◇◇◇◇◇　データ解析のためにもちいる統計　◇◇◇◇◇

/統計学テキスト/記述統計始め/descriptiveA1-a.tex

目次

1　はじめに

統計学は、バラツキのあるデータに対して、経験的知識や専門的認識をもとに、数学の手法を用いて、数値データの特質や規則性の有無を発見あるいは検証します。統計的手法は、統計手法の開発、実験観測の計画、データの要約や解釈を行う上での理論的な根拠を提供する学問であり、幅広い分野で応用されています。

英語で統計または統計学を statistics といいますが、語源はラテン語で「状態」を意味する statisticum であり、この言葉がイタリア語で「国家」を意味するようになり、国家の人力、財力等といった国勢データを比較検討する学問を意味するようになったといわれています。。現在では、経済学、自然科学、社会科学、医学(疫学、EBM)、薬学、心理学、言語学など広い分野で必須の学問となっていることは論をまたない。また統計学は哲学の一分科である科学哲学においても重要なひとつのトピックスになっている。理由は統計学が科学的な研究においての方法論として基礎的な部分を構成していながら、確率という極めて捉えがたい概念を基盤としていることもあり、不確実性の知覚についてその意味やあり方が帰納の正当性の問題などと絡めて議論されている。

記述統計とは、収集したデータから要約した数値、すなわち統計量とよばれる平均、分散などの数値を計算して、データの状況や規則性を明らかにすることで、データの示す傾向や性質を知ることです。とくにデータについては、これに関する経験や知識が分析を行うためには重要なかぎとなる。新たな発見、知覚認識に結びついていく。

推測統計という分野では、抽出されたデータからその根源となっている諸性質を確率論的に推測する分野で

あり、経験から発見される知覚ではなく、数理的な仮説にもとづく検証や推測を目的とする。

2　統計調査

新聞やテレビにも多くの統計調査が発表されています。いくつかの代表的な統計調査を挙げてみる。政治施策の決定のためには欠かせない景気動向を調べたり、日常生活を楽しむためにも意外な結果もあります。また研究の実験成果を確実に高めるためにも統計データの分析は必要な道具となっています。

(i) 政府による官庁統計: 人口動態 (出生率、死亡率、市町村別人口動態)、消費者物価指数、GDP 成長率
(ii) マスコミによる調査: 世論調査、政党支持率、番組視聴率調査
(iii) 民間会社による市場調査: POS データ、商品動向、需要予測、経済予測

これらの調査での手順は

STEP 1: **調査の企画**　　目的と対象、経費、実施手順、調査項目、設計
STEP 2: **調査の実施**　　調査員の手順説明、抽出方法（２段階層別抽出法など）
STEP 3: **調査結果の処理**　集計整理、コンピュータによるデータ処理、出力結果、結果の分析
STEP 4: **公表**　　調査結果の発表、印刷、製本、出版など

目的と企画：	アンケート調査の目的、その必要性。インターネットなどの代用ではなく、意義の検討。最終的に解明したい問題点を認識することと、その期待すべき成果。 調査対象者、集団の検討、その方法、調査期間。研究活動や調査の解明のために、どういう内容をどのような人たちや集団に対して、どのように調べるか？
調査の準備：	調査票の作成、質問項目の決定、質問数、質問の文章（どのような質問がよいか）に瑕疵（かし）はないか、データの集計した後は、どう手分けをして処理をおこなうか？人員の確保。調査に必要な費用の算出とその金額の確保。チェックシートの作成。
実施：	調査票の配布方法とその回収、調査の状況把握など実施の状況を記録。郵送の場合には収受の記録。督促の連絡をどの時期にどうすべきか？
解析：	データ入力、統計手法の活用、知識の習得。全体的な問題点の把握と認識。単純集計、クロス集計、回帰分析、要因分析。
報告と公表：	報告書の作成、これまでの記録を整理。対象集団と調査の方法と結果。目的に対しての方法とその解析結果。適用した統計手法の妥当性。

3　統計データの種類、分類

調査対象に割り振った変数、その測定により得られたデータが表す基準を尺度水準（しゃくどすいじゅん）という。これらが表現する情報の性質に基づき数学・統計学的に分類する基準である。データ（あるいは変数、測定）の尺度あるいは単位の構造から、ふつう次のような種類（水準）に分類される。この尺度水準によって、統計に用いるべき基本の統計量や統計検定法が異なることに注意する。

3.1 質的データ

質的データ は、カテゴリデータともよばれ、観測データとしては選択あるいは観測した結果が項目（カテゴリー）になっているデータである。

- **名義尺度：** 単なる番号であり、順番の意味をもたない。電話番号、背番号など。

この水準では数字を単なる名前として対象に割り振る。2 つの対象に同じ数字がついていればそれらは同じカテゴリに属する。変数値間の比較は等しいか異なるかでしか行えない。順序もないし加減などの演算もできない。例としては電話番号、背番号、バスの系統番号など。中心的傾向の指標として使えるのは最頻値のみである。統計的バラツキは変動比や情報エントロピーで評価できるが、標準偏差などの概念はありえない。名義尺度でのみ測定されるデータはカテゴリデータとも呼ばれる。

＊ なおカテゴリデータを、ある性質が「あるかないか」という表現に直し、さらにこれを「1 か 0 か」で表現したものをダミー変数という。ダミー変数またはそれから算出されるスコア（点数）を、順序尺度以上の水準に準じて扱う方法もよく用いられる。

- **順序尺度：** 順序が意味を持つ番号。階級や階層など。

この水準では対象に割り振られた数字は測定する性質の順序を表す。数字は等しいかどうかに加え、順序（大きいか小さいか）による比較ができる。しかし加減などの演算には意味がない。物理学的な例にはモース硬度がある。その他の例にはレースの着順などがあるが、これでは到着時間の差は記録できない。心理学や社会科学の測定のほとんどは順序尺度で行われる。例えば社会的態度 (保守的か進歩的かなど) や階級は順序水準で測定されるものである。また客の嗜好（アイスクリームのバニラ味とチョコレート味とどちらが好きか）のデータもこれで表現できる。順序尺度の中心的傾向は最頻値 (モード) や中央値 (メジアン) で表されるが、中央値の方が多くの情報を与える。順序尺度で測定されるデータは順序（または順位）データと呼ばれる。

＊ 以上の名義尺度および順序尺度で表されるデータを合わせて質的データともいう。また各カテゴリに属す対象の個数という形のデータにまとめると数量データと呼ばれ、これは分割表で表示できる。これらに対して用いられる統計検定法はノンパラメトリックなものに限られる。

3.2 量的データ

実験結果や観測値など、調査で得られた結果が数値データであるものをさす。尺度構造として、間隔尺度と比率尺度に分ける。また変数値が実数と整数のそれぞれの場合について、実数型と離散型データとよばれる。

- **間隔尺度：** 数直線上に大きさの順にならべてプロットしたとき、順序に加えて、数値間の間隔にも意味がある（単位がある）が、ゼロには絶対的な意味はない。摂氏・華氏温度、知能指数など。

対象に割り振られる数字は順序水準の性質を全て満たし、さらに差が等しいということは間隔が等しいということを意味する。つまり測定値のペアの間の差を比較しても意味がある。加減の演算にも意味があるが、尺度上のゼロ点は任意で負の値も使える。例にはカレンダーの日付がある。値の間の比には意味がなく、直接の乗除の演算は行えない。とはいえ差の比には意味がある。中心傾向は最頻値、中央値あるいは算術平均で表され、算術平均が最も多くの情報を与える。間隔尺度で測定されるデータは間隔データと呼ばれる。摂氏または華氏で測る温度も間隔尺度である。社会・人文科学分野で用いられる間隔尺度の例では知能指数（IQ）がある。

• **比率尺度**：ゼロを基準とする絶対的尺度で、間隔だけでなく比率にも意味がある場合につけられたもの。絶対温度、金額など。

対象に割り振られた数字は間隔尺度の性質を全て満たし、さらにその中のペア（２つの比較）の比（割合）にも、乗除の演算にも意味がある。比率水準のゼロ点は絶対的である。ほとんどの物理学的量、つまり質量、長さやエネルギーは比率水準である。また温度も絶対温度で測れば比率尺度である。比率尺度で測定される変数の中心的傾向は最頻値、中央値、算術平均あるいは幾何平均で表されるが、間隔尺度と同じく算術平均が最も多くの情報を与える。比率尺度で測定されるデータは比率データと呼ばれる。比率尺度で表される社会的変数には年齢、ある場所での居住期間、収入などといったものがある。

＊正しい意味で単位を有するのは間隔尺度と比率尺度のみであり、従ってこれらは真の尺度とも呼ばれる。これらのデータを合わせて量的データ（質的データに対して)、数値データ（数量データに対して）ともいう。

スタンレー・スティーヴンズ（Stanley Smith Stevens）により 1946 年の論文「測定尺度の理論について」”On the theory of scales of measurement”で提案された分類がよく用いられる。変数に対して可能な数字の演算は、変数を測定した尺度水準に依存し、その結果、特に統計学で用いるべき要約統計量および検定法も変数の尺度水準に依存する。スティーヴンズは低い方から順に以上の 4 つの尺度水準を提案しており、高い水準はより低い水準の性質を含む形になっている。また高い水準でのデータを低い水準に変換して扱うことができる。

参考：Wikipedia, Excel help

4　度数分布表とヒストグラム

度数分布（どすうぶんぷ、Frequency Distribution）とは、統計において標本として得られたある変量の値のリストである。一般に量の大小の順で並べ、各数値が現われた個数を表示する表（度数分布表）で示されます。

データの整理をするために、表計算ソフトを用いると便利です。最近は Windows やＭａｃＯＳ、さらにＬｉｎｕＸなどの上で動くソフトがあります。代表的なものがエクセル (MicroSoft) ですが、あるいはオープンオフィスのカルク（OpenOffice.org）（オープンオフィス・ドット・オルグ、あるいは オープンオフィス・オルグ）は関数の命令もほぼ同じで、フリーソフトになっています。検索ソフトで調べてみて下さい。

例 4.1

ある小学生のクラスで測ったデータ、体重や身長の測定値は量的データとして得られます。そのままではなく、データの加工して表示したりすることが多くあります。

たとえば体格指数にはローレル指数＝体重☒÷身長 $(cm)^3\times 10^7$ とBMI ＝体重 (☒) ÷身長 $(m)^2$ などがよく知られています。

BMI	18.5 未満	やせ
	18.5〜25 未満	標準
	25〜30 未満	肥満
	30 以上	高度肥満

標準体重による肥満度は次の式で計算します。「肥満度（%）＝（実測体重－標準体重）÷標準体重× 100」ローレル指数の場合，身長によって肥満の判定基準が変わります。したがって個人の状態の推移を学年をおって記録するには適切とはいえません。BMI は小学生の場合，その判定と実際の児童の状態とがあまりにずれてしまうのだそうです。成長期の基準化にはうまく行かない問題があります。BMI で判定したところ、大半が「やせ」となってしまいまうことが多いです。青年期以降だと実態に合うようです。

図 1: 小学生に対する身長体重の測定値

例 4.2

例えば、100 人がある主張に同意するかの否かを 5 段階の強弱のいわゆるリッカート尺度で回答したとします。これは項目によって分類され、質的データでの順序尺度の単位構造をもちます。このとき、1 は強く同意することを示し、5 は全く同意しないことを示す。その回答群を度数分布で表すと次のようになります:

階級	同意の度合	集計マーク	回答数
1	強く同意する	正正丅	12
2	ある程度同意する	正正正正正一	26
3	どちらとも言えない	正正正	15
4	ある程度同意できない	正丅	7
5	全く同意できない	正	5

この単純な表には 2 つの弱点がある。変量が連続的な値をとりうる場合と多数の項目に分かれ、範囲が広い場合である。このような度数分布表の作成は難しくなる。また連続的な変量に対しては、観測データをいくつかの階級に分類し、その中心となる階級値ごとに度数を修正する。つぎに述べるヒストグラム (Histogram)（柱状グラフとのよばれる）と関連されて表現するほうがよい。

離散型データを表すための棒グラフではなく、連続的な数値を丸めて表現するという意味をもつから、階級の幅が意味をもっていることに注意。また数値で評価された試験結果の得点を 5 段階評価（秀、優、良、可、不可）で分類する数値データの離散化も使われる。

階級によって区分けされた観測データをグラフによって表現します。注意すべきことはこの階級の幅に意味

図 2　集計マーク (tally mark) のいろいろ

をもつことです。それに対して棒グラフで表す場合ではデータが離散的な場合やカテゴリーで与えられている場合であって、階級に区分けするときは、項目の数（x 軸に対応する）が多数であってある程度まとめたほうがよい場合です。

平均と中央値が異なる場合、度数分布に歪み（ひずみ、skewness）があるといいます。正規分布は平均を中心として対称な形状をして歪みがないといいますが、これとは異なり、L 字型と逆 L 字型の場合が相当します。L 字型とはデータの値が（平均値の右側）大きいほうに低く長くあることから、3 次モーメントは正の値となり、この計算から歪度は正値となります。逆に大きい値により多くのデータがかたまり、この付近に度数も大きければ、平均もこの大きいほうになり、平均の周りの 3 次モーメント値は（平均値の左側）小さい値が多数になることから、逆 L 字型の分布では歪度が負の値になります。

度数分布の尖度（せんど、kurtosis, excess）とは、平均値への集中の度合であり、ヒストグラムで表した場合のグラフの尖り（とがり）具合です。正規分布以上に尖っている場合を「急尖的; leptokurtic」と称し、逆の場合を「緩尖的; platykurtic」とよばれます。データのばらつき具合を正規分布と比較するもので、数値 3 が正規分布であり、平べったいならば、平均の周りの 4 次モーメントを計算することから、値が大きくなり、負の値になる場合は尖っていて、平均の周りにデータが固まっていてモーメントの値が小さくなるからです。

ヒストグラム（度数分布図、柱状グラフ、Histogram）とは、縦軸に度数、横軸に階級をとった統計グラフの一種で、データの分布状況を視覚的に認識するために最も基本的でよく用いられます。

累積度数分布表

級中央値に対する度数の対応表は度数分布表ですが、度数を累積していったものは、単調に増えていき、最後の値はデータの総数になります。これは確率分布では分布関数に相当します。中央値、四分位数、十分位数、百分位数などの計算はこの累積度数分布表から、それぞれ、５０％づつ 2 つに分ける、２５％づつ 4 つに分ける、１０％づつ 10 個に分ける、１％ごとの 100 個のデータに大きさの順にしたものとなります。

相対度数分布表

上で述べた累積相対度数に対して、データの総数で割り、百分率、パーセント値に直したもの。このように基準化することで、データ数が多数であっても表示され、2 つ以上の比較も可能となります。

統計図表

いわゆる統計グラフは、統計図表（とうけいずひょう）ともよばれ、複数の統計データの整理、視覚化、分析、解析等に用いられるグラフおよび表の総称を意味します。ここで、グラフとは「図形を用いて視覚的に、複数の数量・標本資料の関係などを特徴付けた物」のことを指します。この意味においてのグラフはしばし「統計グラフ」と呼ばれます。統計図表は、統計データの整理、分析、検定などの過程で用いられる。統計図表を駆使することで、「調査活動によって得られた数量（統計データ）の特徴」（増減の傾向の型，集団の構成など）や、統計データ同士の関係（相関関係など）を視覚的に理解することが出来ます。

統計グラフの種類

統計グラフには、様々な種類がありますが、以下に典型的な統計グラフを示します。

棒グラフ　棒グラフは、資料を質的に（意味的に複数の項目に）分類したときに、各項目間の大きさを比較するために用いる。項目を横軸、各項目の大きさを縦軸に表現する（横軸、縦軸は逆でも良い）。棒で表すことで、各項目の大きさや、大きい値（小さい値）を持つ項目、各項目間の関係などが把握しやすくなる。

柱状グラフ（ヒストグラム）　柱状グラフ（ヒストグラム）は、棒グラフの一種で、資料を量的に（大きさを複数の階級に区分し、各要素がどの階級に属するかという指標で）分類した時に、各階級の散らばりの様子を見るために用いる。柱状で表すことで、集団の偏りや各階級間の散らばりの様子が把握しやすくなる。品質管理などにおいて、度数分布表から度数分布を図示するときによく用いられる。度数が増えるにしたがって、グラフの形状は柱状から曲線へと近づいてゆく。この曲線を度数分布曲線という。

箱ひげ図（ボックスチャート）　データの集まり、データセットを表現するとき、中心的な位置とバラツキの範囲を同時に表現する方法がある。位置を示す平均が点のプロットで示され、データのバラツキ具合を表す標準偏差あるいは四分位数（Q_1, Q_2, Q_3）を一つにまとめて箱の長さで第 1 、第 3 四分位数 Q_1, Q_3 を示し、箱から突き出した棒でデータの最大と最小までの範囲を示す。

円グラフ　円グラフは、資料を特定の項目に分類した時、その一項目での割合を比較する時によく用いられる。円で全体を表すことで、ある項目内・分野内での割合の大小が直感的に把握しやすく、プレゼンテーションなどでよく利用される。又、円グラフでは、全体の数値を 360 として表現することも少なくない。他方で、厳密な比較には向かないため、専門分野ではむしろ使用されない。

折れ線グラフ　時刻経過に伴い、データ変化の状況を表すためには、時刻に対応した数値を直線で結んでいく。それぞれのデータセットに対応した直線の色や頂点のマークをつけて区別すれば、いくつかのデータセットを同時に表現することができる。それぞれの系列グラフ変化を比較するためにもよく用いられる。

絵グラフ　視覚的な強調や興味あるアピールを行うために様々な表現方法で表示する。一般に表示を工夫したもの、分類できないようなものなどが含まれる。

レーダー・チャート　円形図形をもちいた各項目の数値を中心から半径への距離に基準化して表す。多変量データをまとめて表せる。たとえば、各人のテストでの科目ごとの得点結果をまとめて表現したり、

グラフ・ウィザートをもちいて、度数分布表のつくり方を説明します。

1. 得られたデータの配列から、最大 (max) と最小 (min) を求め、まず範囲（range）を計算します。
2. 階級 (class) の数をおおよそ 10 から 20 ぐらいになるよう設定します。もしうまく行かない場合は数を代えてみます。Starjes の公式 $k = 1 + \log_2 n = 1 + 3.3 \log_1 0n$ という目安もあります。
3. 階級幅 (class width) が範囲を階級の数で割ること R/k で計算できます。階級の度数によってはいくつかの階級を合併してまとめたほうがよいこともあります。
4. これから階級、階級値、度数、累積度数、相対度数、累積相対度数をまとめて表計算ソフトで計算します。

図 3　箱ひげ図 (Boxchart)

	A 列	B 列	C 列	D 列	E 列	F 列	G 列
1 行	階級 (下限)	階級 (上限)	階級値	度数	累積度数	相対度数	累積相対度数
2 行	$a_0 \sim$	a_1	x_1	f_1	$F_1 = f_1$	$p_1 = f_1/n$	$P_1 = p_1$
3 行	$a_1 \sim$	a_2	x_2	f_2	$F_2 = f_1 + f_2$	$p_2 = f_2/n$	$P_2 = p_1 + p_2$
4 行	$a_2 \sim$	a_3	x_3	f_3	$F_3 = f_1 + f_2 + f_3$	$p_3 = f_3/n$	$P_3 = p_1 + p_2 + p_3$
5 行	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
6 行	$a_{k-1} \sim$	a_k	x_k	f_k	$F_k = n$	$p_k = f_k/n$	$P_k = 1$
7 行	計			n		1	

$x_i = (a_{i-1} + a_i)/2, (i = 1, 2, \cdots, k)$ $F_k = f_1 + f_2 + \cdots + f_k = n$, $P_k = p_1 + p_2 + \cdots + p_k = 1$

この表の作成には数式メニューに組み込み関数をいれていきます。=max(データ範囲)、=min(データ範囲) から範囲をあらかじめ計算しておきます。階級の幅（上限と下限）や個数を決めます。度数の枠にいれる命令は、セル D2 からセル D6 をアクティブに選んでから、=frequency, データ配列、区間配列 (セル A2 からセル A6)、入力は数式配列ですから、CTRL+SHIFT+ENTER とします。階級の入力は「編集」→「連続データの作成」、また累積度数はセル E2 に「=D2」さらにセル E3 に「=D2+D3」として、E3 を E4 から E6 までコピーペーストをしていきます。相対参照をしながら自動計算してくれます。相対度数はセル F2 には「=D2/D$7」といれて、分母には絶対参照するセルの値、すなわち合計で割れば求まります。この度数分布表から、グラフ・ウィザードで棒グラフを選び、ヒストグラムをつくります。系列で x 軸に表すラベルには階級値を指定し、とくにオプション・コマンドで棒グラフの間隔幅をゼロに指定します。あるいはこの frequency 命令ではなく、分析ツールのヒストグラムでも作ることができます。

統計グラフにはどのような選択の目安を下記に示す。

1. 2 種類の系列（2 変量データの全体を俯瞰する）からなるデータの相関⇒散布図
2. 1 種類の系列（1 変量データで時間とともに変化する）からなるデータの時間的推移（時間との相関)、2 つ以上のグラフ間における比較⇒折れ線グラフ
3. 値の大きさの比較、横軸には項目カテゴリー⇒棒グラフ
4. 集団における内訳や構成比を見る⇒円グラフ

また表計算ソフトでヒストグラムを描くには棒グラフのチャートをつかいますが、横幅に意味があるので、階級値の幅をゼロにしなければいけません。そのためには　データ範囲を選んでから、「挿入→グラフ→縦棒→項目軸ラベルの設定」までは同じですが、グラフを描いてから、マウスのアイコンをグラフの棒にもってい

図 4　折れ線グラフとレーダー・チャート

図 5　ヒストグラム

き、”右クリック”で、「データ系列の書式設定→オプション→棒の間隔」でこの数値をゼロにすれば、間隔が詰まります。

5　データの代表値

データの集まりから、この分布を特徴づける基本統計量として平均 (average,mean)、標準誤差 (SD,standard deviation)、中央値（メジアン,median）、最頻値 (モード)、標準偏差、分散 (variance)、尖度 (kurtosis, excess)(bulge 内部からのふくらみ具合)、歪度 (skewness)、範囲 (range)、最小 (min)、最大 (max)、合計 (sum)、標本数 (size) が表示されます。

1. 最初の準備として表計算ソフトの設定。Excel の「ツール」→「アドイン」→「分析ツール」と「分析ツール-VBA 関数」にチェックがあることを確認
2. 「ツール」→「分析ツール」→「基本統計量」→「OK」
3. 基本統計量のメニューでは入力範囲に「データの範囲」を指定し、データの先頭行からすべてデータ値のときには「先頭行をラベルとして使用」のチェックマークをはずす。出力オプションは「新規またはつぎのワークシート」になっていますが、もし出力先を同じシートにする場合には、セルを指定します。基本統計量を出力するためには必ず「統計情報」にチェックを入れます。最後には「OK」とします。

出力された基本統計量の説明をします。標本の大きさ n とし、標本データを $\{X_i, i = 1, \cdots, n\}$ とします。これを大きさの順に並び替え（ソート, sort）したものを順位統計量 (Order Statistics) といい、$\{X_{(i)}, i = 1, \cdots, n\}$ と表します。

元データ	大きさ順に並び替え	順序統計量
$X_1, X_2, \cdots, X_n$	$\rightsquigarrow\rightsquigarrow\rightsquigarrow$	$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$

また和の計算は順序が交換可能ですから、$X_1 + X_2 + \cdots + X_n = X_{(1)} + X_{(2)} + \cdots + X_{(n)}$ となることに注意しておきます。もし X のとり得る値が k 個の $x_1, x_2, \cdots, x_k$ のいづれかである場合にはこれを階級別に整理して (k は固定された階級の数) には

階級値 (変数)	x_1	x_2	$\cdots$	x_k	計
度数	f_1	f_2	$\cdots$	f_k	n
相対度数	p_1	p_2	$\cdots$	p_k	1

の形でデータが与えられたとできます。大文字と小文字でこの 2 通りを区別していますが、$f_i = 1, i = 1, 2, \cdots, n$ とすれば、後者は前者の場合に帰着されます。これはそれぞれのデータがすべて異なった値をとっていることになります。

$$\sum_i X_i = X_1 + X_2 + \cdots + X_n = x_1 * f_1 + x_2 * f_2 + \cdots + x_k * f_k = \sum_j x_i f_j,$$
$$\sum_i f_i = f_1 + f_2 + \cdots + f_k = n$$
$$\sum_i X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2 = x_1^2 * f_1 + x_2^2 * f_2 + \cdots + x_k^2 * f_k = \sum_j x_j^2 f_j,$$
$$\frac{1}{n}\sum_i X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = x_1 * p_1 + x_2 * p_2 + \cdots + x_k * p_k = \sum_j x_j * p_j,$$
$$\frac{1}{n}\sum_i X_i^2 = \frac{1}{n}(X_1^2 + X_2^2 + \cdots + X_n^2) = x_1^2 * p_1 + x_2^2 * p_2 + \cdots + x_k^2 * p_k = \sum_j x_j * p_j^2,$$

平均　算術平均を計算します。$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\sum_{j=1}^k x_j f_j = \sum_{j=1}^k x_j p_j$ AVERAGE 関数の値と同じ。

標準誤差 SE　$SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n(n-1)}} = \sqrt{\frac{\sum_i X_i^2 - n\overline{X}^2}{n(n-1)}} = \sqrt{\frac{\sum_j x_j^2 f_j - n\overline{X}^2}{n(n-1)}} = \frac{\sqrt{\sum_j x_j^2 p_j - \overline{X}^2}}{\sqrt{n-1}}$

中央値（メジアン）Me　データを大きさの順に並べたとき、中央に位置する値。$Me = X_{((n/2)+1)}$(偶数)、$= X_{((n+1)/2)}$(奇数) データ数が奇数・偶数によって真ん中の値がかわる。数式メニューでは $= median$(データ範囲)、あるいは四分位数の 2 番目の値 $= quartile$(データ範囲, 2)

最頻値（モード）Mo　データの中で、最も頻度が高く (最大度数) 現れた値。MODE 関数と同じ。もし頻度の最大が 2 つ以上で同じ値を取るとき（つまり最大値がないとき）は、"#N/A"(Not Available) と表示される。度数の値 $\max_i f_i$ が最大となるような変数の値。データの中で、最も頻度が高く現れた値。MODE 関数と同じ。単峰ならばひとつに定まるが、もしひとつに定まらないような双峰形では存在しないとする。データがすべて、異なる値を取るときは、"# N/A"(Not Available) と表示される。

標準偏差 SD　$SD = s = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{(n-1)}}$ 標本分散の平方根を取ったもの。STDEV 関数と同じ。VAR 関数の平方根を取ったもの。

分散（標本不偏分散）s^2　VAR 関数と一致。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\binom{n}{2}}\sum_{i<j}(X_i - X_j)^2 = \frac{n}{n-1}\left(\sum_{j=1}^{k} x_j^2 p_j - \overline{X}^2\right)$$

VARP 関数は、分母が $(n-1)$ でなく、n を用いている。

$$v^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{k} x_j^2 f_j - \overline{X}^2 = \sum_{j=1}^{k} x_j^2 p_j - \overline{X}^2$$

尖度（せんど）KW　KURT 関数で与えられる量。$\frac{1}{n}\sum_i \left(\frac{X_i - \overline{X}}{s}\right)^4 - 3$ とがり具合を測る尺度となります。数 3 を引いている理由は、標準正規分布のばあいには、尖度が 3 となるから基準値として与えているから。つまり正の値ならば、正規分布の尖り具合と比べ、より鋭くなっている。負ならば、尖り具合が平坦であるという目安を与える。

歪度（わいど）SK　SKEW 関数で与えられる量。$\frac{1}{n}\sum_i \left(\frac{X_i - \overline{X}}{s}\right)^3$ 左右対称性を測る尺度となります。たとえば、L 字型の形状の分布では正の値をとり、逆 L 字型では負の値をとります。分布が L 字型であれば、代表値の大きさは　モード＜中位数＜平均　の順になり、もし逆 L 字型では、順序が、　平均＜中位数＜モード　の順になります。

範囲 R　データの最大値から、最小値を引いたもの。$R = \max(\text{データ範囲}) - \min(\text{データ範囲})$ このデータ範囲、最大、最小とデータの大きさ (標本数) から、度数分布表の階級幅や階級のクラスを決めていく。

最大　データの最大値。MAX 関数。データ $x_1, x_2, \cdots, x_n$ に対して $\max_i x_i$ を計算する。

最小　データの最小値。MIN 関数。データ $x_1, x_2, \cdots, x_n$ に対して $\min_i x_i$ を計算する。

合計　データの合計値。SUM 関数。データ $x_1, x_2, \cdots, x_n$ に対して $\sum_i x_i$ を計算する。

標本数　データ数を表すもので、n, N などがよく用いられる。COUNTA 関数で与えられる値。

信頼区間（95.0%）　信頼係数 95% に対する信頼区間の幅の 1/2。上の平均プラスこの値と平均マイナスこの値で出来る区間が信頼係数 95% の信頼区間を構成する。推定値の含まれる範囲を確率で評価したもの。

上記は基本統計量による出力ですが、これ以外にも重要な統計量があります。

トリム平均　TM(trimmed mean) データに異種的な要素が加わっていると、平均値は大きく影響を受ける。これを除去して残りのデータについて平均値をとることがある。切り落し平均ともよばれる。データ全体の上限と下限から一定の割合のデータを切り落とし、残りの項の平均値を返します。TRIMMEAN 関数は、極端な観察データを分析対象から排除する場合に利用します。

書式　TRIMMEAN(配列, 割合) : (1) 配列　対象となるデータを含む配列またはセル範囲を指定します。(2) 割合　平均値の計算から排除するデータの割合を小数で指定します。たとえば、全体で 20 個のデータを含む対象に対して割合に 0.2 を指定した場合、$20 \times\ 0.2 = 4$ となり上限から 2 個、下限から 2 個の合計 4 個のデータが排除されることになります。

平均偏差　MD (Mean Deviation) $= \frac{1}{n}\sum_i |X_i - \overline{X}|$ 各データと平均値との差に対する絶対値の和であるが、このように絶対値を考えることはデータと平均値との距離であり、ばらつきの尺度としてはごく自然なものである。

Z スコアー (偏差値)　$Z_i = \dfrac{X_i - \overline{X}}{s_x}$ 一次変換により、シフト (平均値の移動）とスケール（分散の拡大縮小）により平均値が 0 となり、分散あるいは標準偏差を 1 に変換する。さらに平均 50、標準偏差 10(分散 100) にしたもの、$10Z_i + 50$ を変量 X の偏差値という。知能指数は平均 100、標準偏差 15 にしたもの。

四分位数、四分位偏差　Q_1, Q_2, Q_3 とは、分布全体を 25% ずつの 4 つに分けて、最小値から 25% になるもの Q_1 を第一四分位数、第 2 四分位数は中央値と一致する 50% のところ、すなわちちょうど半分のところである。第 3 四分位数、75% のところは第 3 四分位数とよばれる。

数式入力では=QUARTILE(配列, 戻り値) となる。

ただし (戻り値, QUARTILE 関数の戻り値) = (0, 最小値) , (1, 第 1 四分位数 (25%)) (2, 第 2 四分位数 = 中位数 (50%)) (3, 第 3 四分位数 (75%)) (4, 最大値)

四分位偏差とは $Q_3 - Q_1$ であり、箱ひげ図（ボックスチャート）にも表示される。これ以外にもデータ数が大きければ、十分位数 (decitile) や百分位数 (percentile) などが用いられる。箱ひげ図 (ボックスチャート）では最小値、第 1 四分位数、中央値（第 2 四分位数)、最大値の 4 つの値ももちいて表示する。

変動係数　CV(Coefficient of Variation)= $\dfrac{s_x}{\overline{X}}$

測定のバラツキ (標準偏差) が平均値に対してどの程度の比率（割合）であるかという数値を表す。平均に対する相対誤差を示す量としても用いられる。とくに平均値が異なる集団のバラツキ具合を比較することに使われる。例えば、ＣＶ 10% とは、正規分布の場合、平均値±（平均値の 10%）範囲内、つまり、「平均値× 0.9」～「平均値× 1.1」の範囲内に全測定の 68.3% が入るという意味ですから、測定値が「0.31」であれば、100 回のうち 68.3 回が、「0.31 ± 0.031」の範囲内に入っている。（注：確率変数 X が $N(\mu, \sigma^2)$ に従う時、平均 μ からのずれが $\pm 1 \times \sigma$ 以下の範囲に　X が含まれる確率は $0.6826 \fallingdotseq 68.3\%$ $\pm 2 \times \sigma$ 以下だと 0.9544 となる）

6　共分散と相関係数

２変量の間のデータを分析するための代表値や分析手法はとくに使われることが多いし、共分散、相関係数などを述べる。ここでの共分散や相関係数はしばしば因果関係の根拠として扱われることが多い。 しかし、共分散自身はたんに 1 つの対象の 2 つの測定値が対応しているということの指標に過ぎないので、因果関係があるかどうかは示してくれない。あくまで数値から意味を解釈する程度である。一方、「共分散構造分析」など、複数の共分散を分析する手法では因果関係があるかどうかを直接検証する手法があるが、専門書を参照していただきたい。

標本共分散 (sample covariance): =COVAR(配列 A, 配列 B)　表計算の計算命令式:= $COVAR(x$ 範囲, y 範囲) と書き、数式では

$$Cov(x, y) = \frac{1}{n-1} \sum_i (X_i - \overline{X})(Y_i - \overline{Y})$$

を計算してくれる結果である。データ数が n であるが、割り算が $n-1$ となっている理由は、この数値が推定の不偏性から起因している。2 組の対応するデータ X, Y 間で、平均からの偏差の積を求め、平均値を計算したのも。平均からの大きい値は偏差が正で、小さければ負の値になるから、データ間の同符号の数値が多ければならば正、異符号が多ければ負の値となり、およその変量の（直線的な）増減傾向を捉えることができる。共分散が 0 に近い値ならば、顕著な関連性はないと考えられる。つまりこの

関連性は直線的なもの（1 次関数）に関する傾向の強弱を意味する。

一連の個別の対象物に対して測定される N 個の異なる測定変数がある場合、相関分析ツールと共分散分析ツールは同じ設定で使うことができます。相関分析ツールと共分散分析ツールは共に、測定変数の各組み合わせ間のそれぞれ相関係数または共分散を示すマトリクスが、出力テーブルとして得られます。相関係数が -1 から $+1$ までの範囲に収まるのに対し、対応する共分散はこの範囲に収まらない点が異なります。相関係数と共分散は共に、2 つの変数が一緒に変化する範囲で測定されます。共分散分析ツールは測定変数のそれぞれの組み合わせについて COVAR ワークシート関数の値を計算します。たとえば N=2 の 2 つの測定変数のみの場合は、共分散分析ツールではなく COVAR 関数を直接使用する方法が適しています。共分散分析ツールの出力テーブルで対角線上の i 行と i 列の値は、それ自身の i 番目の測定変数の共分散を表します。これは、VARP ワークシート関数で計算されるその変数に対する母集団の分散の値と同じです。共分散分析ツールを使うと、測定変数の組み合わせそれぞれについて 2 つの測定変数が一緒に変化する傾向があるかどうかを調べることができます。一方の変数の大きな値がもう一方の変数の大きな値と関連する傾向があるか (正の共分散)、一方の変数の小さな値がもう一方の変数の大きな値と関連する傾向があるか (負の共分散)、両方の変数の値が関連しない傾向があるか (0 に近い共分散) などを調べることができます。

標本相関係数 (sample correlation coefficient): =CORREL(**配列 A, 配列 B**) 表計算の計算命令式: $= correl(x$ 範囲$, y$ 範囲$)$ で定義式では

$$\rho = \frac{Cov(x,y)}{s_x s_y} = Cov\left(\frac{x}{s_x}, \frac{y}{s_y}\right)$$

ここで $\dfrac{x}{s_x}$, $\dfrac{y}{s_y}$ は $\dfrac{X_i - \overline{X}}{s_x}, \dfrac{Y_i - \overline{Y}}{s_y}, i = 1, 2, \cdots, n$ とし、これを基準化データともよばれる。

変量を直接計算する共分散ではなく、基準化したデータの場合に対する共分散を求めたもの。共分散は元の値の大きさで数値が決まるので単位が違う変数を複数比較するときなどに解釈が難しい。 そこで関係を見る場合にはこの相関係数を使うことが一般的である。

N 個の対象物それぞれに対して各変数の測定を行う場合、CORREL ワークシート関数と PEARSON ワークシート関数は共に 2 つの測定変数間の相関係数を計算します。いずれかの対象物に対する観察が行われないと、分析時にその対象物が無視されます。相関分析ツールは、N 個の対象物それぞれに対して 3 つ以上の測定変数がある場合に特に役立ちます。この分析を行うと、測定変数の可能な組み合わせそれぞれに対して適用された CORREL (または PEARSON) 関数の値を示した相関マトリクスが、出力テーブルとして得られます。共分散と同じように、相関係数は 2 つの測定変数が一緒に変化する範囲で測定します。共分散とは異なり、相関係数は 2 つの測定変数を表現する単位とは関係なくその値の基準が決められます。たとえば、2 つの測定変数が重量と高さの場合、重量がポンドからキログラムに変更されても相関係数の値は変わりません。相関係数のすべての値は、-1 から $+1$ までの範囲に収まる必要があります。相関分析ツールを使うと、測定変数の組み合わせそれぞれについて 2 つの測定変数が一緒に変化する傾向があるかどうかを調べることができます。一方の変数の大きな値がもう一方の変数の大きな値と関連する傾向があるか (正の相関)、一方の変数の小さな値がもう一方の変数の大きな値と関連する傾向があるか (負の相関)、両方の変数の値が関連しない傾向があるか (0 に近い相関) などを調べることができます。

図 6　散布図の例

-1 〜 -0.7	強い負の相関	0 〜 0.2	ほとんど相関なし
-0.7 〜 -0.4	かなりの負の相関	0.2 〜 0.4	やや正の相関あり
-0.4 〜 -0.2	やや負の相関あり	0.4 〜 0.7	かなりの正の相関
-0.2 〜 0	ほとんど相関なし	0.7 〜 1	強い正の相関

スピアマンの順位相関係数 (Speaman's rank correlation coefficient)　変量の値が順位で与えられているとき、$r_s = 1 - \dfrac{6\sum_i (x_i - y_i)^2}{n(n^2-1)} = 1 - \sum_i (x_i - y_i)^2 / \binom{n-1}{3}$ ここでデータ x_i, y_i は $1, 2, \cdots, n$ のうちのいづれか一つとなっている。もし同順位のある場合では、補正因子を入れて、それらの平均順位として調整する。同順位の補正因子=[COUNT(範囲) + 1 - RANK(数値, 範囲, 0) - RANK(数値, 範囲, 1)]/2 ただし RANK(数値, 範囲, 順序) で、順序：数値の順位を決めるため、範囲内の数値を並べ替える方法を指定します。順序に 0 を指定するか、順序を省略すると、範囲内の数値が ...3、2、1 のように降順に並べ替えられます。 順序に 0 以外の数値を指定すると、範囲内の数値が 1、2、3、... のように昇順で並べ替えられます。

2 変量の度数分布表は行列の形式で、横軸と縦軸にそれぞれの階級と階級値を並べ、要素には各変量に対する度数（頻度）を表す。2 変量のヒストグラムにおいては 3 次元の鳥瞰図等をもちいて表現するが、陰の部分は表示できない。度数が多くないときには、平面上にデータ点をプロットして表現できる。相関図（そうかんず）または散布図（さんぷず）とは、縦軸、横軸に 2 項目の量や大きさを対応させ、データを点でプロットしたものである。各データは 2 項目の量や大きさを表現している。

散布図には、視覚的に 2 項目の分布、相関関係を大まかに把握できる特長がある。データ群が右上がりに分布する傾向であれば正の相関があり、右下がりに分布する傾向であれば負の相関がある。相関係数が 0 であれば無相関となる。

クロス集計表の重要性

2 変量の関係を最もよく表現できるのは「クロス集計表」です。古典的かつシンプルな方法なので、「原始的」と見えますが、これが最も情報の欠落の少ない優れた方法なのです。

これをみると、「とても重視する」と答えた人が「満足」と「不満」に 2 極分化していることがわかります。ここになんらかの情報が隠れているかもしれません。さらに分析する必要がありそうですね。この例とは逆に「身長と学力に強い相関がある」というデータがあります。学年別に分けて相関係数を算出するとほとんど無相関だったという笑い話です。相関係数だけでなく、七面倒くさい式で表される統計量はなにか「ありがたい」ものとして、その結果だけをむやみに大事にしがちです。しかし、自分の感覚と「ズレ」があるものは疑う、またそうした感覚を養うことが大事であることはいうまでもありません。

図 7: クロス集計表の例

表計算ソフトでクロス集計を行なうにはピボットテーブルを使います。ピボットテーブルでは、大量のデータを持つ表から必要な項目を選択して、縦軸と横軸を交差させた集計表を作成できます。例えば、縦軸に「商品名」、横軸に「店舗」の項目を配置してそれぞれの「売上」を集計できます。

図 9　ピボットテーブルの名称

図 8　ピボットテーブルの起動

7　不平等解析

ジニ係数とは イタリアの経済学者 Corrado Gini が 1 9 1 2 年に発表した。計算は累積分布関数から求め、公平さと不公平さの違いの程度を表すものと解釈され、0 と 1 の間の値となる。所得の例では変数値の

図 10　ピボットテーブルの結果

所得金額に対して、その所得金額以下の割合を縦軸に表していく。ジニ係数の値 0 とは、所得金額が公平に配分されている状況を表す。すなわち変数値の金額に対して、割合の増加が直線になる。もう一つの極端な係数値が 1 とは、ある人がすべての所得金額を占めており、その他の人々が所得金額ゼロとなる場合である。このようにジニ係数が低い値であることは、社会の富の配分（所得金額の値）がほぼ平等に分けられていることを示し、数値が高いことは、全体の少数部分が高い富を占有していることとなる。Wikipedia (Gini coefficient) を参照。また 100 倍してパーセント表示にしたものをジニ指数（Gini Index）をよぶ。http://www.sustainablemiddleclass.com/Gini-Coefficient.html による近年のジニ指数の値比較：

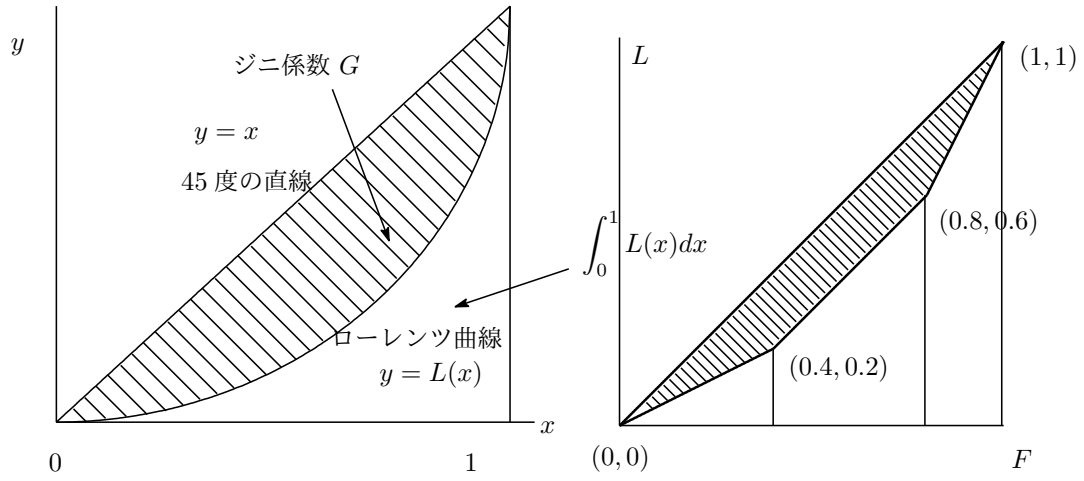
Japan 24.9	United Kingdom 36.0	Sweden 25.0	Iran 43.0
Germany 28.3	United States 46.6	France 32.7	Argentina 52.2
Pakistan 33.0	Mexico 54.6	Canada 33.1	South Africa 57.8
Switzerland 33.1	Namibia 70.7		

このようなジニ係数 (指数）の結果は、経済学での貧富の指標、公平さの変化を調べるために、社会変化や政治活動のひとつとして公表されている。たとえば、発展国としてヨーロッパ諸国では係数値は 0.24 から 0.36 であるのに対して、アメリカ合衆国では 0.4 を超えている。つまりアメリカとは、貧富の差の大きい国である。また政治哲学や政策の補填の意味でも、このジニ係数は役立つと考えられる。しかし大きな国と小さな国で比較を行なうときには、当然誤解をもたらすことを注意しておかねばならない。世界全体のジニ係数は、およそ 0.56 から 0.66 の間といわれている。Bob Sutcliffe (2007), Postscript to the article‘World inequality and globalization’(Oxford Review of Economic Policy, Spring 2004), ¡http://siteresources.worldbank.org/INTDECINEQ/Resources/PSBSutcliffe.pdf¿. Retrieved on 2007-12-13

ジニ係数の計算式：

$$G = 1 - 2\int_0^1 L(x)dx = \frac{1/2 - \int_0^1 L(x)dx}{1/2} \quad \text{or} \quad = 1 - 2\sum_i L(x_i)(i.e.L(x)\text{ の面積})$$

ここで $L(x)$ はローレンツ曲線で、



つぎで定める：

(1) 離散型分布のとき；$i = 1, 2, \cdots, n$

値	x_i	x_1	x_2	$\cdots$	x_n
確率	$p_i = f(x_i)$	p_1	p_2	$\cdots$	p_n
累積確率	F_i	$F_1 = p_1$	$F_2 = p_2 + F_1$	$\cdots$	$F_n = p_n + F_{n-1} = 1$

$$L_i = \sum_{j=1}^{i} x_j f(x_j) \Big/ L = \sum_i \frac{L_{i+1} + L_i}{2}(F_{i+1} - F_i) \quad (i = 1, 2, \cdots, n)$$
ただし $L = \sum_{j=1}^{n} x_j f(x_j)$

性質 : $L_0 = 0 \leq L_1 \leq L_2 \leq \cdots \leq L_{n-1} \leq L_n = 1$ 増加関数で、下にとつの形をし、0 から 1 まで変化する。

(2) 連続型分布のとき ; $-\infty < x < \infty$

値	x
確率	$f(x)$
累積確率	$F(x) = \int_{-\infty}^{x} f(t)dt$

$$L(x) = \int_{-\infty}^{x} tf(t)dt \Big/ L \quad (0 \leq x \leq 1) \text{ ただし } L = \int_{-\infty}^{\infty} tf(t)dt$$

性質 :

(i) $L(0) = 0 \leq L(x) \leq \cdots \leq L(y) \leq L(1) = 1, 0 < x < y < 1$

(ii) $0 < L(x) \leq x, 0 < x < 1$

つぎにジニ係数はデータ $x = (x_i, i = 1, 2, \cdots, n)$ とそれに対する累積相対度数 $F = (i/n, i = 1, 2, \cdots, n)$ との共分散 s_{xy} を含む式で表現できることを注意する。あるいは少し変形するとジニ係数 G は、変動係数 CV(x) と相関係数 R_{XF} との積として表現できることを述べる。

$$G = \text{定数} \times CV(x) \times R_{XF}$$

母集団の大きさが n で、データを $X = \{x_1, x_2, \cdots, x_n\}$ とする。平均を $\overline{x} = \frac{1}{n}\sum_i x_i$ とし、各データの偏差 (データ間の距離) $\|x_i - x_j\|, i, j = 1, 2, \cdots, n$ とすると、散布度の指標を表す。この総計を個数で割った総計の偏差 (平均差)D は

$$D = \frac{1}{n^2}\sum_{i,j} |x_i - x_j| = \frac{2}{n^2}\sum_{i<j} |x_i - x_j|$$

ここで和の記号では $\sum_{i,j}$ は n^2 通りの組 $\{(i, j); i, j = 1, 2, \cdots, n\}$ にわたる 2 重和で、$\sum_{i<j}$ は $n(n-1)/2$ 通りの組 $\{i < j; i, j = 1, 2, \cdots, n\}$ とする。

ジニ係数 G の定義式は

$$G = \frac{D}{2\overline{x}}$$

であり、さらにデータ $\{x_1, x_2, \cdots, x_n\}$ を昇順に並び替えた順序統計量 (order statistics) を

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

と表す。n 次行列として三角形成分の各列和 (縦) を求め、行に加え合わせれば計算を行うと、つぎのように変形できる。

$$\begin{aligned}
\textstyle\sum_{i<j} |x_i - x_j| &= |x_1 - x_2| + |x_1 - x_3| + \cdots + |x_1 - x_n| \\
&\quad + |x_2 - x_3| + \cdots + |x_2 - x_n| \\
&\quad + \cdots + \cdots \\
&\quad + |x_{n-1} - x_n| \\
&= \big(x_{(2)} - x_{(1)}\big) + \big(x_{(3)} - x_{(1)}\big) + \cdots + \big(x_{(n)} - x_{(1)}\big) \\
&\quad + \big(x_{(3)} - x_{(2)}\big) + \cdots + \big(x_{(n)} - x_{(2)}\big) \\
&\quad + \cdots + \cdots \\
&\quad + \big(x_{(n)} - x_{(n-1)}\big)
\end{aligned}$$

したがって右辺式を整理して

$$\begin{aligned}
& \textstyle\sum_{i<j} |x_i - x_j| \\
= \;& (n-1)x_{(n)} + ((n-2)-1)\, x_{(n-1)} + \cdots + (1-(n-2))\, x_{(2)} + (0-(n-1))\, x_{(1)} \\
= \;& \textstyle\sum_i i \times x_{(i)} - \sum_i (n-i+1)x_{(i)} \\
= \;& 2\textstyle\sum_i i \times x_{(i)} - n(n+1)\overline{x} = 2\sum_i i\left(x_{(i)} - \overline{x}\right)
\end{aligned}$$

ここでは、$\sum_i x_i = \sum_i x_{(i)} = n\overline{x}$, $\sum_i i = n(n+1)/2$ を用いた。

順序統計量データ $x = (x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ とデータ $F = (F_i) = \left(\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n}\right)$ とし、これらの平均、分散さらに共分散を計算する。

$$\overline{F} = \frac{1}{n}\sum_i \frac{i}{n} = \frac{n+1}{2n}$$

$$\begin{aligned}
s_F^2 &= s_{FF} = \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{(F_i - F_j)^2}{2} = \frac{1}{n-1}\sum_i (F_i - \overline{F})^2 = \frac{n+1}{12n} \\
s_{XF} &= \frac{1}{n-1}\sum_i (x_{(i)} - \overline{x})(F_i - \overline{F}) \\
&= \frac{1}{n-1}\sum_i (x_{(i)} - \overline{x})\left(\frac{i}{n} - \frac{n+1}{2n}\right) \\
&= \frac{1}{n-1}\sum_i (x_{(i)} - \overline{x})\left(\frac{i}{n}\right)
\end{aligned}$$

より、したがって変量 X と F との相関係数 ρ は

$$\rho = \frac{s_{XF}}{s_X s_F} = \frac{1}{n-1}\sum_i (x_{(i)} - \overline{x})\left(\frac{i}{n}\right) \Big/ \left(s_X\sqrt{\frac{n+1}{12n}}\right)$$

$$\therefore D = \frac{4}{n^2}\sum_i i\left(x_{(i)} - \overline{x}\right), \qquad \frac{nD}{4} = \sum_i \left(\frac{i}{n}\right)(x_{(i)} - \overline{x}) = S_{xy}$$

これをジニ係数の定義式に代入すると

$$G = \frac{D}{2\overline{X}}$$

において、

$$D = \frac{1}{\binom{n}{2}}\sum_{i<j} |x_{(i)} - x_{(j)}| = \frac{4}{n(n-1)}\sum_i i\left(x_{(i)} - \overline{x}\right)$$

$$\frac{D}{2\overline{x}} = \frac{2}{(n-1)\overline{x}}\sum_i \left(\frac{i}{n}\right)(x_{(i)} - \overline{x}) = \frac{2}{\overline{x}}s_{XF} = \frac{2}{\overline{x}}\rho s_X s_F = \frac{2}{\overline{x}}\rho s_X\sqrt{\frac{n+1}{12n}} = \sqrt{\frac{n+1}{3n}}\rho CV(X)$$

もし n が十分に大きいならば、$\sqrt{\frac{n+1}{3n}} \approx 0.577$ と近似できるから、ジニ係数 G

$$G = \frac{\sum_{i,j} |x_i - x_j|/n^2}{2\overline{x}} = \frac{1}{n^2\overline{x}}\sum_{i<j} |x_i - x_j| \approx 0.577\rho CV$$

8　回帰分析

表計算ソフトの回帰分析ツールは、線形回帰分析を行います。回帰分析では、R-2 乗値を使って、観測値のデータが最適な直線に当てはめられます。このツールを使って、複数の独立変数が 1 つの従属変数に与える影響を分析することができます。たとえば、スポーツ選手の年齢、身長、体重などの要素が成績に与える影響を分析できます。成績データに基づいて、これらの要素それぞれが成績に影響した比率を割り当てたり、回帰分析の結果を使って、ほかのスポーツ選手の成績を予測することもできます。生ビールの生産計画では、従属変数を数ヶ月先の生ビール生産量とし、独立変数：各年の生産量、各月の生産量、長期天気予報（気温）などとして予測をします。またクレジットカードでのローン評価では、従属変数；信用評価を独立変数；勤務年数、年収、年齢、持ち家/借家などから求める例もあります。回帰分析ツールは LINEST ワークシート関数を使用します。

回帰分析とは、ある原因に対し、結果となる数字がどのような関係を持っているかを調べます。例えば、原因となる値を X として、結果となる値を Y とすると、次のような式で表すことができるとします。

線形 (1 次）式 $Y = a \cdot X + b$、　べき乗の式 $Y = a \cdot X^b$、　指数関数 $Y = a \cdot e^X$、　対数関数 $Y = a \cdot logX$

基本は一次の線形関係式で、これらは変換によって帰着できます。ここで R2 値とは、ある現象がその回帰式で表される確率というように考えれば OK です（したがって、例のグラフの場合 9 割方は回帰式で説明がつくということになります)。R2 値は一般的には、0.5〜0.8 なら、回帰式成立の可能性がありそうで、0.8 以上ならばかなり成立の可能性が高いことを示します。(直線近似の場合は、R2 ＝相関係数の 2 乗になります）

[注意] 分析ツールは Microsoft Office Excel のアドイン (アドイン：Office に独自のコマンドや独自の機能を追加する追加プログラムです。) プログラムで、Microsoft Office または Excel をインストールすると利用可能になります。ただし、Excel で分析ツールを使用するには、最初に分析ツールを読み込む必要があります。このためには (Microsoft Office ボタン) をクリックし、[Excel のオプション] をクリックします。[アドイン] をクリックし、[管理] ボックスの一覧の [Excel アドイン] をクリックします。

●エクセルを使った回帰分析

1. エクセルで回帰分析をするときは、グラフ上で、近似曲線を追加するを選択します。近似曲線の描き方
2. 近似曲線の種類を選びます。近似曲線の種類
3.「グラフに数式を表示する」と「グラフに R-2 乗値を表示する」を追加します。近似曲線の R 値

●重回帰分析

重回帰分析とは、次の式のように、ある結果となる変数 Y に対して、原因となる変数が 2 つ以上ある場合 $(X_1, X_2, \cdots, X_n)$ と定数 b で行う回帰分析のことです。

$$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_n \cdot X_n + b$$

つぎのような例を考えます。

各行について（A 列）(Y) には、（B 列）(X_1) と（C 列）(X_2) の変数と定数 b からなる関係があるとします。$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + b$ において、a_1, a_2, b を得られたデータの関係式から推定します。スプレッドシートには入力された値が次であったとします。分析ツールを用いた結果が「概要」として、右側の表です。

図 11　回帰直線

10 個のデータ

	A 列	B 列	C 列
1 行	Y	X1	X2
2 行	10	18	10
3 行	12	17	11
4 行	3	3	2
5 行	14	26	15
6 行	4	7	5
7 行	10	18	9
8 行	6	10	6
9 行	11	15	13
10 行	8	15	7
11 行	11	14	14

回帰統計	
重相関 R	0.98
重決定 R2	0.96
補正 R2	0.95
標準誤差	0.78
観測数	10.00

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2.00	110.61	55.31	90.29	0.00
残差	7.00	4.29	0.61		
合計	9.00	114.90			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.82	0.65	1.26	0.25	−0.72	2.36	−0.72	2.36
X 値 1	0.25	0.07	3.53	0.01	0.08	0.42	0.08	0.42
X 値 2	0.49	0.11	4.47	0.00	0.23	0.74	0.23	0.74

重回帰分析は、エクセルの「ツール」→「分析ツール」の中にある回帰分析を使うと求められます。あるいは「データ」→「データ分析」→「回帰分析」

図 12　回帰直線

ここから、次のような関係が導かれます。

$$Y = 0.25 \cdot X_1 + 0.49 \cdot X_2 + 0.82$$

重回帰分析では、次の 2 つの値に注意する必要があります。（いずれも上のやり方で自動的に出てくる数字です）

■ t 値 t 値、エクセルを使って、回帰分析をすると自動的に出てきます。この値が大きい変数は、結果として出てくる変数（例の場合 Y）との関係性が強くなります。また、この値が 2 を超えているかどうかが、原因となる変数（例の場合 X）として採用するかどうかの判断材料になります。

■ P 値 P 値も、エクセルと使うと自動的に出てきます。この値が、0.05 よりも大きいときは、原因となる変数（例の場合 X）として採用しないほうがよいとされています。

分散分析の結果はモデル全体が意味のあるものであるかを検定した結果です．帰無仮説は「すべての係数＝0」となっています．係数の欄の t 値，p 値は推定された「係数が 0 である」という帰無仮説を検定したものです．検定結果から帰無仮説が棄却できない，すなわち統計的に 0 でないとはいえない（0 かもしれない）となると，Y と X の関係がないことになるので，この回帰分析に意味がなくなります．判断の仕方は以下の通り．

(1)　$P(T \leqq t)$ < 実験者が設定する棄却域の確率　帰無仮説を棄却

(2)　t 境界値 <「t」の絶対値　帰無仮説を棄却

t 境界値は tinv 関数を用いて得ることができます．確率と自由度から、「t 境界値= tinv（確率，自由度）」を計算します。

確率には多くの場合，5%(0.05)，1%(0.01) といった確率を用います．自由度には，「$n-k-1$」の数が入ります．ここで，n はサンプル数，k は回帰式に用いた独立変数の数です．この tinv 関数が返す数値は両側検定のものです．片側検定の値を求める場合は，引数に用いる確率を 2 倍します．5% 片側の数字の場合は，0.1 となります．

また、重回帰分析では、相関の強い変数を 2 つ以上採用することは避けるべきだと考えられます。例えば、家賃の変数として、駅からの距離と地価を変数にした場合、駅からの距離と地価に強い相関があると、どちらかの数字が t 値または P 値の基準が NG になってしまいます。

●相関係数・回帰分析を用いる際の注意点

■分析の前提・次のアクションにつながるかを押さえる回帰分析は、現象の傾向を表すのに非常に有効なツールです。しかし、回帰分析はやり方次第で、いくらでもそれらしい線を引くことができます。うまく近似曲線をひけたとしても、その近似線で説明できる前提を押さえたり、その近似曲線がわかることで次にどんなアクションにつなげられるかを考えたりすることが非常に重要です。

■相関があるからといって因果関係があるわけではない回帰分析で高い相関が発見できても、それらに因果関係があるとまでは言い切れないません。実際にグラフに示したみたり、定性的に考えて第 3 の因子を考えたりすることが重要になります。

■相関から外れたところの扱いに注意回帰分析をしてグラフを見ると、近似線から明らかに外れたデータが出てくる場合があります。こうしたデータにはビジネス上の大きなヒントが隠されている場合があるので注意して掘り下げてみることも必要です。もちろん単なるノイズとして、データを省ける場合もあります。

■相関の高低の判断はビジネスの種類によって違う相関があるというためには、相関係数が上述のように一般的に絶対値で 0.7 くらい（R2 値だと 0.5 くらい）必要ですが、ビジネスの性質によってはそれ以下でも相関関係を深堀して考える場合があります。例えば相関があった場合のリスクが極めて大きい場合などは、相関係数（あるいは R2 値）が低くてもしっかり内容を調査していきます。

独立変数（説明変数）を選択する際、マーケティングやアンケートでよく使う一般的な重回帰の場合、複数の説明変数同士は無相関という仮定が入っている。そのため、説明変数同士が関連性の高い場合、多重共線性と呼ばれる状態になるため、係数が直感に反する値になることがあるので注意する必要がある。例えば、小学校での定期テスト得点から重回帰で分析する場合、理科の点数を従属変数に、数学と国語とを説明変数にした場合、数学が増えると理科の点数が増え、国語の点数が高ければ理科の点数が減るといった意味の係数が出ることがある。これは数学と国語との点数の間に強い相関がある（一般に、どちらの成績も学習習慣や知能の影響を強く受ける）ことで起こりうる。この場合のように説明変数間の相関が高いと係数が不安定になりやすい。

実務的対応としては、一方を除いて分析するのが最も手軽である。また、数学と国語の平均点と、数学と国語の得点の差というように和と差に数字を加工すると、この二つは相関がたいてい低く、かつ解釈しやすい。数学と国語の得点の差は、数学の方が高い生徒の方が理科の点数が高い傾向があるというように理解できるためである。ただし、このような正の相関を持つ変数同士の差得点は元の変数よりも信頼性が落ちるので、サンプル数を増やすなどの対応が求められる。

また、適切な予測力を実質的には持たない変数であっても、説明変数に加えると予測式自体の説明力（R2）は上がることが多いので注意する必要がある。そのため、単なる R2 ではなく、その分を調整した修正 R2 を参照する、ステップワイズ法等で投入する説明変数を取捨選択する、AIC を見るなどの対応が求められる。

9　インターネットによる統計データ資料

統計データを公開しているホームページには、代表的なものとして、総務省統計局の統計データ

`http://www.stat.go.jp/data/index.htm`

があります。

千葉県の統計情報

`http://www.pref.chiba.jp/outline/statistics/index-j.html`