χ二乗検定

F 検定や t 検定などの分散分析では、等分散性と正規性が検定の必須条件です。しかし、等分散性のないデータはよくあります。例えば、医薬品の投薬試験のデータは、効果のあった人の割合で表されますが、投薬量が少ない時はすべての人に効果がありません。投薬量が増えると、効果のある人の数が増えますが、ある程度の投薬量になっても、何らかの理由でその薬の効果がない人には効果がありませんから、ある一定の段階で投薬量を増やしても、効果のある人の割合はそれ以上高くなりません。つまり、投薬量が少ない時はデータの変動が小さく、中間の段階で変動が大きくなり、さらに投薬量が増えるとまた変動が小さくなります。図 1 は割合のデータが持っ ている特性を説明するためにに作ったグラフです。

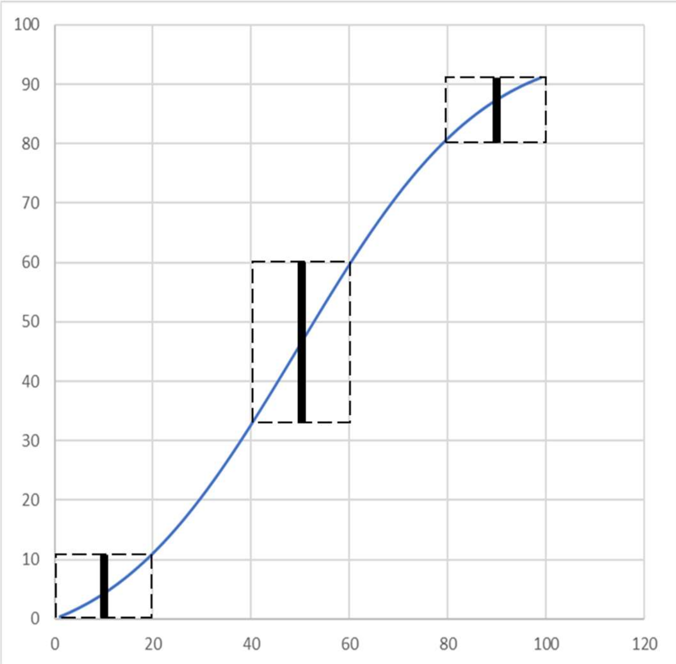

図 1 .割合のデータの特徴

ここでは、正規分布の累積確率曲線を李よしましたが、別に正規分布である必要はありません。単に一般の累積割合の S 字曲線と理解してください。つまり、両側でなだらかで、中間で険しい曲線です。薬の投薬効果は人の体重や代謝活性・生理状態で変わります。ある投薬量を与えたグループの人々は体重や生理状態が異なる様々な人がいます。体重や代謝レベルを単位にしたときには、実質的に投薬量が微妙に違っているのです。つまり、実質的にはこの図の横方向に変動があります。横方向の変動を一定と考えて、この横方向の変動内での縦方向の変動を示したのが黒い太い線です。この太い線を比較すると、両側で短く、中央部で最大になっていることがわかります。S 字曲線はこのような性質を持っています。投薬試験の場合、投薬量の増加に伴って効果が下がることは普通あり得ないから、データは累積比率のよ うになって、100%と近くなると、もうそれ以上値が増加することはありません。し

たがっ て、変動も小さくなるのです。変動が大きいのは、第二次導関数が正から負に代わる点（変曲点)、第二次導関数が 0 になる点です。こういう性質を持ったデータは珍しくありません。 そういう場合、データを何らかの関数で変換して、等分散性を持たせて分散分析をするというのも一つの方法です。しかし、等分散性を持つようなうまい変換を思いつかないかもしれません。多くの統計の教科書では、比率データの場合、分散分析ではなくて、期待値と測定値の差の分散比であるカイ二乗検定をすることを薦めています。すでに説明したようにカイ二乗分布は実測値と期待値の差の分散とデータ分散の比の分布ですから一種の分散分析なのです。しかし、等分散性がないデータの分析に使えますし、名義変数や順序変数で表現されたデータの分析にも使えます。もちろん他にも、様々な分析法があります。等分散性がなければ、即、カイ二乗検定にすると機械的に考える必要はありませんが、広い範囲で使える便利な検定法です。カイ二検定が使えるのは、何らかのかたちで期待値が与えられることです。カイ二乗検定は、何らかの期待値が考えられる場合に有力で汎用性の広い検定法だと考えればよいでしょう。

いくつかのサブサンプル間に違いがあるかないかという検定の場合に、そのサブサンプルについての期待値があたえられれば、カイ二乗の観測値を次の式で求めます。

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

f_i：サブサンプ iの実測値 e_i: サブサンプルi の期待値

この式は、多くの統計学の教科書で紹介されています。しかし、この式の意味を詳しく解説したものは極めて少ないと思います。割合や比率のデータは、二項分布的です。二項分布では、期待値が、$e_i = \frac{\sigma^2}{2}$になっていることはすでにχ二乗分布の講義で説明しました。

カイ二乗値は実は次のような分散比の和なのです。

$$\chi^2 = \sum_{i=1}^{n} \frac{2(f_i - \mu_i)}{\sigma^2}$$

確率的な現象の例として良く出てくるのはサイコロです。サイコロのデータを使って、カイ二乗検定のやり方の簡単な例を示します。検定したいのは、サイコロがインチキのない正確に作られたサイコロであるかどうかです。そこで、帰無仮説はサイコロが正確に作られているということにします。この帰無仮説にしたがうと、600 回サイコロを転がしたときの、各目が出てくる回数の期待値は１００ずつです。しかし、実際には確率的な変動があるから、すべての目が 100 回ずつ出ると いうことはないでしょう。実際に得られた回数（実測値）と期待値の間には差があります。 この場合のカイ二乗の計算例を表 1 に示しました。

表１．サイコロを転がした時に出てくる各目の観測値と期待値

目	観測値	期待値	差	差の二乗	χ^2値
	A	B	C＝A-B	D＝C^2	E＝D/B
1	120	100	20	400	4
2	85	100	-15	225	2.25
3	113	100	13	169	1.69
4	115	100	15	225	2.25
5	80	100	-20	400	4
6	87	100	-13	169	1.69
合計	600	600	0		15.88

この例では実測されたカイ二乗値は 15.88 、一つひとつの目について６通りのデー タがあるから自由度は $df = n - 1 = 6 - 1 = 5$ です。統計の本などにある自由度５のカイ二乗値の限界値は、$p \leq 0.01$ で 15.0863 です。ですから、この値から、帰無仮説を棄却して、このサイコロは正確なサイコロではないという結論が出せます。これは、考え方を示すための例です。実際にこんなことをする人はいないでしょう。サイコロの場合、期待値は試行回数の６分の１だということが初めから解っていますが、期待値をどのようにするかは、何が帰無仮説かによって変ります。ここが、カイ二乗の難しいところでもあり、面白いところでもあります。表２に、もう少し現実的で具体的な例を挙げました。A という薬と B という薬の投薬試験で、A と B の薬の効果に違いがあるかを検定します。A の薬を 116 人の人の投薬し、66 人に効果が ありました。一方、97 人の人に B の薬を投薬したところ 56 人効果がありました。

表２ ２つの治療薬の効果試験の結果

	A	B	合計
効果あり	66	56	122
効果なし	50	41	91
合計	116	97	213

このような事例では、まず、何を検定するのか帰無仮説を考えて、それにしたがって期待値を計算します。この場合、帰無仮説は薬の効果に差がないということになります。。効果に差がないのだから、A の薬でも B の薬でも、同じ割合の人に効果があるはずだから、A の薬でも B の薬でも、効果のあった人の割合は、両方を合わせた全員の中で、効果があった人の割合で、効果のなかった人の割合は、両方を合わせた全員の中で効果のなかった人の割合です。

$$\text{効果があった人の割合：}\ \frac{122}{203} = 0.57277$$

$$\text{効果がなかった人の割合：}\frac{91}{203} = 0.42723$$

これらの割合を使って、表 3.のように、それぞれの期待値を求めます。

表 3. A,B の効果に差がない場合の期待値

116×0.57277　　97×0.57277

	A	B	合計
効果あり	66.44131	55.55869	122
効果なし	49.55869	41.44131	91
合計	116	97	213

116×0.42723　　97×0.42723

以下、表 4 に示したように、実測値の期待値の差、実測値と期待値の差の二乗、実測値と期待値の 2 乗を期待値で割って、カイ二乗を求める。カイ二乗の総和を求めるという手順で、計算します。

表 4. カイ二乗の計算手順。

期待値			
	A	B	合計
効果あり	66.44131	55.55869	122
効果なし	49.55869	41.44131	91
合計	116	97	213
期待値と実測値の差			
	A	B	合計
効果あり	-0.44131	0.441315	0
効果なし	0.441315	-0.44131	0
合計	0	0	0
期待値との差の 2 乗			
	A	B	
効果あり	0.194759	0.194759	
効果なし	0.194759	0.194759	
χ^2値			
	A	B	
効果あり	0.002931	0.003505	0.006437
効果なし	0.00393	0.0047	0.008629
合計	0.006861	0.008205	0.015066

66-66.44131

$(-0.44131)^2$

$$\frac{(f-e)^2}{e} = \frac{0.194759}{66.44131}$$

$$\sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

このように計算してχ^2の観測値は 0.01506622 です。自由度は薬が 2 通り、効果ありナシという結果が 2 通りなので (2 − 1)(2 − 1) = 1 となります。カイ二乗の限界値の表を見るま でもなく、カイ二乗値は小さく、帰無仮説を棄却できないので、A、B 二つの薬の投薬

効果に差があるとは結論できません。

この例は、比較的単純な比較ですが、もう少し複雑なデータでも、同じ考え方でカイ二乗検定が出来ます。次の例は、人々が自分がどんな階層に属しているかという意識を、町ごとに比べるという例です。階層意識というのは収入などの具体的な数値とは関係なく、貧乏な人でも、自分は上流階級に属していると感じていたりします。これは、所属している社会の文化と関係しています。そこで、そのような意識が住んでいる地域によって違うかどうか、A、B、C、D の４つの町で、「あなたは、上流階級、中流階級、下流階級のどの階級に属しますかという質問をしました。回答結果は、表 5 の実データの欄のようになりました。この結果から、４つの 町の間に違いがあるかどうかを検定します。計算のプロセスは、

１．クロス集計表を作る

２．和を求める

３．全体で、上流、中流、下流と意識している人の割合を計算する

４．それぞれの街の人数の和と割合から、それぞれも町ごとに、上流、中流、下流と思っている人の期待値を計算する。

５．実データと期待値の差を計算する

６．実データと期待値の差の２乗を計算する

７．実データと期待値の差の二乗を期待値で割る。

８．合計して、全体のカイ二乗を求める

という手順になります。

この計算をした、エクセルシートを示します。

表５．町ごとの階層意識の違いに関するカイ二乗検定の計算（エクセルシート）

$\frac{50}{240}$

	実データ				合計	割合
階層	町					
	A	B	C	D		
上流	10	20	5	15	50	0.208333
中流	35	30	30	50	145	0.604167
下流	15	0	20	10	45	0.1875
合計	60	50	55	75	240	1

60×0.208333

	期待値				
階層	町				
	A	B	C	D	
上流	12.5	10.41667	11.45833	15.625	50
中流	36.25	30.20833	33.22917	45.3125	145
下流	11.25	9.375	10.3125	14.0625	45
合計	60	50	55	75	240

$10 - 12.5$

	実データと期待値の差				
階層	町				
	A	B	C	D	
上流	-2.5	9.583333	-6.45833	-0.625	0
中流	-1.25	-0.20833	-3.22917	4.6875	0
下流	3.75	-9.375	9.6875	-4.0625	0
合計	0	0	0	0	0

$(-2.5)^2$

	実データと期待値の差の２乗				
階層	町				
	A	B	C	D	
上流	6.25	91.84028	41.71007	0.39063	140.191
中流	1.5625	0.043403	10.42752	21.9727	34.00608
下流	14.0625	87.89063	93.84766	16.5039	212.3047
合計	21.875	179.7743	145.9852	38.8672	386.5017

$\frac{(f-e)^2}{e} = \frac{6.25}{12.5}$

$\sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$

	χ^2乗値				
階層	町				
	A	B	C	D	
上流	0.5	8.816667	3.640152	0.025	12.98182
中流	0.043103	0.001437	0.313806	0.48491	0.84326
下流	1.25	9.375	9.100379	1.17361	20.89899
合計	1.793103	18.1931	13.05434	1.68352	34.72407

χ^2: 34.72407

町の自由度：4-1=3

社会階層の自由度：3-1=2

全体の自由度：(4-1)(3-1)=6

自由度６のχ^2臨界値：18.5476

以上の結果から、帰無仮説を棄却し、４つの町の社会階層意識には差があると結論できます。