

第 7 回 カイ二乗分布, F 分布とその応用

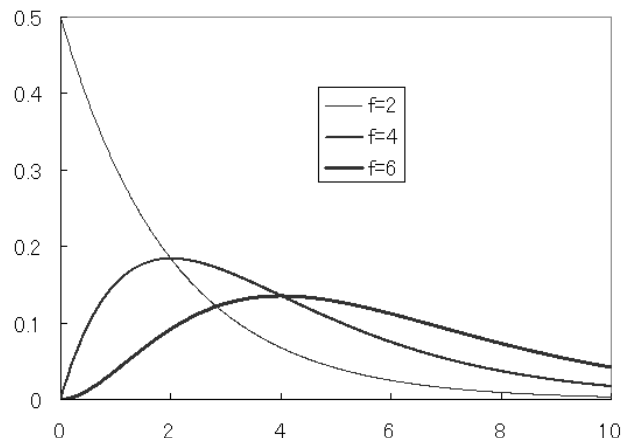
★ 教材「生物統計学__分散・標準偏差の区間推定 2013」を予習しながら空所を埋めておくこと

A. カイ二乗分布

1. カイ二乗分布 (χ^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 (χ^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**.



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.

$$\text{母分散 } \sigma^2 = V = 57.91g^2$$

$$\text{母標準偏差 } \sigma = SD = 7.6g$$

トマトの重さ(g)			
	100.0		
	110.0		
	105.0		
	95.0		
	98.0		
	118.0		
	103.0		
	92.0		
	100.0		
	94.0		
	110.0		
	105.0		
		エクセルでの計算式	
標本平均	102.5	=AVERAGE(K6:K17)	
標本分散	57.90909	=VAR(K6:K17)	
標本標準偏差	7.609802	=STDEV(K6:K17)	

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 DEVSQ と CHIINV を使うと計算できる. ここではデータをシートに代入すると計算できるようにした (100 個以内のデータに限る).

2013 年 11 月 19 日

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，データと信頼率をシートに代入（コピー）すると自動的に計算する．

標本番号	測定値			
1	100.0	カイ二乗分布に基づく区間推定		
2	110.0			
3	105.0	信頼率%	95 %	
4	95.0			
5	98.0	分散上限	166.9397	
6	118.0	分散下限	29.06015	
7	103.0			
8	92.0	標準偏差上限	12.92052	
9	100.0	標準偏差下限	5.390747	
10	94.0			
11	110.0			
12	105.0			
13				

母分散の区間推定

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

母標準偏差の区間推定

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

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．

予習問題：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．

母標準偏差の区間推定

$$(\quad) \leq \sigma(\%) \leq (\quad)$$

★ 教材「生物統計学__分散・標準偏差に関するカイ二乗検定 2013」を予習しながら空所を埋めておくこと

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， σ^2 は新しい飼育法を導入したあとの母集団の分散， σ_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 DEVSQ と CHIDIST 関数で計算できる。ここではデータをシートに代入すると計算できるようにした (100 個以内のデータに限る)。2 つの有意確率を計算する。そのうち小さい方の有意確率を使う。帰無仮説が成り立つとして、今回の結果が得られる確率はエクセルで以下のように計算できる。

カイ二乗検定による母分散の検定	
母標準偏差(帰無仮説)	0.07
母分散(帰無仮説)	0.0049
有意確率p-値A	0.005263381
有意確率p-値B	1.994736619
有意確率は小さい方を採用する	

③ 検定結果

p-値は () である。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 $0.005235 \leq \sigma^2 \leq 0.053806$

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

予習問題：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/m^2 と異なるかを有意水準 5%で検定せよ。

帰無仮説 H_0 :

対立仮説 H_1 :

p-値=

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

検定結果：

5%の有意水準で帰無仮説は (棄却できる ・ 棄却できない)

収量のばらつきは以前の品種と (同じである ・ 異なる ・ 異なるとはいえない)

★ 教材「生物統計学__カイ二乗検定の応用（メンデル遺伝）2013」を予習しながら空所を埋めておくこと

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。

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。

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

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）。今回は上の式を意識しないでも p 値を計算できるようにした。

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

例：F₁ 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。

帰無仮説：「分離比は 3 : 1 である」。

対立仮説：「分離比は 3 : 1 でない」

	観測値	理論値
赤花の数	80	82.5
白花の数	30	27.5
合計	110	
p-値	0.659658	

検定：p-値＝

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。

予習問題：F₁ 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。

帰無仮説：

対立仮説：

検定結果： p-値は（ ）である。したがって、有意水準 5% において、帰無仮説は（ 棄却された ・ 棄却できない ）ので、赤花：白花が 3:1 に（ 分離する ・ 分離しない ・ 分離するとはいえない ・ 分離しないとはいえない）。

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、右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。

	観測値
A-B-	65
A-bb	50
aaB-	30
aabb	15
合計	160

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

2 遺伝子座の場合でも、データをシートの黄色いところに観測数を代入すると計算できるようにした。

	観測値	理論値	分離比
A-B-	65	90	9
A-bb	50	30	3
aaB-	30	30	3
aabb	15	10	1
合計	160		
p-値	4.49E-05		

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 5% 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。

予習問題：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
赤一高一	赤色・草丈	80	9
赤一低低	赤色・草丈	35	3
白白高一	白色・草丈	30	3
白白低低	白色・草丈	15	1

帰無仮説：

対立仮説：

検定結果： p-値は () である。したがって、有意水準 5% において、帰無仮説は (棄却された ・ 棄却できない) ので、9:3:3:1 に (分離する ・ 分離しない ・ 分離するとはいえない ・ 分離しないとはいえない)。

3:1 や 9:3:3:1 以外の分離比の場合、3 遺伝子座以上の場合もカイ二乗検定で検定できる。ここでは方法は省略する。

3. 計数（比率）データの標本（群）の差の検定

着果率，発芽率，生存率のように全体のうち，何パーセントが着果，発芽，生存したかを示すデータを頻度データあるいは計数（比率）データという．頻度データの場合は調べた標本数が多くても，少なくともパーセント表示にしてしまうとその違いがなくなってしまうので，できるだけパーセントにしないで統計的検定をする方がよい．すなわちもし 2 つのデータがどちらも 30% であったとしても，片方が 100 個体から得たデータ（すなわち 30 個体が該当した），片方が 10 個体から得たデータ（すなわち 3 個体が該当した）ではデータの意味が全然違うからである．

ここでは昔からよく使われた方法であるカイ二乗検定を使う方法を説明する．しかし，カイ二乗検定は近似法であり，検出力も劣るので，コンピュータが利用できる昨今では二項検定やフィッシャーの正確確率検定を使うことも増えている．しかし，現在でもよく使われるので，カイ二乗検定をとりあげる．

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

右の表にある 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, B 2 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．

帰無仮説：2 つの飼育方法には生存率に違いはない．

対立仮説：2 つの飼育方法には生存率に差がある．

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

エクセルのシートに 2 つの飼育法による生存数と死亡数をそれぞれ入れるだけで p 値を計算するようにしてある．なお 2 種類の p 値を計算するので，大きい方の p 値を検定には利用する．

2つの飼育法による生存数の違いをカイ二乗検定するための表			
飼育法	生存数	死亡数	計
A	255	73	328
B	190	30	220
計	445	103	548
カイ二乗値その1	5.857828		
カイ二乗値その2	6.987332		
p値その1	0.015508		大きい方のp値を採用する
p値その2	0.008209		

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．

飼育法 A の生存率は 77.7%，飼育法 B の生存率は 86.4% である．パーセント表記すると飼育法 B の方が生存率が高いことがすぐ読み取れる．しかし，検定する場合はパーセント表記したデータでは多くの情報が失われるのであまり望ましくはない．

予習問題：カブトムシを宍道湖北側の山から採集してきた腐葉土と宍道湖の南側の山から採集してきた腐葉土で幼虫を飼育，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．腐葉土の採集地点の違いで生存数に違いがあるかを有意水準 5% で検定せよ．

腐葉土	生存数	死亡数
宍道湖北側	247	44
宍道湖南側	335	90

帰無仮説：

対立仮説：

p 値

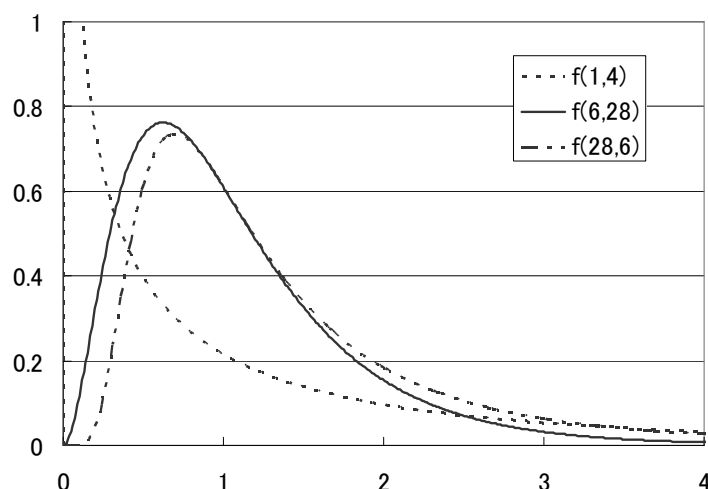
検定結果

★ 教材「生物統計学__分散に関する F 検定 2013」を予習しながら空所を埋めておくこと

C. F 分布と F 検定

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, B 2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 FTEST で計算できる。ここではデータをシートに代入すると計算できるようにした（それぞれ 100 個以内のデータに限る）。

標本番号	測定値A	測定値B		
1	73	71	有意確率p値	0.502413
2	74	68		
3	70	73		
4	77	71		
5	72	71		
6	72	72		
7	69	69		
8	73	72		
9	72	68		
10	75			

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	

帰無仮説 H_0 :

対立仮説 H_1 :

p-値=

検定結果 :

5%の有意水準で帰無仮説は (棄却できる ・ 棄却できない)

品種Bのばらつきは品種Aと (同じである ・ 異なる ・ 異なるとはいえない)

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が利用できる.

E. 宿題

宿題は <https://moodle.cerd.shimane-u.ac.jp/moodle/> を見てください.