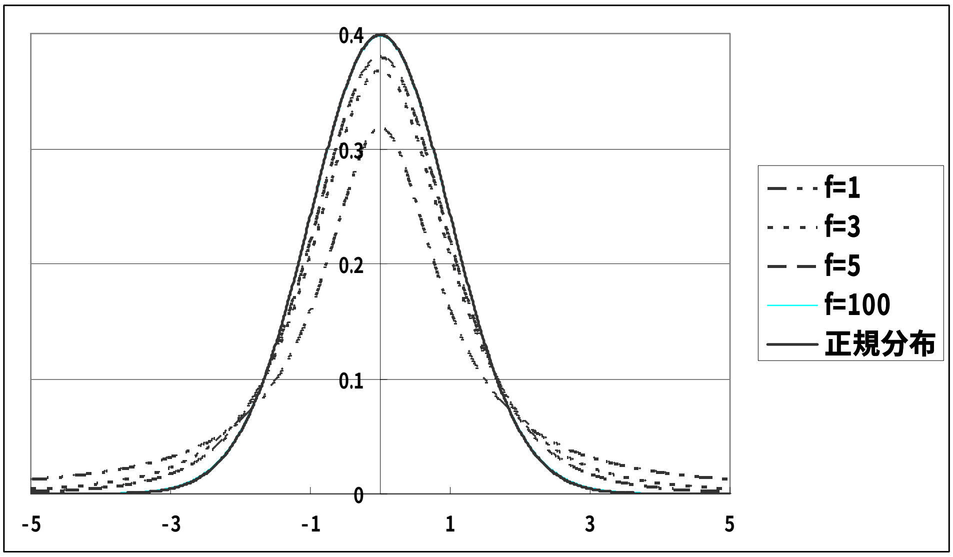
第７回　ｔ分布・カイ二乗分布

Ａ．　ｔ分布（小標本に関する平均の推定と検定）

１．ｔ分布

前回の授業では，標本が十分に大きいあるいは母分散が既知であることを条件に正規分布を用いて検定した．しかし，標本が小さい場合には，標本分散から母分散を推定するときの不確実さを加味したｔ分布を用いて推定・検定しなければならない．ｔ分布は標本分散の自由度ｆ(ふつう f=n-1)によって分布が決まる．標本数 n が増えるとｔ分布は正規分布に近づき，n=∞のときに正規分布と一致する．



２．ｔ分布による区間推定

母平均μの推定方法（<u>母分散σ^{2}は未知である</u>）

基本的な考え方は正規分布を用いた区間推定と同じである．

点推定：すなわち標本平均をそのまま母平均の点推定に使う．

区間推定：信頼率 P のときのμの信頼区間は以下の式を用いて計算する．

$$\bar{x} - t(f,1-P)SE < \mu < \bar{x} + t(f,1-P)SE$$

標準誤差 $SE = s/\sqrt{n} = \sqrt{V/n}$

f は自由度：$f = n-1$

n は標本数

P は信頼率（95%信頼区間なら 0.95）

エクセルで計算するときは

$t(f,1-P) = TINV(1-P, n-1)$ という TINV 関数を用いる．

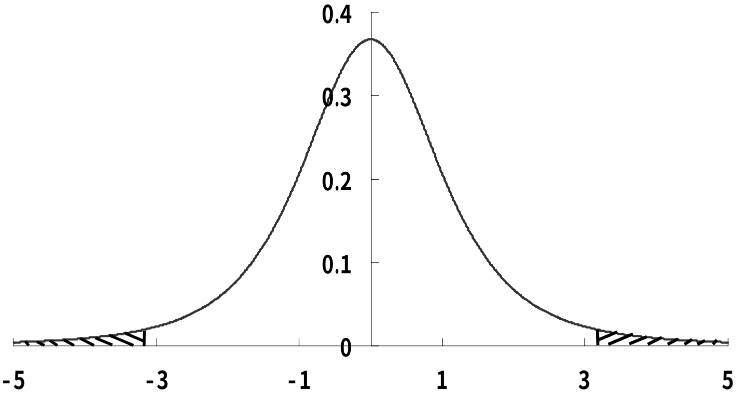


図 ｔ分布での 95%信頼区間

したがって，信頼率 P のときのμの信頼区間は，

$$\bar{x} - TINV(1-P, n-1) * SE < \mu < \bar{x} + TINV(1-P, n-1) * SE$$

例：Ａ公園の桜から 6 本を無作為に選びの木に付く花の数を数えた．123,156,168,190,211,234 の 6 つのデータを得た．Ａ公園の桜の花の数（平均）を 95%信頼区間および 99%信頼区間をつけて推定せよ．

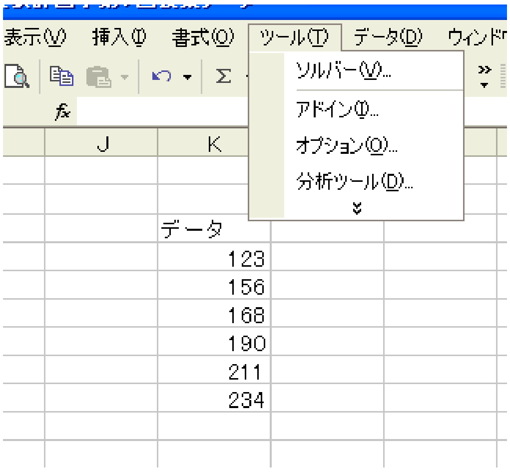
95%信頼区間をつけた推定値は

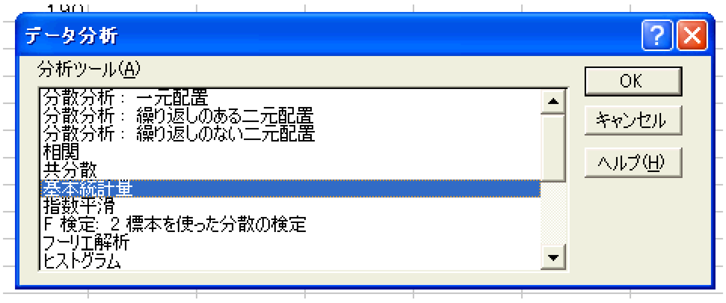
99%信頼区間をつけた推定値は

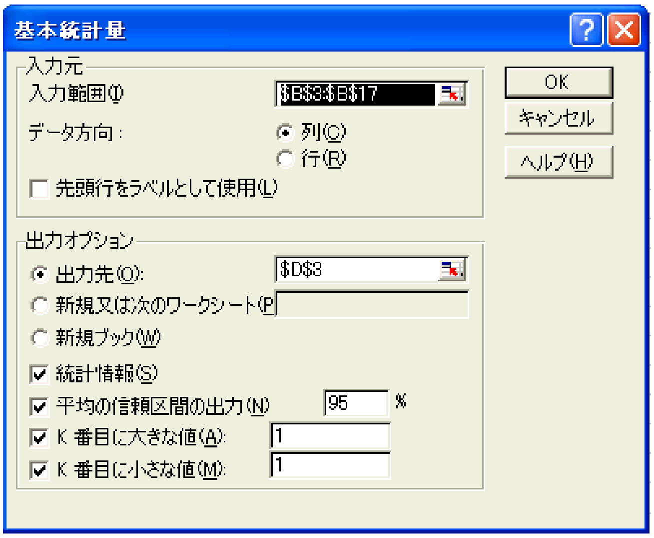
★　母標準偏差が既知であるとしたときにはどのくらい推定値の信頼区間が小さくなるか

練習：M大学の学生から 10 人を無作為に選び，100m 走をし，13.6±3.3 秒を得た．95%信頼区間を求めよ．

★　エクセルの分析ツール→基本統計量による計算の仕方







３．ｔ検定

①　ある決まった平均に対する検定

例：Ｔ食堂のラーメンの大盛りはライバル店Ｋレストランより 50g 多いと主張している．Ｋ君はＴ食堂で 10 回ラーメンの大盛りを注文し，こっそり重さを調べた結果，Ｋより 47±4g 多いという結果を得た．50g 多いというＴ食堂の主張を検定せよ．

帰無仮説：　H_0：　$\mu = 50g$

対立仮説：　H_1：　$\mu \neq 50g$

有意水準を設定する．この場合，有意水準を 5%としてみよう．

Ｐ値を計算する．

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次のように計算できる．

Ｐ値 $= TDIST(ABS(\mu - \bar{x})/SE,\ n-1, 2) = 0.041792$

練習（前述の練習と同じデータ）：M大学の学生から 10 人を無作為に選び，100m 走をし，13.6±3.3 秒を得た．大学当局の主張は 12 秒だが，これを検定せよ．

② 2つの母集団からの小標本の検定

ａ．対応のないデータのときのｔ検定

２つの独立した母集団から得た２つの小標本の平均に関する検定はｔ分布に基づいて行う．基本的な手順は正規分布による検定と同じである．標準偏差の扱いが少し複雑になる．

例：Ｔ牧場とＷ牧場のニワトリの卵を 10 個ずつ調査し，それぞれ右下の表のようなデータを得たとなった．両牧場の卵の重さの母平均は違うのかを検定せよ．

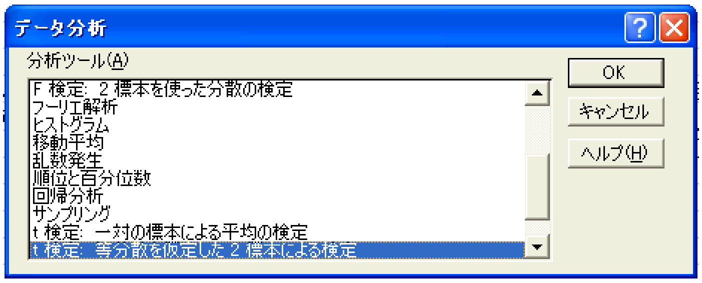
T牧場	W牧場
66.3	64.4
67.5	65.3
68.8	65.8
66.5	66.2
65	64.1
69.1	69.6
70.3	64.4
64	64.4
69.6	68.8
70.5	64.1

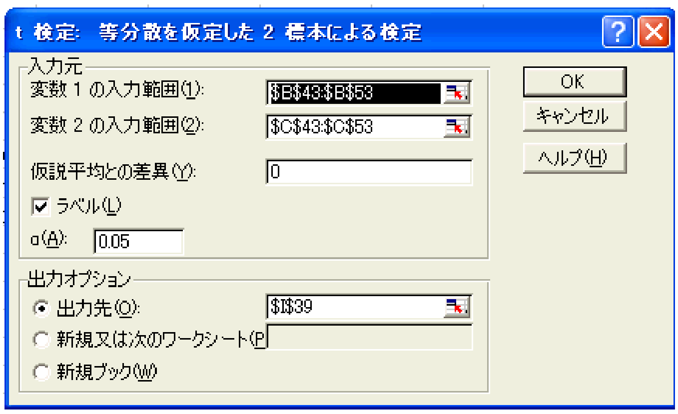
帰無仮説　H$_0$：　$\mu_T = \mu_W$

対立仮説　H$_1$：　$\mu_T \neq \mu_W$

ここでは有意水準を 5%としてみよう．

帰無仮説が成り立つとしたときに今回のデータが得られる確率であるＰ値はエクセルの分析ツールで計算できる．





t-検定：等分散を仮定した2標本による検定

	T牧場	W牧場
平均	67.76	65.71
分散	5.084889	3.936556
観測数	10	10
プールされた分散	4.510722	
仮説平均との差異	0	
自由度	18	
t	2.15832	
P(T<=t) 片側	0.022328	
t 境界値 片側	1.734063	
P(T<=t) 両側	0.044656	
t 境界値 両側	2.100924	

両側検定でのＰ値は 0.044655＜5%なので，帰無仮説は有意水準 5%で棄却された．しかし，有意水準 1%にすると棄却できない．以上のことから，有意水準 5%でＴ牧場とＷ牧場のニワトリの卵の重さ（の母平均）は異なると結論できる．

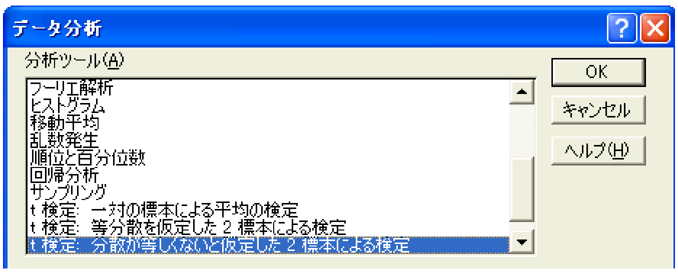
練習：Ａ地区とＢ地区それぞれ地区ぐるみで健康のために減塩に取り組んだ．無作為に選んだ標本から摂取食塩量を調査した．その結果は右下の表のようになった．食塩摂取量に差はあるのかを検定せよ．

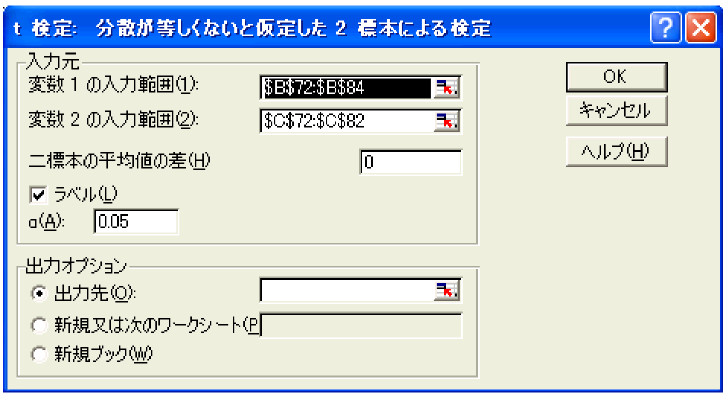
Ａ地区	Ｂ地区
6.0	5.6
6.2	6.4
5.8	6.8
5.6	6.6
6.0	6.0
5.9	5.8
6.4	6.2
6.8	6.6
5.4	5.9
5.8	6.3
6.2	
6.4	

★　ｔ検定の注意点

２つの独立した母集団の母平均に差があるかどうかを検定するｔ検定の場合，２つの母集団の母分散が等しいことが前提となっている．しかし，実際には標本数がほぼ同じ場合には母分散が異なっていてもそれほど検定に問題がないことがわかっている．したがって，母分散が異なっており，標本数も大きく異なる（おおむね２倍以上）場合にはこの方法を用いると問題がある．この場合には Welch の検定を使う．

なお母分散に差があるかどうかを検定する方法はＦ検定という，次回，学ぶ予定である．Welch の検定はホームページで詳しく紹介するが，下のような検定である．





t-検定：分散が等しくないと仮定した2標本による検定

	Ａ地区	Ｂ地区
平均	6.041667	6.22
分散	0.148106	0.152889
観測数	12	10
仮説平均と	0	
自由度	19	
t	-1.07284	
P(T<=t) 片	0.148391	
t 境界値 片	1.729131	
P(T<=t) 両	0.296782	
t 境界値 両	2.093025	

b．対応のあるデータのときの t 検定

例：A，B の２つのハカリで同じ品物を量る．同様に 10 個の品物についてそれぞれ量って，右の表のような結果を得た．２つのハカリの指示には差があるか．

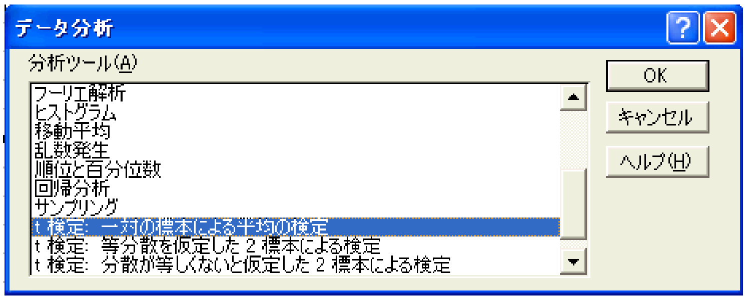
品物 No.	はかりA での値	はかりB での値	重さの差 d
1	427	419	8
2	401	391	10
3	408	410	-2
4	417	406	11
5	416	406	10
6	396	400	-4
7	437	432	5
8	386	391	-5
9	406	402	4
10	402	395	7
平均	409.6	405.2	4.4

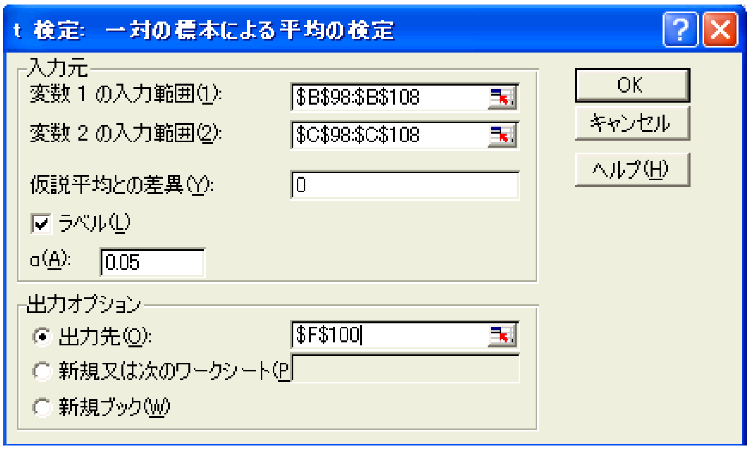
前項の対応のないデータでの検定を行うと 5%で有意でないという結果が出る．しかし，A，B のハカリの差を品物ごとに取ると傾向がありそうだとわかる．このように２つの標本のデータが対応する場合，２つの標本は独立していないといい，対応するデータの対の差 d を検定しなければならない．

対応のあるデータの差 d について検定する．

帰無仮説：　H_0：$\mu_d = 0$　A と B の２つのはかりの指示は同じである．

対立仮説：　H_1：$\mu_d \neq 0$　A と B の２つのはかりの指示は異なる．





t-検定：一対の標本による平均の検定ツール

	はかりA での値	はかりB での値
平均	409.6	405.2
分散	226.4889	164.1778
観測数	10	10
ピアソン相関	0.918931	
仮説平均との差異	0	
自由度	9	
t	2.310462	
P(T<=t) 片側	0.023101	
t 境界値 片側	1.833114	
P(T<=t) 両側	0.046202	
t 境界値 両側	2.262159	

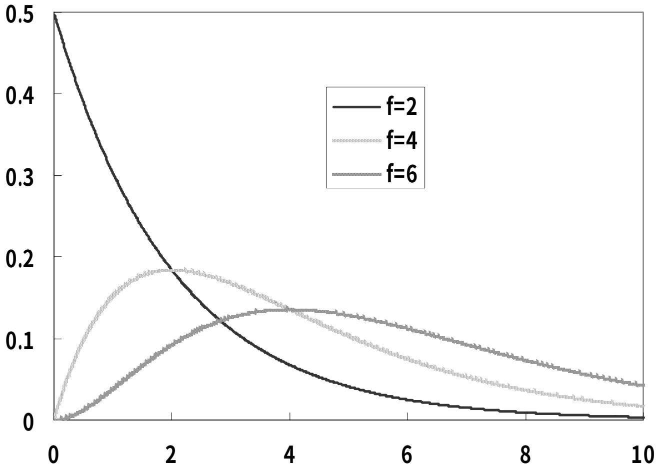
Ｐ値が 0.046202＜5%であるから，有意水準 5%で帰無仮説は棄却され，Ａ，Ｂ２つのハカリの指示に 5%の有意水準で相違があると結論できる．

Ｂ．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．

母分散$\sigma^2 = V = 57.91g^2$

母標準偏差$\sigma = SD = 7.61g$

	A	B	C
1			
2		トマトの重さ(g)	
3		100	
4		110	
5		105	
6		95	
7		98	
8		118	
9		103	
10		92	
11		100	
12		94	
13		110	
14		105	
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)

② 母分散の区間推定

母分散の p%信頼区間は次の式に従う.

$$\frac{S}{\chi^2(f,\frac{100-p}{200})} \le \sigma^2 \le \frac{S}{\chi^2(f,\frac{100+p}{200})}$$

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

例：母分散の 95%信頼区間は次の式に従う.

$$\frac{S}{\chi^2(f,0.025)} \le \sigma^2 \le \frac{S}{\chi^2(f,0.975)}$$

エクセルでは母分散の p%信頼区間を以下のように計算する.

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

例：トマト 12 個を計ると右のような値を得た（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.

	A	B	C
1			
2		トマトの重さ(g)	
3		100	
4		110	
5		105	
6		95	
7		98	
8		118	
9		103	
10		92	
11		100	
12		94	
13		110	
14		105	
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)

95%信頼区間をつけた区間推定

$$\frac{637}{\chi^2(11,0.025)} \le \sigma^2 \le \frac{637}{\chi^2(11,0.975)}$$

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 57.91g^2，母標準偏差 7.61g）が区間推定のまん中には来ない.

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.

有意水準を 5%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.

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

$$\text{P 値} = 2 \times CHIDIST(\chi_0^2, f) = 2 \times CHIDIST(\frac{S}{\sigma_0^2}, n-1)$$

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

$$\text{P 値} = 2 \times (1 - CHIDIST(\chi_0^2, f)) = 2 \times (1 - CHIDIST(\frac{S}{\sigma_0^2}, n-1))$$

１)，２）とも片側検定の場合は，2×をとる.

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 $0.005235 \leq \sigma^2 \leq 0.053806$
母標準偏差の区間推定 $0.07235 \leq \sigma \leq 0.23196(\%)$

C．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，生ー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．

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

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，上側検定になる．（下側は分離比に当てはまりすぎるということを意味する）

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，χ^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．

$$\chi^2_{adj} = \sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}, \quad \text{P 値} = CHIDIST(\chi^2_{adj}, 1)$$

★　メンデル遺伝での分離比の検定

メンデルの分離の法則に従ったデータが得られたかを検定するのに，カイ二乗検定を用いる．

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	理論値
赤ー高ー	赤色・草丈高い	65	9	90
赤ー低低	赤色・草丈低い	50	3	30
白白高ー	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	理論値	
2	赤ー高ー	赤色・草丈高い	65	9	90	
3	赤ー低低	赤色・草丈低い	50	3	30	
4	白白高ー	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)		
11						

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．

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．

自由度が 1 なので，補正する．

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

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．

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	理論値
赤一高一	赤色･草丈高い	80	9	90
赤一低低	赤色･草丈低い	35	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	

D．宿題

１．第５回の宿題で調査したデータを用いて，２種類の卵の重さの母平均が同じであるかを有意水準 5%でｔ検定せよ．また，それぞれの卵について，その母平均および母標準偏差を 95%信頼区間をつけて，推定せよ．

２．第５回の宿題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平均ならびに母分散についての検定を行え．

４．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？