任意標本調査法（一）

関　弥三郎

一　母集団分布
(1) 統計集団の記述
(2) 量的構造の簡約表章
(i) 算術平均値
(ii) 標準偏差（分散）
(iii) 正規分布
(iv) 歪度、尖峰度
(3) 質的構造の簡約表章
(4) 相関々係の簡約表章
(i) 相関係数
(ii) 級内相関係数

従来社会統計調査に於ては悉皆調査、即ち社会的集団現象を構成するすべての要素を観察する事が原則とせられ、その一部分のみの観察による標本調査は、悉皆調査の困難な場合にとられる代用法に過ぎないとされて来たのである。それは悉皆調査による時は社会的集団現象の正確な統計を獲得する事が出来るが、標本調査による時は誤差――そしてその大いさを知る事は出来ない――を伴う統計しか得られないからであった。しかし統計調査の実際に於ては大規模な悉皆調査の可能な場合は極めて少く、大部分は代用法たる標本調査によらざるを得ない状態であり、且悉皆調査もそれの実行に於て多くの技術的誤差が発生するため、実際上は必ずしも正確な統計を得るものではないのである。

しかるに第二次世界大戦の直前より、近代数理統計学の成果たる抽出理論に基く任意標本調査法が発達し、

標本調査による推定の誤差を規定し得るようになったため、標本調査によっても悉皆調査に劣らず十分正確に社会的集団現象を認識する事が可能となり、今や標本調査は単なる代用法ではなく悉皆調査と並ぶ正確な調査方法となったのであって、社会統計調査のあらゆる方面に於て急速に発達しつつあるのである。

しかし任意標本調査法による社会的集団現象の認識には一定の限界があるのであって、且それの実行には一定の条件が整っていなければならず、又精巧な方法である丈に誤用する危険も大きいため、標本調査の実施に際してはよくその限界並に必要条件を弁え、いやしくも誤用、濫用に陥らざるよう注意しなければならないのである。しかるに任意標本調査法の発達は日未だ尚浅く、殊に我が国に於ては近々数年間の事柄に過ぎないため、それの実行に必要な基礎資料の整備、充実は勿論、経験のある優れた調査立案者の養成が、それの普及と共に重大な問題となって来るのである。

このような任意標本調査法に対する社会的需要の増大に鑑み、茲に任意標本理論を解説してそれの理解の一助たらしめんとするのである。社会統計調査法に於ける任意標本調査法の地位並にそれの方法的特質に就ては、前号の「統計調査法」に於て考察したため、茲では任意標本理論の概要を出来る丈平易に説明する事にしよう。

一、母集団分布

統計集団の大いさ並に構造の絶対数による正確な把握は、集団構成要素の悉皆観察によってのみ可能であるが、統計集団の構造をそれの簡約表章である平均値、比例数等によって把握する事は、勿論誤差を伴うが、集団構成要素の部分観察即ち標本調査によっても可能である。

そして抽出理論に基く任意標本調査法に於ては、調査せんとする統計集団を母集団とし、それより抽出された統計単位の集団を標本とし、両者を任意抽出操作 random sampling operation を媒介として結附け

る事によって、標本の構造を示す値（平均値、比例数等、これを統計量という）より、悉皆観察によって明らかにされるであろう母集団の構造を示す値（平均値、比例数等、即ち母数）を、一定の信頼度に於て、誤差の範囲を明示して推定せんとするのである。従って先ず母集団の構造の簡約表章は如何にして求められるかを知らねばならないのである。

(1) 統計集団の記述

統計的方法による統計集団の把握は、それを構成する要素即ち統計単位の個別観察、分類、集計の手続を経て行われるのであって、統計集団の性質は一方向の集団性より観察した統計集団の数量的構造として明らかにされるのである。即ち一方向の集団性に就てすべての統計単位を観察して各々の有する性質に応じて標識を与え、次で標識毎にそれを有する単位を分類しその数を合計する事によって、各標識を同じ性質の単位の数によって具体的に規定し（この値を統計値という）、最後に集団性によってこれ等の統計的規定を受けた標識を統括する事によって、統計集団の大いさが明らかになり、且それは標識を同じくする、従って同質的な統計単位の部分集団より成るものとしてその構造を明らかにし、これ等部分集団相互の乃至は部分集団の統計集団全体に対する数量的関係によって、一方向に於ける統計集団の性質が数量的に規定されるのである。

今この事を明らかにするために従業員80人の一事業所を想定し、この勤労者集団の男女構成、給与状態、扶養家族数等の性質を明らかにせんとする場合を考えよう。先ず規定せられた集団性である「体性（男女）」「一ヵ月間賃銀額」及び「扶養家族数」をすべての従業員に就て観察して、それぞれ「男」「一五、六〇〇円」「三人」、「女」「六、二五〇円」「ナシ」等の標識を与え、その結果を標識毎に分類、集計し、それを集団性によって統括する事によって第一―第三表の統計表が得られたとする。

第一表はこの集団は「体性（男女）」の集団性に就

第一表　男女別従業員数

男女別	人数
男	60人
女	20
計	80人

第二表　扶養家族数別従業員数

家族数	人数
0人	25人
1	28
2	14
3	9
4	3
5	1
計	80人

第三表　一ケ月間賃銀階級別従業員数

賃銀階級	人数
千円 5 以上 千円 6 未満	2人
6 〃 7 〃	3
7 〃 8 〃	6
8 〃 9 〃	10
9 〃 10 〃	15
10 〃 11 〃	13
11 〃 12 〃	10
12 〃 13 〃	7
13 〃 14 〃	5
14 〃 15 〃	3
15 〃 16 〃	2
16 〃 17 〃	1
17 〃 18 〃	1
18 〃 19 〃	1
19 〃 20 〃	1
計	80人

て見た時は、男と女の部分集団より成って居り、それは80人中男60人と女20人という数的構成に於てある事を示して居り、第二表は「扶養家族数」の集団性に就て見た場合は、扶養家族を有しない者から五人有する者迄に分れ、その各〻は80人中25人、28人……1人の数的関係にある事を明らかにしているのであり、第三表は「一ケ月間賃銀額」に就て見た時の集団の数量的構造を、それと同じようにして示しているのであって、かくしてこの事業所内の勤労者集団の諸性質が数量的に規定されているのである。

しかしながらそのような数量的相互関係の概観は容易でないため、統計集団の性質の簡約な表章が必要となるのであって、それは集団性が質的な性質の場合には比例数を求める事により、量的な性質の場合には平均値、標準偏差等を計算する事によって可能となるのである。

(2)　量的構造の簡約表章

量的集団性に就て統計集団を観察する時、（例えば人口集団を「年齢」に就て、勤労者集団を「賃銀額」に就て観察する場合）個々の統計単位に与えられる標識は数量値であってこの量的標識を変量 variate といい、それには二の一定値の間に多くの任意の値をとり得る連続量なる場合と、分離した値のみをとりその中間の値をとり得ない非連続量（離散量）なる場合とがある。前節の例に於て賃銀額は連続的変量であり、扶

養家族数は非連続的変量である。

連続的変量の場合、統計単位の分類は各単位に与えられた数量値（即ち観察標識）によってもよいが、集団が小さく統計単位の数が多くない時は、それでは少数の単位より成る多くの部分集団に分割され、統計集団の構造が不明瞭になる場合が多いのである。そこで一定の範囲の数量値をまとめて分類標識とし、その範囲内の変量を有する単位はすべて同一の部分集団に分類する事によって、部分集団の数を減らしてその大きさを増し、以て統計集団の構造を明確に表現せしめる事が必要となるのである。しかるに非連続的変量の場合は、観察標識によっても部分集団の数は余り多くならないから、一般にこのような分類標識を設定する必要は起らないのである。

前例に於て集団性「一ヶ月間賃銀額」に就て統計単位を観察した結果を、観察標識によつて分類すると第四表の如くであり、これを「五千円―六千円」「六千円―七千円」……の分類標識によって整理する事によつて第三表が得られたのであるが、統計集団の構造は第四表によるよりも第三表による方が明確に表現され得るであろう。

第四表

賃銀額	人数
千円	人
5.2	1
5.7	1
6.3	2
6.7	1
7.2	2
7.5	1
7.8	3
8.0	2
……	……
計	80人

しかし「扶養家族数」に就て観察した場合は、観察標識を以てそのまゝ分類標識とする事によって第二表が得られたのである。

統計集団の単位を量的標識によって部分集団に分類した結果を度数分布 frequency distribution といい、それを表示したものを度数分布表という。そして各部分集団を級 class、級に属する単位の数を度数 frequency といい、各級の量的標識の幅を級間隔 class interval、その量的標識の限界を級限界 class limit という。

(i) 算術平均値

度数分布として示された統計集団の量的構造の簡約表章は、すべての統計単位は数量値を異にするのであるから、そのうち代表的、典型的な数量値（中数 Mittelwerte）を規定する事によって可能となるであろう。そして中数としては平均値 mean（算術平均値、幾何平均値、調和平均値、平方平均値）、中位値 medium 最頻値（又は並み数）mode が求められる。これ等のうち算術平均値が最も数学的処理が進んで居り、任意標本理論に於ても一般にそれが取扱われるため、茲では算術平均値のみを説明する事にする。

算術平均値 arithmetic mean は、すべての統計単位の変量の総和を単位の総数で割った値であって、統計集団全体の量的大いさ（即ち単位の変量の総和）が、全統計単位に均一に配分された場合の仮想的な値である。単位の変量を x、統計集団の単位の総数を N とすれば、算術平均値 m の計算式はその定義によって次の如くである。

$$m=\frac{x_1+x_2+\cdots\cdots+x_N}{N}=\frac{\sum_{i=1}^{N}x_i}{N} \qquad (1)$$

茲に $\sum$（シグマ sigma）は総和を示す記号であって、$\sum_{i=1}^{N}x_i$ は x を1より N 迄加える事を意味する。

例えば $x_1=2,\ x_2=3,\ x_3=5,\ x_4=5,\ x_5=7,\ x_6=9$ とすれば、

$$\sum_{i=1}^{6}x_i=2+3+5+5+7+9=31$$

$$\sum_{i=2}^{5}x_i=3+5+5+5=20$$

この $\sum$（即ち系列の各項の和）の計算に習熟する事は抽出理論を学ぶ上に於て極めて重要である。故に茲ではそれの練習の意味を兼ねて、簡約表章の計算方法を少し詳細に説明するであろう。

それに先立つて次の事を注意して置こう。

$$\sum_{i=1}^{N}ax_i=ax_1+ax_2+\cdots+ax_N,\ a\text{ は任意の常数(constant)}$$

$$=a(x_1+x_2+\cdots\cdots+x_N)=a\sum_{i=1}^{N}x_i$$

即ちすべての変数 x に或る一定の値（常数）a が掛っている場合は、その総和は x のみの総和の a 倍に等しいのである。又

$$\sum_{i=1}^{N}(x_i\pm y_i)=(x_1\pm y_1)+(x_2\pm y_2)+\cdots\cdots+(x_N\pm y_N)$$

$$=(x_1+x_2+\cdots\cdots+x_N)\pm(y_1+y_2+\cdots\cdots+y_N)$$

$$=\sum_{i=1}^{N}x_i\pm\sum_{i=1}^{N}y_i$$

即ち二の変数 x と y の和（又は差）の合計は、x の合計と

yの合計の和（又は差）に等しい。

しかしながら多くの場合統計単位の有する変量は、個々的に与えられないで度数分布に整理されている。その時は度数fをウエイト（重み weight）とした加重平均値 weighted mean を求めればよいのである。変量x_1がf_1単位、x_2がf_2単位、……x_kがf_k単位あるとすれば、

$$m=\frac{x_1f_1+x_2f_2+\cdots\cdots+x_kf_k}{f_1+f_2+\cdots\cdots+f_k}=\frac{\sum_{i=1}^{k}x_if_i}{\sum_{i=1}^{k}f_i},\quad \sum_{i=1}^{k}f_i=N$$

$$=\frac{\sum_{i=1}^{k}x_if_i}{N}\qquad(2)$$

によって求められる。第二表より平均扶養家族数を求めると第五表の如くである。

度数分布表

変量	度数
x_1	f_1
x_2……	f_2……
x_k	f_k
計	N

この場合一変量x_iを有する単位f_iの全体に於ける存在の割合（相対的度数）は$p_i=\frac{f_i}{N}$によって求められる。従って(2)式は次のようになる。

$$\frac{\sum_{i=1}^{k}x_if_i}{N}=\sum_{i=1}^{k}x_i\frac{f_i}{N}=\sum_{i=1}^{k}x_ip_i\qquad(2')$$

即ち算術平均値は変量とそれの相対的度数との積の総和である。

第五表　平均扶養家族数の計算表

変量 x_i	度数 f_i	x_if_i
0人	25人	0人
1	28	28
2	14	28
3	9	27
4	3	12
5	1	5
計	80人	100人

$$m=\frac{100}{80}=1.25$$

∴ 平均扶養家族数 1.25人

若し変量が連続的であって度数分布の級が級間隔を有する時は、各級間隔の中央の変量値を以て級の代表値として今と同様にして計算するのである。級間隔が大きくなく又集団の単位数が余り少くない時は、このような便法によるために生ずる誤差は極めて小である。第三表より平均賃銀は第六表のようにして計算される。

算術平均値の簡便計算法

計算式(2)は次のように変形する事が出来る。

$$m=\frac{\sum_1^k x_i f_i}{N}=\frac{\sum_1^k (x_i-a+a)f_i}{N},\quad \text{但 } a \text{ は任意の常数}$$

$$=\frac{\sum_1^k (x_i-a)f_i+\sum_1^k af_i}{N},\quad \text{分子第二項}:\sum_1^k af_i=a\sum_1^k f_i=aN$$

$$\therefore =\frac{\sum_1^k (x_i-a)f_i}{N}+a \qquad (3)$$

即ち算術平均値は任意の値 a と各変量との差を加重平均した値に a を加えたものである。故に a を適当に選ぶ事によって偏差 (x_i-a) が簡単な値になるならば、これによって算術平均値の計算は極めて容易に行えるのである。（普通 a は平均値に近いと考えられ、そして加重計算の容易な偏差の得られるような値に決定される。故にこれを仮平均値 working mean という。）

先の平均賃銀の計算（第六表）に於て、中央値 x_i に500円の端数があるため $x_i f_i$ の計算が煩雑となったのである。そこでこの端数を消すために（且平均値に近いと考えられる）$a=10.500$ 円として、簡便式によって平均賃銀を求めると第七表のように簡単になる。

第六表　平均賃銀の計算表

賃銀階数 x_i～x_i			中央値 x_i	人数 f_i	$x_i f_i$
千円		千円	千円	人	千円
5	以上 6	未満	5.5	2	11.0
6	〃 7	〃	6.5	3	19.5
7	〃 8	〃	7.5	6	45.0
8	〃 9	〃	8.5	10	85.0
9	〃 10	〃	9.5	15	142.5
10	〃 11	〃	10.5	13	136.5
11	〃 12	〃	11.5	10	115.0
12	〃 13	〃	12.5	7	87.5
13	〃 14	〃	13.5	5	67.5
14	〃 15	〃	14.5	3	43.5
15	〃 16	〃	15.5	2	31.0
16	〃 17	〃	16.5	1	16.5
17	〃 18	〃	17.5	1	17.5
18	〃 19	〃	18.5	1	18.5
19	〃 20	〃	19.5	1	19.5
計				人 80	千円 856.0

$$m=\frac{856}{80}=10.7$$

∴　平均賃銀 10,700 円

第七表　平均賃銀の計算表

x_i～x_i	x_i	f_i	x_i-a	$(x_i-a)f_i$
	千円	人	千円	千円
	5.5	2	−5	−10
	6.5	3	−4	−12
	7.5	6	−3	−18
	8.5	10	−2	−20
	9.5	15	−1	−15
	10.5	13	0	0
	11.5	10	1	10
	12.5	7	2	14
	13.5	5	3	15
	14.5	3	4	12
	15.5	2	5	10
	16.5	1	6	6
	17.5	1	7	7
	18.5	1	8	8
	19.5	1	9	9
計		人 80		千円 16

$a=10{,}500$ 円

$$m=\frac{16}{80}+10.5=10.7$$

∴　平均賃銀 10,700 円

註　算術平均値よりの偏差 (x_i-m) の総和は 0 である。

即ち

$$\sum_1^N (x_i-m)=(x_1-m)+(x_2-m)+\cdots\cdots+(x_N-m)$$

$$=\sum_{1}^{N} x_i - Nm$$

しかるに $m=\dfrac{\sum_{1}^{N} x_i}{N}$ ∴ $\sum_{1}^{N} x_i = Nm$

これを代入して $\sum_{1}^{N} x_i - Nm = Nm - Nm = 0$

算術平均値のこの性質は極めて重要である。

(ii) 標準偏差（分散）

かくてすべての統計単位の数量値を代表する値が求められたのであるが、次は各単位の有する値がどの程度異るかを知る事が必要であろう。この単位の変量の分布の範囲（即ち分散度）の測定は、四分位偏差 quartile deviation、平均偏差 mean deviation、標準偏差（又は分散）等を求める事によって行われるのであるが、これ等のうち最も多く利用され且茲で必要なものは標準偏差（又は分散）である。

標準偏差 standard deviation は、算術平均値に対する各単位の変量の偏差 deviation の自乗を平均し、それを平方に開いた値であって、通常 σ（sigma ギリシャ文字Σの小文字）を以て示される。故に計算式は次の如くである。

$$\sigma=\sqrt{\frac{(x_1-m)^2+(x_2-m)^2+\cdots\cdots+(x_N-m)^2}{N}}$$

$$=\sqrt{\frac{\sum_{1}^{N}(x_i-m)^2}{N}} \quad (4) \quad (但,\ \sigma>0\ のみをとる)$$

変量が度数分布として与えられている時は

$$\sigma=\sqrt{\frac{(x_1-m)^2f_1+\cdots\cdots+(x_k-m)^2f_k}{f_1+\cdots\cdots+f_k}}$$

$$=\sqrt{\frac{\sum_{1}^{k}(x_i-m)^2f_i}{N}} \quad (5)$$

$$=\sqrt{\sum_{1}^{k}(x_i-m)^2p_i} \quad (5'),\ 但\ p_i=\frac{f_i}{N}$$

すべての単位の変量の分布の範囲を算術平均値を基準に測る場合、算術平均値よりの偏差の合計は0であるためそのままでは分散度を測る事は出来ない。故に偏差を自乗して符号を消してから平均し、後平方根をとって平均的な偏差を求めたものが標準偏差である。

標準偏差の簡便計算法

標準偏差の計算に於ては算術平均値よりの偏差を自

乗しなければならないため、偏差を簡単な値に直す事によって計算は極めて簡易となるであろう。故に標準偏差の簡便計算は次の式によって行われ得る。(5)式より

$$\sigma^2=\frac{\sum_1^k(x_i-m)^2f_i}{N}=\frac{\sum_1^k(x_i-a+a-m)^2f_i}{N}$$

分子: $\sum_1^k\{(x_i-a)+(a-m)\}^2f_i$

$=\sum_1^k\{(x_i-a)^2+2(x_i-a)(a-m)+(a-m)^2\}f_i,$

$(a-m)=\text{const.}$

$\therefore\quad=\sum_1^k x_i-a)^2f_i+2(a-m)\underbrace{\sum_1^k(x_i-a)f_i}_{=(a-m)N}+(a-m)^2\sum_1^k f_i$

$\therefore$(3) より

$\left[\sum_1^k(x_i-a)f_i=(m-a)N=-(a-m)N\right]$

$=\sum_1^k(x_i-a)^2f_i-2(a-m)^2N+(a-m)^2N$

$=\sum_1^k(x_i-a)^2f_i-(a-m)^2N$

故に

$$\sigma^2=\frac{\sum_1^k(x_i-a)^2f_i-(a-m)^2N}{N}$$

$$=\frac{\sum_1^k(x_i-a)^2f_i}{N}-(a-m)^2 \qquad (5)$$

又は

$$=\frac{\sum_1^k\left\{I\left(\frac{x_i-a}{I}\right)\right\}^2f_i}{N}-(a-m)^2$$

$$=\frac{I^2\sum_1^k\left(\frac{x_i-a}{I}\right)^2f_i}{N}-(a-m)^2 \qquad (6')$$

茲に a は仮平均をとり、I は級間隔である。故に(6')式は偏差を級間隔で除して極めて簡単な値にした上で計算するのである。この簡便式によって賃銀分布の標準偏差を求めると第八表の如くである。

第八表　賃銀分布の標準偏差の計算表

x_i	f_i	$\frac{x_i-a}{I}$	$\left(\frac{x_i-a}{I}\right)^2$	$\left(\frac{x_i-a}{I}\right)^2f_i$
千円	人			
5.5	2	−5	25	50
6.5	3	−4	16	48
7.5	6	−3	9	54
8.5	10	−2	4	40
9.5	15	−1	1	15
10.5	13	0	0	0
11.5	10	1	1	10
12.5	7	2	4	28
13.5	5	3	9	45
14.5	3	4	16	48
15.5	2	5	25	50
16.5	1	6	36	36
17.5	1	7	49	49
18.5	1	8	64	64
19.5	1	9	81	81
計	人 80			618

$m=10{,}700$円
$a=10{,}500$円
$I=1{,}000$円

$$\sigma^2=\frac{1.000^2\times618}{80}-(-200)^2$$
$$=7{,}685{,}000$$
$$\therefore\quad\sigma=2{,}772\text{ 円}$$

尚(6)式に於て $a=0$（即ち仮平均を用いない）とすると

$$\sigma^2=\frac{\sum_{i}^{k}x_i^2f_i}{N}-m^2 \qquad (6'')$$

即ち標準偏差は算術平均値よりの偏差をとらなくとも、変量の自乗の加重平均値より算術平均値の自乗を引く事によって求める事が出来る。若し変量 x が簡単な値である時はこの計算式によるとよい。今扶養家族数分布の標準偏差をこの式で求めると第九表の如くである。

第九表　分布の扶養家族数標準偏差の計算表

x_i	f_i	x_i^2	$x_i^2f_i$
人	人	人	人
0	25	0	0
1	28	1	28
2	14	4	56
3	9	9	81
4	3	16	48
5	1	25	25
計	80 人		238 人

$$m=1.25\text{人}$$

$$\sigma^2=\frac{238}{80}-1.25^2$$

$$=1.4125$$

$$\therefore\ \sigma=1.19\text{人}$$

標準偏差の自乗、即ち算術平均値に対する各単位の変量の偏差の平方平均を分散 variance という。故に分散（V で表す）は

$$V=\sigma^2=\frac{\sum_{i}^{k}(x_i-m)^2f_i}{N} \qquad (7)$$

分散度としての実際的意義は後述する如く標準偏差の方が重要であるが、平方根は数学的取扱が不便なため に理論的には分散が主に用いられる。

標準偏差の値が小さい時は、算術平均値に対する偏差（x_i-m）が小さいものが多い、即ち大多数の変量は平均値の周りに集つていて大差がない事を意味するのであるから、その算術平均値はよく凡ての統計単位の変量を代表する事が出来るであろう。そして標準偏差の値が大きい時は、平均値より相当異った値の単位が多いのであるから、算術平均値の代表性は小さいといわねばならぬ。かくて標準偏差の大小は算術平均値がどの程度よく統計集団の代表値たり得るかを表す尺度であると見る事が出来る。しかしながらその場合戦後の賃銀額が何千円、何万円という時の賃銀分布の標準偏差は、戦前の賃銀額何十円、何百円という時のそれよりも大きい事は、特に標準偏差を計算し比較しなくとも明らかであって、その値の大小は平均値の代表性の尺度とはなり得ないのである。又賃銀分布の標準偏差と扶養家族数分布の標準偏差とでは、変量の性質が全然異るため比較は不可能である。従って変量の大

いさが著しく異る場合や変量の種類が異る時は、標準偏差を直接比較しても変量の分散度の大小従って平均値の代表性の大小を判断する事は出来ないのである。

かくて平均値の代表性を表すには、標準偏差が算術平均値よりの偏差によって計算した分散度であるため、標準偏差を算術平均値で除して相対化した値の方がよいのである。これを変異係数 coefficient of variation といい c（又はC、V）で表す。

$$c=\frac{\sigma}{m} \qquad (8)$$

変異係数が小さい程変量の分布は平均値の周りに密であり、従ってその平均値は代表性が大である事は明らかであろう。今第二表の扶養家族数分布の変異係数 c_1 と第三表の賃銀分布の変異係数 c_2 を比較すると、賃銀分布の方が分布の状態は相対的に稠密である事が判るのである。

$$c_1=\frac{1.19}{1.25}=0.951=95.1\%$$

$$c_2=\frac{2,772}{10,700}=0.259=25.9\%$$

(iii) 正規分布

度数分布に於て量的標識は大小の差を有する数量値であり一定の自然的序列を有するために、各標識を有する単位の数即ち度数の変化は一定の分布形態を決定するのである。この度数分布の形態は、変量（標識）と度数との関係を変量の変化に従って度数が変化する函数であると考える事によって、一定の数学式によって表現する事が出来、その数学式を度数分布函数という。[註] そして度数分布の形態は若干の規則的な類型に従うのが実際であるため、何れの類型に属するかが明らかにされた以上は、その型の分布形態を表す数学式のパラメーター（媒介変数 parameter）の値を求める事によって、その度数分布の形態は十分に規定し得る事となるのである。

註 かくして得られた数学式の表す理論的度数分布は、必ずしも実際の度数分布と完全には一致しないのであつて、従つて分布函数は量的構造を近似的に表すに過ぎないの

である。しかし場合によつてはこの理論的度数分布は、実際の度数分布の観察数の不足による不完全さを補ひ、理想的な度数分布の状態を表すものと考える事が出来るのである。例えば既に述べたように、連続的変量の場合でも観察単位数が少いため、度数分布は非連続的にしか取扱ひ得ないのであるが、それに当嵌めた度数分布は連続的な度数曲線を描き、連続的変量に対する度数の分布を与えるのである。

度数分布の類型としては対称分布（正規分布）、非対称分布、J字型分布、U字型分布、複峰性分布等に分け得るのであるが、茲では標本調査法に於て必要な対称分布（正規分布）のみに就て説明しよう。

対称分布は度数の集中点が全分布範囲の中央にあって、それを中心に度数が左右に対称的に減少する場合である。（第一図参照）度数分布が対称分布をなす場合は自然現象に多く、社会現象の場合は大部分非対称分布をなすのである。理論的には対称分布は度数分布のうち一番重要なものであって、早くから最もよく研究されている分布である。対称分布は普通正規分布 normal distribution といわれ、対称分布函数の表す曲線を正規曲線 normal curve（又は Gauss 曲線、（正規）誤差曲線）という。

正規分布を表す函数は変量を x、度数を y[1]とすると次の如くである。

$$y=f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (9)$$

この式に於てパラメーターは m と σ とであって、m は度数分布の算術平均値、σ はその標準偏差である。[2]（そして常数 π は円周率であってその値は 3.1459…、e は Napier の常数であって 2.71828…である。）或る一の正規度数分布をこの正規函数で表すためには、その度数分布の算術平均値及び標準偏差を求めその値を(9)式に代入すればよいのであって、m の位置、σ の大小によって異る正規分布が表されるのである。そして算術平均値は度数の集中点（即ち最頻値）に一致するのであって、[3]その相対的度数は $\frac{1}{\sqrt{2\pi}\sigma}$ である。即ち

$$x=m,\quad y=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(m-m)^2}{2\sigma^2}}=\frac{1}{\sqrt{2\pi}\sigma} \qquad \because\ (e^0=1)$$

(1) (9)式は正確には相対的度数を表すのであつて、νの値は単位総数N個中或る変量xを有する単位の数fの存在の割合f/Nを示しているのである。故に(9)式に単位総数Nを乗ずる事によつて、任意の変量を有する単位の数が求められるのである。

(2) 先づ正規分布の算術平均値がmである事を証明しよう。(2)式より算術平均値は変量と相対的度数の積の総和であるが、この場合は変量xが連続的変量であるため総和は積分となるのであり、そして積分限界は一般性を得るために$-\infty$より$+\infty$迄とするとよいのである。故に

$$\int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}dx$$

変数変換：$t=\dfrac{x-m}{\sqrt{2}\sigma}$，$x=\sqrt{2}\sigma t+m$，$dt=\dfrac{dx}{\sqrt{2}\sigma}$

$$\therefore \quad = \int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}(\sqrt{2}\sigma t+m)e^{-t^2}dt$$

$$= \frac{\sqrt{2}}{\sqrt{\pi}}\sigma\underbrace{\int_{-\infty}^{\infty} te^{-t^2}dt}_{\because\ 0} + \frac{m}{\sqrt{\pi}}\underbrace{\int_{-\infty}^{\infty} e^{-t^2}dt}_{\sqrt{\pi}} = m$$

（被積分函数が奇函数なるため）

次に標準偏差がσなる事を証明するのであるが、この場合も(5)式より算術平均値の偏差の平方と相対的度数との積の総和の平方である事からして、今と同様次のようにして求められる。

分散：$\displaystyle\int_{-\infty}^{\infty}(x-m)^2f(x)dx$

$$= \int_{-\infty}^{\infty}(x-m)^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}dx$$

変数変換：$t=\dfrac{x-m}{\sqrt{2}\sigma}$，$(x-m)^2=2\sigma^2t^2$，$dt=\dfrac{dx}{\sqrt{2}\sigma}$

$$\therefore = \int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}2\sigma^2t^2e^{-t^2}dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\underbrace{\int_{-\infty}^{\infty} t^2e^{-t^2}dt}_{\frac{\sqrt{\pi}}{2}} = \sigma^2$$

∵ 被積分函数が偶函数なるため

$$\left(\int_{-\infty}^{\infty} t^2e^{-t^2}dt = 2\int_0^{\infty} t^2e^{-t^2}dt,\right.$$

$$\left. = \Big[-te^{-t^2}\Big]_0^{\infty} + \underbrace{\int_0^{\infty} e^{-t^2}dt}_{\frac{\sqrt{\pi}}{2}} = \frac{\sqrt{\pi}}{2}\right)$$

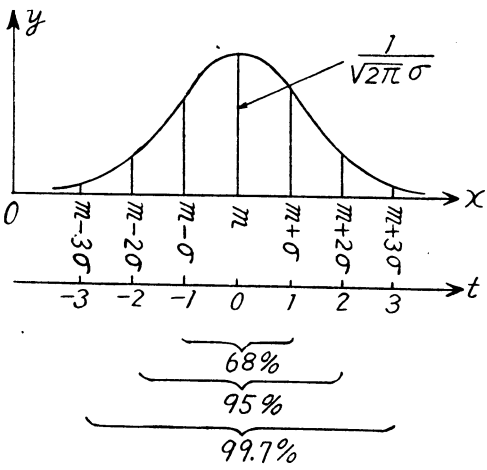
故に　標準偏差：σ

(3) この事は数学的には(9)式を微分して極大点を求める事によつて証明されるのであるがその説明は略する。

正規分布に於て$m-\sigma$より$m+\sigma$迄の間の変量を有する単位の数の割合$P(m-\sigma\leqq x\leqq m+\sigma)$は、その

区間に於ける正規曲線下の面積である。故に正規函数 $f(x)$ を $m-\sigma$ より $m+\sigma$ 迄積分する事によってその値は得られるのである。そして正規函数 $f(x)$ は算術平均値 m に関して対称であるため、m より $m+\sigma$ 迄の区間の面積を知れば、それを二倍する事によって区間（$m-\sigma$, $m+\sigma$）の面積は求められる。（第一図参照）故に

第一図　正規分布



$$P(m-\sigma\leqq x\leqq m+\sigma)=\int_{m-\sigma}^{m+\sigma}f(x)dx=2\int_{m}^{m+\sigma}f(x)dx$$

$$=2\int_{m}^{m+\sigma}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}dx \quad (10)$$

しかしながらこの積分は m と σ の値が決られば求められない。そこで正規函数 $f(x)$ に於て $t=\frac{x-m}{\sigma}$ なる変換を行って $\varphi(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ という形に直す。（この事を正規函数（乃至は変数）の標準化という）この場合 m 及び $m+\sigma$ は新変数 t に於てはそれぞれ 0 及び 1 となる。蓋し

$x=m$ の時は　$t=\frac{m-m}{\sigma}=0$

$x=m+\sigma$ の時は　$t=\frac{m+\sigma-m}{\sigma}=1$

であるからである。（この関係は第一図の x 軸と t 軸との比較によって了解せよ。）すると(10)式は

$$=2\int_{0}^{1}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \quad (10')$$

となって m、σ を含まないためその値が求められるのであり、それは 0.68268 である。要するに

$$P(m-\sigma\leqq x\leqq m+\sigma)=\int_{m-\sigma}^{m+\sigma}f(x)dx=2\int_{m}^{m+\sigma}f(x)dx$$

$$=2\int_0^1 \varphi(t)\,dt=0.68268$$

同様にして

$$P(m-2\sigma\leqq x\leqq m+2\sigma)=2\int_m^{m+2\sigma} f(x)dx$$
$$=2\int_0^2 \varphi(t)dt=0.95450$$
$$P(m-3\sigma\leqq x\leqq m+3\sigma)=2\int_m^{m+3\sigma} f(x)dx$$
$$=2\int_0^3 \varphi(t)dt=0.99730$$

かくて正規分布に於ては $m\pm\sigma$ の間に全単位数の68%強が存在し、$m\pm2\sigma$ の間には全単位数の95%強が、$m\pm3\sigma$ の間には99.7%強が存在するのである。

チェビシェフの定理 (Tchebycheff's Theorem)

今述べた正規分布に於ける標準偏差と度数分布との関係は、一般の度数分布に於ては次の如くである。

変量	度数
x_1	f_1
x_2	f_2
……	……
x_k	f_k
計	N

上の度数分布に於てその算術平均値を m、標準偏差を σ とすれば

$$m-\lambda\sigma\leqq x\leqq m+\lambda\sigma,\ (\text{但}\ \lambda\ \text{は}\ 1\ \text{より大なる実数})$$

を満足する単位の数 N_1 は

$$N_1\geqq\left(1-\frac{1}{\lambda^2}\right)N \qquad (11)$$

従つてその割合は

$$\frac{N_1}{N}\geqq\left(1-\frac{1}{\lambda^2}\right) \qquad (11')$$

即ち算術平均値を中心に標準偏差 σ の λ 倍の幅を附した範囲内には、全体の $100\left(1-\frac{1}{\lambda^2}\right)$% 以上の単位が含まれている事を示しているのである。これをチェビシェフの定理という。そして度数分布が連続的な度数分布函数によつて表されている場合は、チェビシェフの定理は次の如く表される。$f(x)$ を連続的分布函数としその単位総数を N、算術平均値を m、標準偏差を σ とすると

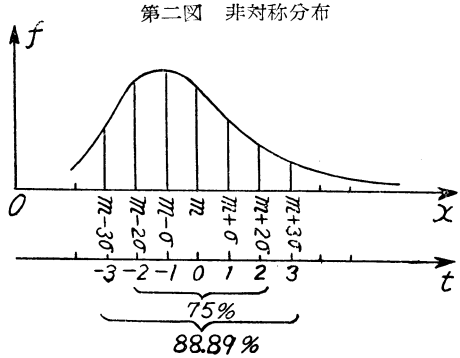
$$\int_{m-\lambda\sigma}^{m+\lambda\sigma} f(x)dx\geqq\left(1-\frac{1}{\lambda^2}\right)N \qquad (12)$$

なる関係が成立つのである。(第1図参照)

今 λ に 1 2 3 …… 10 の値を与え、その範囲内に存在する単位の割合を (11') 式より求めると次の如くである。

	範囲	総度数に対する割合 %
$\lambda=2$	$m\pm2\sigma$	75.00
$\lambda=3$	$m\pm3\sigma$	88.89
⋮	⋮	⋮
$\lambda=5$	$m\pm5\sigma$	96.00
⋮	⋮	⋮
$\lambda=10$	$m\pm10\sigma$	99.00

第二図　非対称分布



これを正規分布の場合の度数とその分布範囲との関係に比較する事により、若し度数分布が正規分布をなすならば、同じ割合の単位数を含む範囲が著しく狭くなる事が判るであろう。

変数の標準化

統計集団の二の量的性質の比較（即ち二の度数分布の比較）に於て、両集団の量的標識（変量）の測定単位が異る時は、そのままでは比較する事は出来ない。例えば賃銀階級別度数分布と勤続年数別度数分布とでは、一方の測定単位は金額であるが他方のそれは年齢であるため、直接比較は不可能である。又仮令測定単位は同じであっても、一方の変量が他方の変量と著しくかけ離けている時も亦、直接比較し得ない事は、戦後の家計費階級別度数分布と戦前の夫との比較を考える事によって明らかであろう。

しかし次のような変数変換を行う時は、何れの度数分布の変量も無名数の、且同じスケールの単位で表されるので比較が可能となるのである。即ち上の度数分布に於て算術平均値を m、標準偏差 σ とする時、変数変換

$$t_k=\frac{x_k-m}{\sigma} \qquad (13)$$

を施すと、新変数 t によって表される度数分布に於は算術平均値0、標準偏差1である。（この事を変数の標準化という。）

x	f
x_1	f_1
x_2	f_2
⋮	⋮
x_k	f_k
計	N

t	f
t_1	f_1
t_2	f_2
⋮	⋮
t_k	f_k
計	N

これを証明すると

$$\text{算術平均値}: \frac{\sum_1^k t_i f_i}{N} = \frac{\sum_1^k \left(\frac{x_i - m}{\sigma}\right) f_i}{N} = \frac{\sum_1^k x_i f_i}{N\sigma} - \frac{m \sum_1^k f_i}{N\sigma}$$
$$= \frac{m}{\sigma} - \frac{m}{\sigma} = 0$$

$$\text{分散}: \frac{\sum_1^k (t_i - 0)^2 f_i}{N} = \frac{\sum_1^k \left(\frac{x_i - m}{\sigma}\right)^2 f_i}{N} = \frac{1}{\sigma^2} \frac{\sum_1^k (x_i - m)^2 f_i}{N}$$
$$= \frac{\sigma^2}{\sigma^2} = 1$$

∴ 標準偏差＝1

そしてこの変数変換の意義は $(x_i - m)$ で原点を m に移し、$\frac{1}{\sigma}(x_i - m)$ でそれを σ を測定単位として表したものである。（第11図でいえば x 軸が t 軸に変る事である。）

かくて新変数 t は測定単位に無関係の無名数となり、(13)式の分子、分母共に同じ測定単位の数値であるから）且チェビシェフの定理により -10 より $+10$ の範囲に全度数の殆ど全部が存在するのであるから、このような変数変換を施された度数分布は相互比較が可能となるのである。

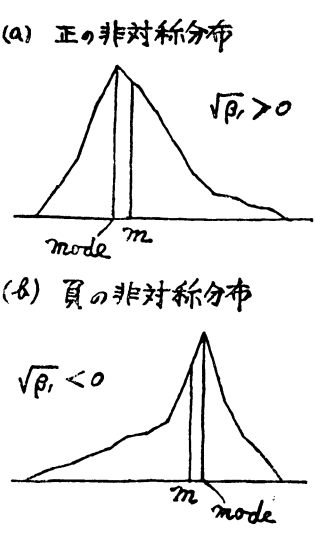
(iv) 歪度、尖峰度[註]

以上算術平均値、標準偏差を求める事により或は又分布函数を規定する事によって、度数分布の特徴を簡約に示す値が得られたのであるが、尚更に非対称分布の場合には算術平均値（又は最頻値）を中心とする度数分布の釣合の状態、即ち非対称の方向と程度を表す歪度 skewness が計算される。その計算方法の一は次の如くである。

$$\pm\sqrt{\beta_1} = \frac{\frac{1}{N}\sum_1^k (x_i - m)^3 f_i}{\sigma^3} \qquad (14)$$

（但 $\sqrt{\beta_1}$ は分子と同符号にとる。この符号が非対称性の方向を示すのである。）

第 三 図



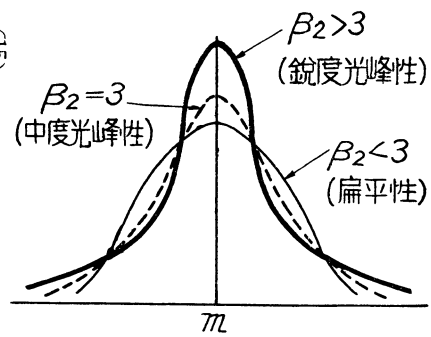
正規分布の場合は、$\sqrt{\beta_1} = 0$ である。

又度数分布曲線の尖りの大小（それは算術平均値の周りに於ける変量の集中度を表す）を示す尖峰度 kurtosis を求める事も行われる。それは

$$\beta_2=\frac{\frac{1}{N}\sum_{i}^{k}(x_i-m)^4 f_i}{\sigma^4} \qquad (15)$$

によって計算される。正規分布の場合は $\beta_2=3$ である。

第四図



註 これ等の測度は本稿に於ては殆ど必要でないのでその具体的意義の説明は省略する。図表によつて大略の概念を得、詳細は統計書を参照されたし。

(3) 質的構造の簡約表章

統計集団の構造を質的な集団性に就て明らかにした場合、（例えば人口集団を「体性」「職業」に就て構造分析した場合）標識を異にする統計単位は質的に異った現象であるため、統計集団全体に於て各標識（質）の単位がどれ丈あるかその存在の割合——即ち比例数——を求める事によって、統計集団の構造を簡明に表現する事が出来るであろう。[註] このような統計集団の構造を示す比例数又は相対的度数又は頻度 relative Häufigkeit od. Häufigkeit ともいわれ、一定の条件の下に於ては経験的確率として考える事が出来るのである。

今質的集団性Aに就て統計集団を観察した場合、統計集団は標識 $a_1 a_2 \cdots\cdots a_k$ の部分集団より成り、その各 x の単位数は $n_1 n_2 \cdots\cdots n_k$ である事が明らかにされたとすると、（そして総単位数 $n_1+n_2+\cdots\cdots+n_k=N$ とする）各標識の単位の存在の割合（相対的度数）は

$$p_1=\frac{n_1}{N},\ p_2=\frac{n_2}{N},\ \cdots\cdots,\ p_k=\frac{n_k}{N}$$

である。そしてすべての標識の相対的度数の和は1である事は明らかであろう。

$$p_1+p_2+\cdots\cdots+p_k=\frac{n_1+n_2+\cdots\cdots+n_k}{N}=1$$

註 尚比例数は統計集団の量的構造の記述にも用いられる。

前の例でいえばその事業所に於て賃銀額が五千円から六千円迄の者の割合が幾ら、六千円から七千円迄の者の割合が幾ら等々として、80人の従業員中各賃銀階級に属するものゝ割合を求める事も、その集団の量的構造の認識に大いに役立つであろう。しかしながら量的構造の場合は部分集団の数が多いために、比例数よりも算術平均値、標準偏差等による方が遙かに簡明に集団の構造を表示する事が出来るので、普通量的構造の記述には比例数は用いられないのである。

変量	度数
1	n_1
0	n_2
計	N

→

標識	単位数	比率
α	n_1	p
β	n_2	q
計	N	1

質的構造の場合は標識が数量的でないためにこれ以上の数理的取扱は不可能である。しかしながら統計集団の質的構造が、α標識とα標識でないもの即ちβ標識の二の部分集団（それぞれの大きさをn_1 n_2、比率をp qとする）に分けられる時は、α標識に1、非α標識従ってβ標識に0なる値を与えると、統計集団は非連続的変量0及び1の二の級より成り、度数がn_1 n_2の度数分布によって表されて質的構造は量的構造と見る事が出来、従ってそれの算術平均値及び標準偏差等を求める事が出来るであろう。今算術平均値mを計算すると

$$m=\frac{1\times n_1+0\times n_2}{n_1+n_2}=\frac{n_1}{N}=p$$

となり、平均値mは（1を与えられた）標識αの比率pと一致する。次に標準偏差σを計算すると

$$\sigma^2=\frac{(1-p)^2n_1+(0-p)^2n_2}{n_1+n_2},\qquad \text{しかるに } p+q=1 \ \therefore\ 1-p=q$$

$$\therefore\ =\frac{n_1}{N}q^2+\frac{n_2}{N}p^2=pq^2+qp^2$$

$$=pq(q+p)=pq$$

$$\therefore\ \sigma=\sqrt{pq}$$

であり、標準偏差σはα標識の比率とβ標識の比率との積の平方根である。今度はβ標識に1を与え非β標識即ちα標識に0を与えると、算術平均値mはβの比率qに一致する。

$$m=\frac{0\times n_1+1\times n_2}{n_1+n_2}=\frac{n_2}{N}=q$$

標準偏差σは前と同様 $\sigma=\sqrt{pq}$ である。

$$\sigma^2=\frac{(0-q)^2n_1+(1-q)^2n_2}{n_1+n_2}=\frac{n_1}{N}q^2+\frac{n_2}{N}p^2$$

$$=pq^2+qp^2=pq(q+p)=pq$$

$$\therefore \sigma=\sqrt{pq}$$

以上により明らかな如く統計集団の質的構造が二の部分集団より成る場合は、0、1の標識を与える事によって量的構造と同様に取扱う事が出来るのであって、この事は実際的意義はとに角も理論的には極めて重要である。

(4) 相関々係の簡約表章

統計集団の性質は、一方向の集団性に就て明らかにされた統計集団の数量的構造として表示されるのであるが、それを更に他の方向の集団性に就て二重に構造分析する事によって、統計集団のより詳細な性質が明らかにされるのである。例えば「一カ月間賃銀額」なる集団性に就て構造分析された勤労者集団を、更に「体性」によって分析する場合がこれである。元来男子と女子とは作業能率の点から、又仕事の性質を異にする事から当然賃銀水準を異にするのであって、同じく九千円の賃銀であっても男子の場合は中以下の賃銀であるが、女子の場合は中以上であるといわねばならず、従って「賃銀額」によって分けられた各部分集団には、このような質的に異る単位が混在しているので、そのままでは勤労者集団の性質を十分に示しているとはいえないのである。故に勤労者集団の「賃銀額」構造を更に「体性」によって分析し、各部分集団を男女別に細分して当該勤労者集団の男子の賃銀構造と、女子の賃銀構造とを別個に表示しなければならないのである。

一般に集団性相互の事物的、論理的関連に基いて二以上の集団性を関係附ける事によって、統計集団はより小さな、しかしより同質的な単位より成る部分集団に細分され、それ等相互の数量的関係によって統計集団のより詳細な性質が明らかにされるのである。そしてその場合関係附けられる二の集団性が何れも量的な性質のものである時は、そこに見られる統計的関係を相関々係 correlation といい、この重複度数分布の表

示を相関表という。

一般に相関表の形式は第十表の如くである。最下及び最右の合計欄はそれぞれ、統計集団をα集団性及びβ集団性に就て分析した時の度数分布を表すのであり、そして中央の欄は標識x_iを有する単位$f_{i\bullet}$個より成る部分集団がβ集団性に就て分析された時、y_1を有する単位f_{i1}、y_2を有する単位f_{i2}…y_lを有する単位f_{il}より成る部分集団に細分された事を示しているのである。

第十表　相　関　表

β \ α	x_1	x_2	…	x_i	…	x_k	計
y_1	f_{11}	f_{21}	…	f_{i1}	…	f_{k1}	$f_{\bullet 1}$
y_2	f_{12}	f_{22}	…	f_{i2}	…	f_{k2}	$f_{\bullet 2}$
⋮	⋮	⋮		⋮		⋮	⋮
y_j	f_{1j}	f_{2j}	…	f_{ij}	…	f_{kj}	$f_{\bullet j}$
⋮	⋮	⋮		⋮		⋮	⋮
y_l	f_{1l}	f_{2l}	…	f_{il}	…	f_{kl}	$f_{\bullet l}$
計	$f_{1\bullet}$	$f_{2\bullet}$	…	$f_{i\bullet}$	…	$f_{k\bullet}$	N

但し　$\sum_{j=1}^{l} f_{ij} = f_{i\bullet}$　又　$\sum_{i=1}^{k} f_{ij} = f_{\bullet j}$

$$\sum_{i=1}^{k} f_{i\bullet} = N \qquad \sum_{j=1}^{l} f_{\bullet j} = N$$

$$\therefore \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = N$$

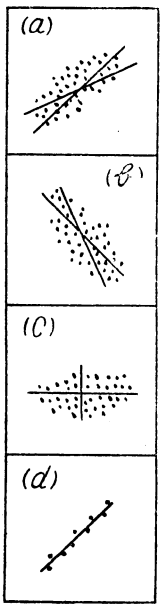
相関表に於ては各集団性に従属する標識は何れも変量であり一定の自然的序列を有するため、相関度数分布は一定の分布形態を確定する。そしてそれは一方向の集団性αに従属する変量をx、他の方向の集団性βに従属するそれをyで表す事によって、xy平面上の点の分布として図示する事が出来るのである。

相関度数分布の形態には次の場合がある。

(1)　x、y共に小なるところからx、y共に大なるところにかけて分布する傾向を有する場合。（第五図(a)）これは変量xの大きな値を有する統計単位は同時に又、より大きな変量yの値を有するものが多い場合である。この相関々係を正相関という。

(2)　xは小なるもyは大なるところから、x大y小なるところにかけて分布する傾向を有する場合。（第五図(b)）これは一般にxのより大きな値を有する単位は逆により小さなyの値を有する場合である。この相

第五図



関々係は負相関といわれる。

(3) 以上のような特定の方向に集中する傾向を有せず、全般的に一様に分布する場合。（第五図(c)）これはxの大なる値を有する単位が、多くyに就てはより大（又はより小）であるという如き一定の傾向を有しない場合、即ちxとyとは相互に無関係（独立）なる場合である。この時は相関々係なし（又は無相関）という。

(4) 相関分布の中心を走る線（回帰線）の周りの分布の範囲が小くなった極限の場合。（第五図(d)）これはxの或る値x_iを有する単位はすべてyに就ても只一の値y_iしか有しない場合であって、xとyとは一義的な相互関係に於てある場合である。この相関々係を完全相関という。

それではこのような相関々係の簡約表章は如何にして行われるのであろうか。[注] 先ず相関々係の形態を規定する事が必要である。相関分布に於ては変量xの一の値x_iに対する変量yの値は、一価的でなくして多価的であり一定の範囲内の値をとるのである。このようなxとyとの相互関連の形態は、各x_iに対するyの平均値を求める事によってx、yの関係を一価的な函数関係となし、それを表す数学式を求める事によって規定する事が出来るのである。この場合xの変化に対するyの平均値の変化が略々直線的と看做し得る時は、その相関々係を直線相関 linear correlation といい、この平均値を結ぶ線に直線を当嵌めた場合その直線を回帰直線 regression line という。若しxに対するyの平均値の動きが直線的でなく曲線的であるならば、その相関々係は曲線相関 non-linear correlation であるといい、それに当嵌められた曲線を回帰曲線 regression curve という。以上はxの変化に対するyの変化としてx、yの相関々係を捉えたのであるが、逆にyに対するxの変化としてx、yの相関々係を捉える事が出来る。そして両者は（完全相関でない限り）異るので、一の相関分布に於ては二の回帰線が得られるのである。

この回帰直線（又は曲線）を求める事によって相関

々係の形態は簡明に表されるのであるが、それは変量xに対する変量yの平均的な変化を表すのみであって、yは回帰線の周囲に一定の範囲で分布しているのである。回帰線を中心に相関分布の範囲が狭く回帰線の周りに密に分布している時は、xの変化に伴うyの変化は回帰線によって十分代表し得るであろう。そしてこの場合はxの変化とyの変化は相互によく関連し合って居り、統計集団のこの二の性質は相互に密接な関連を有する事を示しているのである。しかるに回帰線の周りに粗く広範囲に亘つて分布している時は、xとyとの関係は回帰線によっては十分代表し得ないであろう。そしてこの場合は統計集団の二性質の相互関連は左程密接ではないのである。従って次にこの相関分布の範囲を規定しなければならないのであって、それの測度が相関係数 correlation coefficient（曲線相関の場合は相関比 correlation ratio）である。

かくて相関々係の簡約表章は回帰直線（又は回帰曲線）と相関係数（又は相関比）の二によってなされるのであるが、茲では本稿に於て必要な相関係数のみを説明しよう。

註 関係附けられる集団性が何れか一つ又は双方共に質的性質の場合は、部分集団の相互関聯の簡約表章は比例数によって行われる。

(i) 相関係数

相関係数ρ（ロー rho）は次の式によって計算される。

$$\rho=\frac{\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\bar{x})(y_j-\bar{y})f_{ij}}{\sigma_x\sigma_y} \qquad (16)$$

又は

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} \qquad (16')^{2)}$$

但し

$$\bar{x}=\frac{\sum_{i}^{k}x_if_{i\cdot}}{N},\quad \sigma_x=\sqrt{\frac{1}{N}\sum_{i}^{k}(x_i-\bar{x})^2f_{i\cdot}}$$

… α 集団性に就ての度数分布の平均値、標準偏差

$$\bar{y}=\frac{\sum_{j}^{l}y_jf_{\cdot j}}{N},\quad \sigma_y=\sqrt{\frac{1}{N}\sum_{j}^{l}(y_j-\bar{y})^2f_{\cdot j}}$$

… β 集団性に就ての度数分布の平均値、標準偏差

(1) (16)式に於ける $\sum_{i=1}^{k}\sum_{j=1}^{l}$ の意義は次の如くである。便宜上先づ相関表に於ける度数の計算を例にとつて説明する。二方向の集団性に就て二重に構造分析された統計集団の総単位数 N は、各部分集団の単位数 f_{ij} の合計である。故に相関表の左端の列（縦欄）より順次加えて行くと（第十表参照）

$$N=(f_{11}+f_{12}+\cdots+f_{1l})+(f_{21}+f_{22}+\cdots+f_{2l})+\cdots$$
$$\cdots+(f_{k1}+f_{k2}+\cdots+f_{kl})$$

この場合各括弧内の f は添数 i は同じであるが j は異り、それは1から l 迄あるのである。これを和の記号 Σ を使つて略記すると

$$=(\sum_{j=1}^{l}f_{1j})+(\sum_{j=1}^{l}f_{2j})+\cdots\cdots+(\sum_{j=1}^{l}f_{kj})$$

これは相関表に於て各列の度数の合計即ち変量 x_i を有する部分集団の大いさである。故に

$$=f_{1\cdot}+f_{2\cdot}+\cdots+f_{k\cdot}=\sum_{i=1}^{k}f_{i\cdot}$$

と書ける。そして今述べたように $f_{i\cdot}=\sum_{j=1}^{l}f_{ij}$ であるため、これを代入して

$$N=\sum_{i=1}^{k}\sum_{j=1}^{l}f_{ij}$$

要するに $\sum_{i=1}^{k}\sum_{j=1}^{l}f_{ij}$ は相関表の各列の度数を合計し、その結果（第十表最下欄）を更に横に合計する事を意味するのである。従つて $\sum_{j=1}^{l}\sum_{i=1}^{k}f_{ij}$ は逆に先づ相関表の各行（横欄）の度数を合計し、その結果（第十表最右欄）を縦に合計する事を意味するのである。

以上により明らかな如く(16)式の分子は先づ各列毎に $(x_i-\bar{x})(y_j-\bar{y})f_{ij}$ を合計し、次でその結果を横に合計したものである。

(2) 二方向の集団性による統計集団の観察結果が相関度数分布に整理されず、各統計単位毎に個々的に N 個の変量の組 $(x_1,y_1),(x_2,y_2),\cdots\cdots(x_N,y_N)$ として与えられている時は、平均値よりの偏差の積 $(x_i-\bar{x})(y_i-\bar{y})$ は N 個あるので、(16)式の分子は $\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$ となり(16′)式が得られるのである。(16′)式の方が簡単で見易いので以下の説明はこれによる事にする。

この相関係数の算式の意義は次の如くである。α 集団性に従属する変量 x と β 集団性に従属する変量 y とが、測定単位を異にする時は直接比較する事は不可能である。（例えば学生の身長と体重に就ての相関表に於て、cm/mの身長とkg/gの体重とは直接比較し得ない。）

故に各変量に変数変換を施して標準化しなければならぬ。そこで次の変数変換を行い新変数をそれぞれX、Yとする。

$$X_i=\frac{x_i-\bar{x}}{\sigma_x},\qquad Y_i=\frac{y_i-\bar{y}}{\sigma_y}$$

この変数変換の相関図表に於ける意義は、x y直交座標軸が右上方に移動し点$(\bar{x},\bar{y})$の位置（これは相関分布の中心である事は、$\bar{x}$、$\bar{y}$がそれぞれ平均値である事より明らかであろう。）にXY新座標軸の原点が来る事である。

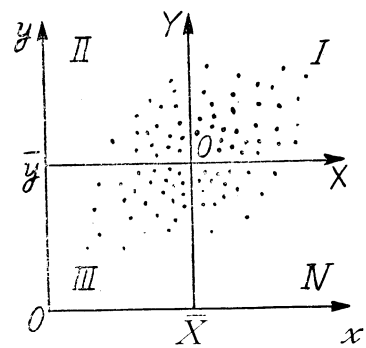
新相関図表に於て、若し相関分布が正相関即ち左下より右上にかけて分布する傾向を有する時は、第一象限と第三象限に多くの単位が存在しているのに対して、第二、第四象限には少しの単位しか存在しないであろう。そして各単位の有する変量の積（$=X_i\cdot Y_i$）をとると、第一、第三象限にある単位の値は正であり、（蓋し第一象限に於ては$X,Y\bigvee 0$ 第三象限に於ては$X,Y\bigwedge 0$ なるため）第二、第四象限にある単位の値は負である。（蓋し第二象限に於ては$X\bigwedge 0,Y\bigvee 0$ 第四象限に於ては$X\bigvee 0,Y\bigwedge 0$ なるため）今この各単位の有する二変量の積をすべての単位に就て合計すると、積の値が正である第一、第三象限に、それが負である第二、第四象限よりも多くの単位が存在するのであるから、その結果は正の値が得られるであろう。即ち

$$\sum_{i=1}^{N}X_iY_i=\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\bigvee 0$$

そしてxとyとの相互関連が密接であればある程、各xの値に対するyの値の変化の範囲が小であるから、第一、第三象限により多くの単位が存在し、第二、第四象限にある単位の数は少くなるためこの値は益〻大きくなるであろう。

逆に相関分布が負相関即ち左上より右下にかけて分布する傾向を有する場合は、新相関図表の第二、第四

第六図



象限に多くの単位が存在し、第一、第三象限には少しの単位しか存在しないであろう。従って各単位の有する変量の積を合計するとその結果は負の値となり、そしてxとyとの相関々係が密接である程、第二、第四象限により多くの単位が存在するため、この値はより小（絶対値はより大）となるであろう。又相関分布が無相関即ち特定の傾向を有せず一様に分布する時は、特定の象限に多くの単位が存在する事なく、第一、第二、第三、第四の各象限に略々同数の単位が、且X軸、Y軸を中心に大体対称的に存在するため、各単位の有する変量の積の合計は正の値と負の値とが相殺されて0に近くなるであろう。

かくて $\sum_{i=1}^{N} X_i Y_i$ を単位総数Nで除して一単位当り平均を求めると、その符号により相関々係の方向を示し、その絶対値の大小によって相関々係の強度を表す事が出来るのである。故に相関係数は次の式によって求められる。

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)=\frac{\frac{1}{N}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}$$

即ち(16')式が得られたのである。

(16')式によって計算される相関係数の絶対値$|\rho|$は無相関の場合の0より、相関度が高まるに従って増大するのであるが、1以上の値をとる事はないのである。即ち

$$|\rho|\leqq 1,\quad (-1\leqq\rho\leqq 1)\qquad (17)$$

これを証明するには先ず $\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma_x}\pm\frac{y_i-\bar{y}}{\sigma_y}\right)^2$ を考える。するとこれは自乗されているためその値は正である。次にこれを展開すると

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma_x}\pm\frac{y_i-\bar{y}}{\sigma_y}\right)^2\\&=\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{x_i-\bar{x}}{\sigma_x}\right)^2\pm 2\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)+\left(\frac{y_i-\bar{y}}{\sigma_y}\right)^2\right\}\\&=\frac{1}{\sigma_x^2}\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right\}\pm 2\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}\\&\qquad+\frac{1}{\sigma_y^2}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2\right\}\\&=\frac{\sigma_x^2}{\sigma_x^2}\pm 2\rho+\frac{\sigma_y^2}{\sigma_y^2}=2\pm 2\rho=2(1\pm\rho)\geqq 0\end{aligned}$$

$\therefore\ 1\pm\rho\geqq 0\quad\therefore\ -1\leqq\rho\leqq 1$

そして$|\rho|=1$, $(\rho=\pm 1)$の場合は、相関々係の密接な極限の場合即ち各x_iに対するyは只一しかなく、x、

yの相互関連は函数関係として捉え得る場合である事は容易に考え得るであろう。

(16)式（又は(16')式）の分子をx、yの共変量 covariance とよび、Cov (x,y) で表す。

$$\mathrm{Cov}(x,y)=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\bar{x})(y_j-\bar{y})f_{ij} \qquad (18)$$

又は

$$\mathrm{Cov}(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \qquad (18')$$

共変量は相関分布の中心$(\bar{x},\bar{y})$を基準とする、各単位の有する変量(x_i,y_i)の分布の範囲を測る測度であって、一変量の度数分布に於て分散が算術平均値を中心として、各単位の変量の分布の範囲を測定するのと同じ機能を、二変量の度数分布即ち相関分布に於て果すのである。これによって相関係数ρは次のように表される。

$$\rho=\frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y} \qquad (19)$$

共変量の簡便計算法

(18')式は次のように変形する事が出来る。

$$\mathrm{Cov}(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$$

$$=\frac{1}{N}\sum_{i=1}^{N}\{(x_i-\bar{x}_0)+(\bar{x}_0-\bar{x})\}\{(y_i-\bar{y}_0)+(\bar{y}_0-\bar{y})\}$$

但、$\bar{x}_0,\bar{y}_0$はそれぞれα集団性及びβ集団性に就ての度数分布に於ける仮平均

分子：$\sum^{N}\{\cdots\cdots\}$

$$=\sum^{N}\{(x_i-\bar{x}_0)(y_i-\bar{y}_0)+(\bar{x}_0-\bar{x})(y_i-\bar{y}_0)+(\bar{y}_0-\bar{y})(x_i-\bar{x}_0)+(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})\}$$

$$=\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y}_0)+(\bar{x}_0-\bar{x})\sum^{N}(y_i-\bar{y}_0)+(\bar{y}_0-\bar{y})\sum^{N}(x_i-\bar{x}_0)+\sum^{N}(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})$$

しかるに

第二項：$(\bar{x}_0-\bar{x})\sum^{N}(y_i-\bar{y}_0)=(\bar{x}_0-\bar{x})(N\bar{y}-N\bar{y}_0)$,

$=-(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})N$, $\because\ \bar{y}=\frac{1}{N}\sum^{N}y_i$

第三項：$(\bar{y}_0-\bar{y})\sum^{N}(x_i-\bar{x}_0)=(\bar{y}_0-\bar{y})(N\bar{x}-N\bar{x}_0)$

$=-(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})N$, $\because\ \bar{x}=\frac{1}{N}\sum^{N}x_i$

第四項：$\sum^{N}(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})=(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})\{\sum^{N}1\}$

$=(\bar{y}_0-\bar{y})(\bar{x}_0-\bar{x})N$

$\therefore\ \sum^{N}\{\cdots\}=\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y}_0)-2(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})N+(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})N$

$=\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y}_0)-(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})N$

故に

$$\mathrm{Cov}(x,y)=\frac{1}{N}\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y}_0)-(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y}) \tag{20}$$

この式によると共変量の計算は容易に行える。若し仮平均を用いない $\bar{x}_0=0,\ \bar{y}_0=0$ とすればよいのであるから

$$\mathrm{Cov}(y,y)=\frac{1}{N}\sum^{N}x_iy_i-\bar{x}\bar{y} \tag{20'}$$

となる。

相関係数の簡便計算法

相関係数は共変量と標準偏差の比であるため、それ等の簡便計算法によって容易に求める事が出来る。故に

$$\rho=\frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y}$$

$$=\frac{\frac{1}{N}\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y})-(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})}{\sqrt{\frac{1}{N}\sum^{N}(x_i-\bar{x}_0)^2-(\bar{x}_0-\bar{x}^2)}\sqrt{\frac{1}{N}\sum^{N}(y_i-\bar{y}_0)^2-(\bar{y}_0-\bar{y})^2}}$$

$$=\frac{\sum^{N}(x_i-\bar{x}_0)(y_i-\bar{y}_0)-N(\bar{x}_0-\bar{x})(\bar{y}_0-\bar{y})}{\sqrt{\sum^{N}(x_i-\bar{x}_0)^2-N(\bar{x}_0-\bar{x})^2}\sqrt{\sum^{N}(y_i-\bar{y}_0)^2-N(\bar{y}_0-\bar{y})^2}} \tag{21}$$

仮平均を用いない時には

$$\rho=\frac{\sum^{N}x_iy_i-N\bar{x}\bar{y}}{\sqrt{\sum^{N}x_i^2-N\bar{x}^2}\sqrt{\sum^{N}y_i^2-N\bar{y}^2}} \tag{21'}$$

(ii) 級内相関係数

概念的に対称な二つの個体がグループ（集落 cluster という）をなし、そのグループより構成されている統計集団に於て、グループ内の単位間の或る量的性質の相互関連の密接さを知らんとする場合、何れか一方の変量を x とし他方の変量を y として、(16') 式によって相関々係を計算する事が考えられるであろう。しかしながらその場合グループ内の単位は概念的に対称なものであるため、何れの単位の変量を x とし他方を y とするかが問題となるのである。例えば双生児の兄弟のグループより成る双生児の集団に於て、双生児の身長の類似の密接さの程度を測らんとする場合、双生児のうち何れの者の身長を x とし他方の者の身長を y とするかは、一義的に決定し得ない問題である。若し兄の身

長をx、弟の身長をyとするならば、相関係数は双生児の兄と弟との間の身長の相関々係を示す事となり、若し丈の高い方をx、低い方をyとすれば、相関々係は双生児の丈の高い者と低い者との間の身長の類似性を表すものとなるのであって、双生児の身長の類似性の測度とはいい得ないのである。そこでこの場合はグループ内の単位が概念的に対称な個体であるため、どれか一方をx、他方をyと固定し得ないのであるから、何れも交互にx並にyとして相関係数を計算すればよいであろう。

又仮令変量の組がxとyとに固定し得る場合であっても、統計集団が三（又は三以上）の量的集団性に就て構造分析された場合、この三変量間の相互関連の密接さの測定は最早や(16′)式では不可能であり、それ等三の変量を相互にx並にyたらしめて相関係数を計算しなければならないのである。

今統計集団を構成するk個のグループの観測値の組を

$$(x_{11}, x_{12}), (x_{21}, x_{22}), \cdots, (x_{i1}, x_{i2}), \cdots, (x_{k1}, x_{k2}) \quad (22)$$

（但x_{ij}の添数iはグループの番号である）とすると、各組に於て先ずx_{i1}をx、x_{i2}をyとし、次に逆にx_{i2}をx、x_{i1}をyとするのであるから(22)のk個の観測値の組は

$$\begin{gathered}(x_{11}, x_{12}), (x_{21}, x_{22}), \cdots, (x_{i1}, x_{i2}), \cdots, (x_{k1}, x_{k2}) \\ (x_{12}, x_{11}), (x_{22}, x_{21}), \cdots, (x_{i2}, x_{i1}), \cdots, (x_{k2}, x_{k1})\end{gathered} \quad (23)$$

の$2k$個の組となる訳である。（各組の左側の値がx、右側の値がyである。）そして(23)の観測値の組の相関係数を(16′)式によって計算するのであるが、それには先ずx数列及びy数列の算術平均値、標準偏差を求めねばならない。(23)より明らかな如くx数列は(22)のk個の観測値の組のすべての値、従って$2k$個の単位（これが統計集団を構成する単位の総数である。）の有する変量値のすべてを含んで居り、この事はy数列に就ても同じであるため、x数列とy数列は算術平均値、標準偏差を同じくし、それは同時に統計集団それ自体の平均値$\bar{x}$、標準偏差σでもある。従って

$$\bar{x} = \bar{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{2} x_{ij}}{2k} = \bar{x}$$

$$\sigma_x=\sigma_y=\sqrt{\frac{1}{2k}\sum_{i=1}^{k}\sum_{j=1}^{2}(x_{ij}-\bar{x})^2}=\sigma$$

そして(16')式は(23)の各組の値の$\bar{x}$よりの偏差を標準化し、その積の合計を平均したものであるので、此の場合は次の様になる。

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}=\frac{\frac{1}{2k}\sum_{i=1}^{k}\{(x_{i1}-\bar{x})(x_{i2}-\bar{x})+(x_{i2}-\bar{x})(x_{i1}-\bar{x})\}}{\sigma\cdot\sigma}=\frac{\frac{1}{2k}\sum_{i=1}^{k}\sum_{j\neq j'}^{2}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})}{\sigma^2}$$

茲に$\sum_{j\neq j'}^{2}$は添数jの異るx_{ij}の組の偏差の積を二つ加える事を意味するのである。このような相関係数を級内相関係数 intraclass correlation coefficient といゝρ'で表す。故に

$$\rho'=\frac{\frac{1}{2k}\sum_{i=1}^{k}\sum_{j\neq j'}^{2}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})}{\sigma^2}\qquad(24)$$

これに対して前項の一般的な相関係数を級間相関係数 interclass correlation coefficient というのである。

以上はグループ内の単位が2個の場合であるが、これが一般的なl個の場合には級内相関係数の計算式は次の如くである。各グループの観測値の組を

$$(x_{11},x_{12},\cdots,x_{1l}),(x_{21},x_{22},\cdots x_{2l}),\cdots,(x_{i1},x_{i2},\cdots,x_{il}),\cdots,(x_{k1},x_{k2},\cdots,x_{kl})\qquad(25)$$

とする。先ず第一のグループに於て一のx_{1j}をx他の任意の$x_{1j'}$をyとして組合せ、且l個のすべてをx（並にy）たらしめる場合の数は、l個の中より重複する事なく2個取出す場合の数として求められる。それはそれは$l(l-1)$であるため、第一のグループは$l(l-1)$組の$(x_{1j},x_{1j'})$の組に分れるのである。この事はk個のすべてのグループに就て同じであるため、結局(25)は$kl(l-1)$組の$(x_{ij},x_{ij'})$の組に分れるのであり、それに就て相関係数を計算するのである。x数列、y数列の算術平均値、標準偏差は、前と同様に統計集団全体（この場合はkl個の単位より成る。）の平均値$\bar{x}$、標準偏差σであるから

$$\bar{x}=\bar{y}=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}x_{ij}=\bar{x}$$

$$\sigma_x=\sigma_y=\sqrt{\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x})^2}=\sigma$$

故に級内相関係数 ρ' は

$$\rho'=\frac{\frac{1}{kl(l-1)}\sum_{i=1}^{k}\sum_{j\neq j'}^{l}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})}{\sigma^2} \qquad (26)$$

茲に $\sum_{j\neq j'}^{l}$ は添数 j の異る x_{ij} の組の偏差の積をすべて（その数は第1グループに就て見たように $l(l-1)$ 個ある。）加える事を意味する。

この級内相関係数の性質を知るためには先ず次の事を知らねばならぬ。統計集団全体の分散は次のように変形される。

$$\sigma^2=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x})^2$$

$$=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}\{(x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{x})\}^2$$

但し $\bar{x}_i=\frac{1}{l}\sum_{j=1}^{l}x_{ij}\cdots$ i-group の平均値

$$=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)^2+\frac{2}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x})+\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(\bar{x}_i-\bar{x})^2$$

しかるに

第二項: $\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x})$

$$=\sum_{i=1}^{k}(\bar{x}_i-\bar{x})\underbrace{\left\{\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)\right\}}_{0}\overset{(1)}{=}0$$

$\because$ 〔算術平均値よりの偏差の総和なるため〕

第三項: $\sum_{i=1}^{k}\sum_{j=1}^{l}(\bar{x}_i-\bar{x})^2=\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2\left\{\sum_{j=1}^{l}1\right\}$

$$=\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2 l$$

$$\therefore\ \sigma^2=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)^2+\frac{1}{kl}\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2 l$$

ここで $\sigma_w^2=\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)^2$, $\sigma_b^2=\frac{1}{kl}\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2 l$

とおけば

$$\sigma^2=\sigma_w^2+\sigma_b^2 \qquad (27)$$

(1) $\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x})$ に於て $(\bar{x}_i-\bar{x})$ は添数 j を含んでいないため、j の計算に於ては常数となり、j に就ての和 $\sum_{j=1}^{l}$ の外部に約し出せる。

ここに σ_w^2 は集落内分散 variance within clusters といい、各集落（グループ）内に於ける変量の集落平均値を中心とする分布の範囲の測度であり、σ_b^2 は集落間分散 variance between clusters といい、各集落平均値の統計集団平均値を中心とする分布の範囲を測定するものである。かくて統計集団が統計単位のグループ即

ち集落より成る時は、統計集団の分散は集落内分散と集落間分散より合成されているのである。従って

(1) 集落内部の単位が完全に同質であり、それ等の有する変量値が同じ（$x_{i1}=x_{i2}=\cdots=x_{il}$）であるならば $\sigma_w^2=0$ 故に $\sigma^2=\sigma_b^2$ であり

$(c\ c\ c)$
$(b\ b\ b)$
$(a\ a\ a)$

(2) 集落平均値が相互に相等しい（$\bar{x}_1=\bar{x}_2=\cdots=\bar{x}_k$）即ち各集落が同質であるならば（その時は各集落内部は可能な限り異質的な単位より成っているであろう。[1]）$\sigma_b^2=0$ 故に $\sigma^2=\sigma_w^2$ となる。

$(a\ b\ c)$
$(a\ b\ c)$
$(a\ b\ c)$

(3) そして集落の構成がこの両極端の中間にある場合、集落内部が同質的（故に σ_w^2 は小）である程 σ^2 は σ_b^2 に依存し、集落間が同質的（故に σ_b^2 は小、この時集落内部は異質的）なる程 σ^2 は σ_w^2 によるのである。

(1) 統計集団に含まれる単位は可成り異質的であるため、それを集落に分けた場合各集落が相互に同質的となるためには、集落内部の単位は出来る丈異質的であらねばならないであろう。

さて級内相関係数 ρ' の算式は次のように変形する事が出来る。

$$\rho'=\frac{1}{\sigma^2}\left\{\sigma_b^2-\frac{1}{l-1}\sigma_w^2\right\} \qquad (28)$$

即ち(26)式より

$$\rho'=\frac{1}{\sigma^2}\left\{\frac{\sum_{i=1}^{k}\sum_{j\neq j'}^{l}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})}{kl(l-1)}\right\}$$

$$\{\cdots\}=\frac{1}{kl(l-1)}\sum_{i=1}^{k}\left\{\sum_{j\neq j'}^{l}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})\right\}$$

しかるに $\left\{\sum_{j=1}^{l}(x_{ij}-\bar{x})\right\}^2$

$$=\sum_{j=1}^{l}(x_{ij}-\bar{x})^2+\sum_{j\neq j'}^{l}(x_{ij}-\bar{x})(x_{ij'}-\bar{x}) \quad [1])$$

$$\therefore \quad =\frac{1}{kl(l-1)}\sum_{i=1}^{k}\left[\left\{\sum_{j=1}^{l}(x_{ij}-\bar{x})\right\}^2-\sum_{j=1}^{l}(x_{ij}-\bar{x})^2\right]$$

$$=\frac{1}{kl(l-1)}\sum_{i=1}^{k}\left[\sum_{j=1}^{l}\{(x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{x})\}\right]^2-\frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x})^2}{kl(l-1)}$$

第一項大括弧内：$\sum_{j=1}^{l}\{(x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{x})\}$

$$=\underbrace{\sum_{j=1}^{l}(x_{ij}-\bar{x}_i)}_{0}+(\bar{x}_i-\bar{x})l$$

第二項：$\frac{1}{l-1}\left\{\frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_{ij}-\bar{x})^2\right\}=\frac{1}{l-1}\sigma^2$

$$\therefore\ \{\cdots\}=\frac{\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2 l^2}{kl(l-1)}-\frac{\sigma^2}{l-1}$$

$$=\frac{l\sigma_b^2}{l-1}-\frac{\sigma^2}{l-1},\quad \because\ \sigma_b^2=\frac{1}{kl}\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2 l$$

$$=\sigma_b^2-\frac{\sigma^2-\sigma_b^2}{l-1}=\sigma_b^2-\frac{\sigma_w^2}{l-1},\quad \because\ \sigma^2=\sigma_w^2+\sigma_b^2$$

故に　$\rho'=\frac{1}{\sigma^2}\left\{\sigma_b^2-\frac{1}{l-1}\sigma_w^2\right\}$

(1)　例えば

$$\left\{\sum_{j=1}^{3}a_j\right\}^2=(a_1+a_2+a_3)^2$$

$$=a_1^2+a_2^2+a_3^2+a_1a_2+a_1a_3+a_2a_1+a_2a_3+a_3a_1+a_3a_2$$

$$=\sum_{j=1}^{3}a_j^2+\sum_{j\neq j'}^{3}a_ja_{j'}$$

かくて級内相関係数は三種の分散 σ^2 σ_w^2 σ_b^2 によって表されるのである。そして $\sigma_w^2=0$ 故に $\sigma^2=\sigma_b^2$ 即ち集落内同質の時は $\rho'=1$ であり、$\sigma_b^2=0$ 故に $\sigma^2=\sigma_w^2$ 即ち集落間同質の時は $\rho'=-\frac{1}{l-1}$ である。一般には統計集団に於ける集落の構造は以上二の両限界の中間であるため、級内相関係数は

$$-\frac{1}{l-1}\leqq\rho'\leqq1\qquad(29)^{1)}$$

なる値をとるのである。

(1)　これは次のようにして証明される。(29)式より

$$\rho'=\frac{1}{\sigma^2}\left\{\sigma_b^2-\frac{1}{l-1}\sigma_w^2\right\}$$

$$=\frac{1}{\sigma^2}\left\{\sigma^2-\sigma_w^2-\frac{1}{l-1}\sigma_w^2\right\},\quad \because\ \sigma_b^2=\sigma^2-\sigma_w^2$$

$$=\frac{1}{\sigma^2}\left\{\sigma^2-\frac{l}{l-1}\sigma_w^2\right\}=1-\frac{l}{l-1}\left(\frac{\sigma_w}{\sigma}\right)^2\leqq1$$

又 $\rho'=\frac{1}{\sigma^2}\left\{\sigma_b^2-\frac{1}{l-1}(\sigma^2-\sigma_b^2)\right\},\quad \because\ \sigma_w=\sigma^2-\sigma_b^2$

$$=\frac{1}{\sigma^2}\left\{\frac{l}{l-1}\sigma_b^2-\frac{1}{l-1}\sigma^2\right\}$$

$$=-\frac{1}{l-1}+\frac{l}{l-1}\left(\frac{\sigma_b}{\sigma}\right)^2\geqq-\frac{1}{l-1}$$

$$\therefore\ -\frac{1}{l-1}\leqq\rho'\leqq1$$

註　集落が只一の場合換言すれば統計集団を集落に分けない場合（故にこの時は l は統計集団単位総数 N に等しい）は $\rho'=-\frac{1}{N-1}$ である。なんとなれば、集落が一であるため(29)式に於て $k=1$, $l=N$ とすればよく、従ってそしてこの時の統計集団の平均値、標準偏差は

$$\rho'=\frac{1}{\sigma^2}\left\{\frac{\sum_{j\neq j'}^{N}(x_j-\bar{x})(x_{j'}-\bar{x})}{N(N-1)}\right\}$$

$$\bar{x}=\frac{1}{N}\sum_{j=1}^{N}x_j,\qquad \sigma^2=\frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})^2$$

である。さて

$$\rho' = \frac{1}{\sigma^2}\left[\frac{1}{N(N-1)}\sum_{j=1}^{N}(x_j-\bar{x})\left\{\sum_{j'\neq j}^{N}(x_{j'}-\bar{x})\right\}\right]$$

$$= \frac{1}{\sigma^2}\left[\frac{1}{N(N-1)}\sum_{j=1}^{N}(x_j-\bar{x})\left\{\underbrace{\sum_{j'=1}^{N}(x_{j'}-\bar{x})}_{=0\ \because\ \text{算術平均値よりの偏差の総計なるため}} - (x_j-\bar{x})^{2)}\right\}\right]$$

$$= \frac{-1}{\sigma^2}\left\{\frac{1}{N(N-1)}\sum_{j=1}^{N}(x_j-\bar{x})^2\right\}$$

$$= \frac{-\sigma^2}{\sigma^2(N-1)} = -\frac{1}{N-1}$$

(1) $(x_j-\bar{x})(x_{j'}-\bar{x})$ は N 個の偏差 $(x_j-\bar{x})$, $(j=1,2,\cdots,N)$ 中の異なる二個の積であつて、それは各 j に就て $(N-1)$ 個ある。故にその $(N-1)$ 個の積の和を $(x_j-\bar{x})$ でくゝると $(x_j-\bar{x})\left\{\sum_{j'\neq j}^{N}(x_{j'}-\bar{x})\right\}$ と書ける。茲に $\sum_{j'\neq j}^{N}(x_{j'}-\bar{x})$ は $(x_j-\bar{x})$ 以外の $N-1$ 個の偏差の合計を意味する。そしてこの事はすべての j に就ていえるため、偏差の積の總和は

$$\sum_{j\neq j'}^{N}(x_j-\bar{x})(x_{j'}-\bar{x}) = \sum_{j=1}^{N}(x_j-\bar{x})\left\{\sum_{j'\neq j}^{N}(x_{j'}-\bar{x})\right\}$$

と書けるのである。

(2) $(x_j-\bar{x})\left\{\sum_{j'\neq j}^{N}(x_{j'}-\bar{x})\right\}$ の $\{\cdots\}$ は、N 個の偏差中の一偏差 $(x_j-\bar{x})$ を除いた $(N-1)$ 個の偏差の合計であるため、それに $(x_j-\bar{x})$ を加えると結局全偏差の合計となる。

$$\sum_{j'\neq j}^{N}(x_{j'}-\bar{x}) + (x_j-\bar{x}) = \sum_{j'=1}^{N}(x_{j'}-\bar{x})$$

$$\therefore\ \sum_{j'\neq j}^{N}(x_{j'}-\bar{x}) = \sum_{j'=1}^{N}(x_{j'}-\bar{x}) - (x_j-\bar{x})$$

この級内相關係數の概念は任意標本理論に於ては重要な役割を演ずるのである。しかし、以上の証明の過程はやゝ難解であるので、結論丈了承すればそれで充分である。