるかを検定せよ．

| 飼料A | 飼料B |
|---|---|
| 73 | 71 |
| 74 | 68 |
| 70 | 73 |
| 77 | 71 |
| 72 | 71 |
| 72 | 72 |
| 69 | 69 |
| 73 | 72 |
| 72 | 68 |
| 75 | |

帰無仮説：$\sigma_A^2 = \sigma_B^2$　２つの標本の母分散は同じである．

飼料 A と飼料 B で飼育したときにハムスターの体重増加量の分散に差がない．

対立仮説：$\sigma_A^2 \neq \sigma_B^2$　２つの標本の母分散は異なる．

飼料 A と飼料 B で飼育したときにハムスターの体重増加量の分散に差がある．

帰無仮説が成り立つと仮定したときに今回のデータが得られる確率 P 値はエクセルの関数から，以下のように計算できる．

p-値＝*FTEST*(変数1の入力範囲，変数2の入力範囲)

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 飼料A | 飼料B | | |
| 2 | 73 | 71 | | |
| 3 | 74 | 68 | | |
| 4 | 70 | 73 | | |
| 5 | 77 | 71 | | |
| 6 | 72 | 71 | | |
| 7 | 72 | 72 | | |
| 8 | 69 | 69 | | |
| 9 | 73 | 72 | | |
| 10 | 72 | 68 | | |
| 11 | 75 | | | |
| 12 | | | | |
| 13 | P値 | 0.502413 | =FTEST(A2:A11,B2:B10) | |

検定の結果

p-値が 0.50 であるから，5%の有意水準で帰無仮説は棄却できない．したがって，5%の有意水準で飼料 A と飼料 B でハムスターの体重増加量のばらつきに差があるとはいえない．

練習　商社Ｓはタイでタマネギを栽培している．スーパーの基準は厳しく，ほとんど同じ直径(cm)のタマネギでないと納入させてくれない．今まで栽培していた品種Ａに代わり，多収で耐病性の強い品種Ｂを導入したいが直径のばらつきは品種Ａと同じではないのかもしれない．無作為に選んだ標本から右のようなデータを得た．品種Ｂのばらつきは品種Ａと異なるのかを有意水準 5%で両側検定せよ．

| 品種A | 品種B |
|---|---|
| 10.2 | 11.0 |
| 10.8 | 11.7 |
| 10.6 | 10.9 |
| 10.5 | 11.6 |
| 10.7 | 11.5 |
| 10.9 | 12.1 |
| 10.4 | 11.8 |
| 10.5 | 12.4 |
| 10.6 | 11.5 |
| 10.5 | |

## Ｄ．統計的検定についての補足

★　正規分布を仮定できない場合の検定

ｔ分布，カイ二乗分布，Ｆ分布は母集団が正規分布するときに標本の統計量（それぞれ母平均，母分散，２つの母分散の比）がどのように分布するかを表したものである．したがって，母集団が正規分布しないときにこれらの分布を用いた推定・検定をすると誤りをおかす可能性がある．ただし多少正規分布にずれた分布でもこれらの分布で検定してもおおむね大きな間違いはないことがわかっている．母集団が正規分布しなくても，数値を対数，指数変換することによって，正規分布に近似する場合は，変数変換してから，これらの方法を用いることができる．

正規分布に近似できない分布ではノンパラメトリック検定（授業では説明しない）が利用できる．

## Ｅ．宿題

１．第６回の宿題４．で調査したデータを用いて，95%信頼区間をつけて，２種類の卵（あるいは別のもの）の重さの母標準偏差をそれぞれ推定せよ．

２．第６回の宿題４．で調査したデータを用いて，２種類の卵（あるいは別のもの）の重さのばらつきが２つの間で異なるかを有意水準 5%でＦ検定せよ．

３．第６回の宿題４．で調査したデータについて，今回返した前回の宿題の講評の最後に書いてある母分散についての検定を行え．

４．種子が黄色で丸いエンドウと種子が緑色でしわのあるエンドウを交雑したＦ１からＦ２を得たところ，黄色で丸い種子のエンドウ，黄色でしわのある種子のエンドウ，緑色で丸い種子のエンドウ，緑色でしわのある種子のエンドウをそれぞれ 185，70，50，15 個体得た．メンデル遺伝に従い，9:3:3:1 に分離しているかを有意水準を 5%として，カイ二乗検定せよ．

５．次回から分散分析といい，同時に得られた３つ以上の標本について，その母集団の平均に有意な違いがあるかを検定する方法を学ぶ．試みに次のような調査のうち１つを選んで調査を行い，一番差の大きい２つの間で t 検定を試して見よ．

① ３つ以上のスーパーの卵(10 個以上をそれぞれ調べる）に重さの違いがあるか？なお，別に卵でなくてもかまわない．

② ３つ以上の品種のイネの１穂穎花数（10 穂以上をそれぞれ調べる）に差があるか？これも別にイネの穂に限らない．

③ ３つ以上の栽培方法あるいは品種の異なる果実の糖度あるいは酸度（10 果以上についてそれぞれ調べる）に差があるか？

提出締め切りは１２月１４日（月）午後１時までに生物資源科学部 2 号館 204 室に提出のこと．

**参考**：カイ二乗検定の応用

★ 計数（比率）データの検定

A，Bの 2 つの方法で飼育した昆虫の生存数と死亡数，ある特産品を 1, 2, 3, 4, ‥‥‥等級品に分け，地域ごとのこれら特産品の出現数などを行・列の 2 元表に分類したとき，この表を分割表という．この分割表についてカイ二乗検定を行うことができる．ここでは２×２の分割表について説明する．

２×２の分割表では次の式で $\chi^2$ を求める

$$\chi^2 = \frac{(ad - bc \pm n/2)^2 \times n}{T_1 T_2 T_A T_B}$$

２×２の分割表の例

| 飼育法 | 生存数 | 死亡数 | 計 |
|---|---|---|---|
| A | a | b | $T_A$ |
| B | c | d | $T_B$ |
| 計 | $T_1$ | $T_2$ | n |

＋，―は括弧の中の絶対値が小さくなる方を選ぶ．この $\chi^2$ を自由度 1 でカイ二乗検定する．なお a, b, c, d の値は 3 より小さいと精度が落ちるとされるので，なるべく観測度数を増やして検定することが望ましい．

例：トノサマガエルのオタマジャクシをA，B2 つの方法で飼育し，成体まで生存した数と死亡数を調査した結果は以下のようになった．2つの飼育方法で生存数に違いがあるかを検定せよ．

| 飼育法 | 生存数 | 死亡数 | 計 |
|---|---|---|---|
| A | 255 | 73 | 328 |
| B | 190 | 30 | 220 |
| 計 | 445 | 103 | 548 |

帰無仮説：２つの飼育方法には生存数に違いはない．

対立仮説：２つの飼育方法には生存数に差がある．

$$\chi^2 = \frac{(255 \cdot 30 - 190 \cdot 73 \pm 548/2)^2 \times 548}{445 \cdot 103 \cdot 328 \cdot 220} = 5.8578...$$

p-値＝$CHIDIST(\chi^2, 1) = 0.0155$

したがって，５％の有意水準で帰無仮説は棄却され，２つの飼育方法には生存数に差があると結論できる．