３章 データと集計

3.1 母集団と標本

統計的な調査には大きく分けて２通りの方法があります。１つは調査対象を全部調べる全数調査で、もう１つは調査対象の中から、ランダムにいくつかの対象を取り出して調べる標本調査です。全数調査の代表的なものには 5 年ごとに政府が実施する国勢調査があります。また、１つの事業所などを対象としたアンケート調査も全数調査の１つです。全数調査は対象の状態が完全に分かるわけですから、これ以上のものはありません。しかし、大規模な調査は大変な金額と手間を必要としますので、なかなか実現できません。それゆえ比較的少ないデータ数で効率を上げることのできる標本調査が重要になってきます。

標本調査は図 3.1 のように、母集団と呼ばれる調査対象から、標本と呼ばれる調査データを取り出します。母集団のデータ数を N、標本のデータ数を n とすると、多くの場合 $N >> n$（N は n よりずっと大きい）を仮定しています。もちろんこれ以外の場合の厳密な計算もありますが、ここでは取り上げません。

母集団の例としては日本の成人男女、日本の６５歳以上、日本の中小企業等、事実上全数調査が困難な対象です。この母集団から乱数表等を使って、偏りがないように標本を抽出しますが、このような抽出法をランダムサンプリングといい、データ収集の基本となっています。ランダムサンプリングされたデータは、後に学ぶ推定や検定に用いることができます。

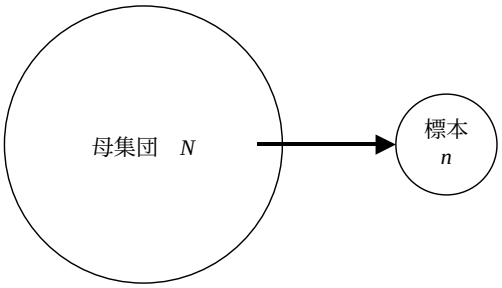


図 3.1 母集団と標本

3.2 データの種類

ここでは統計で扱うデータについて見てみようと思います。データは２つに分類されます。１つは、身長や体重、企業の売上高等、データの値に意味のある量的データです。もう１つは、事象の出現数や出現割合または、アンケート調査の選択問題などのように回答数や回答割合が重要な質的データです。但し、量的データでも階級に分類すると質的データと解釈できます。統計学ではこの 2 種類のデータを集計分析する手法を学びます。

問題

以下のデータを量的データと質的データに分類せよ。

商店の売上高，テストの点数，「はい」と「いいえ」の回答，蛍光灯の寿命，サイコロの目の出現頻度，体力測定のデータ，個人消費額

解答

量的データ　商店の売上高，テストの点数，蛍光灯の寿命，体力測定のデータ，個人消費額

質的データ　「はい」と「いいえ」の回答，サイコロの目の出現頻度

3.3 データの収集

ここではこれらの量的データや質的データを収集する方法を考えてみましょう。まず調べることは、これまでに様々な機関が収集したデータです。同じことをもう一度調べなおすことは、費用と労力の無駄になりますから、これらのデータはできるだけ利用しましょう。このデータには、政府刊行資料や会社固有の資料等があります。データは出版物としてまとめられてもいますが、インターネットに掲載されている場合も少なくありません。次に、実際にデータを得るために測定を実行する場合もあります。例えば、現地での調査、サンプルを実験設備のある場所に持ち帰っての調査、実験等多くの方法があります。また、アンケート調査を実施することもあります。

アンケート調査には様々な方法がありますが、ここでは代表的なものを紹介します。調査方法としては、調査する対象に直接会って質問する面接法、調査世帯などに行って説明し調査票を置いて後から回収に行く留置き法、郵便で調査票を送り回答を送り返してもらう郵送法、調査対象に電話して質問する電話法等、様々な方法が予算や労力、調査期間などに応じて実施されています。できるだけサンプリングの偏りをなくすために、回収率を上げる必要がありますが、郵送法は思ったほど回収率が上がらないのが現実のようです。解答の精度を高めようと思えば、面接法や電話法、予算や人手が少ないなら郵送法といった具合に、調査方法は状況に応じて選びます。

アンケートの質問回答方法にも様々な方法があり、知りたい内容や調査対象の特徴に応じて答え易く、まとめ易いものを選びます。例えば量的データについては値をそのまま書き込んでもらいますが、回答の中から選択して答えてもらう選択的回答法、同じ複数回答の中から順序を付けて選んでもらう序列回答法、全く自由に文章で回答してもらう自由回答法などが代表的方法です。

3.4 データの集計方法

データが入手できたら次はデータの特徴を調べるために集計を行います。集計の方法は質的データと量的データとでは異なりますので、ここでは分けて説明します。また、たくさんあるデータのうち 1 種類のデータ（１つの変数）に絞って見る方法を単純集計、複数種類（主に 2 種類）のデータについて関係を見る方法をクロス集計といいます。

1．質的データ

まず、質的データの単純集計の方法を見てみましょう。質的データとは基本的に事象が何回現れたかとか、アンケートで何人が丸を付けたかという度数ですから、その度数を項目ごとに分割して表示します。ここでその例として、工場における不良品の発生と曜日の関係を見てみます。生の調査データはおそらく下の表 1.1 のように与えられているでしょう。発生曜日については曜日を 1:月曜 2:火曜 3:水曜 4:木曜 5:金曜 のように、番号で表わしている場合が多いでしょう。集計はこのデータから、表 1.2 のように曜日別に度数を分割した表を作ります。

表 1.1 調査データ

標本番号	発生曜日
1	3
2	5
3	4
： ：	： ：
60	3

表 1.2　曜日と不良品の発生件数

曜日	月曜	火曜	水曜	木曜	金曜	合計
発生件数	１２	９	１０	１２	１７	６０

このように１つの変数について度数の分割を行ったものを 1 次元分割表といいます。1 次元分割表をつくると各曜日の比較が容易にできます。合計の部分はない場合もあります。

このままでも集計としてはよいのですが、さらにこれを視覚化することを考えます。発生度数に重点を置く場合は、これから棒グラフを作ります。棒グラフは状況に応じて縦棒グラフでも横棒グラフでも構いません。また、各曜日の割合に重点を置く場合には円グラフを描きます。円グラフには各曜日の割合を%表示で付与するとさらに分か

り易くなります。図 3.2a と図 3.2b に Excel を利用して描いた例を示します。自分で実際にやってみて下さい。

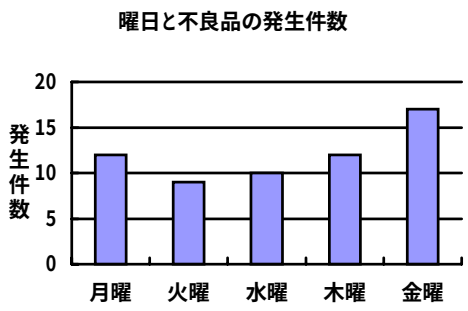


図 3.2a 棒グラフの例

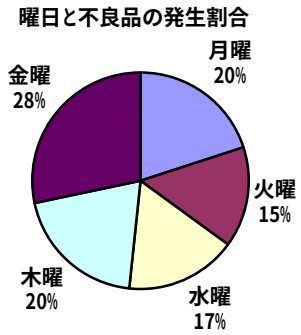


図 3.2b 円グラフの例

問題

Excel を用いて表 1.2 の 1 次元分割表から、図 3.2a,b のグラフを描け。

解答

省略

次に複数の変数を用いてその関係を見るクロス集計について学びます。例としてある案に賛成か中立か反対かをみるアンケート調査の結果について考えます。アンケート調査には 1:男 2:女を選択する項目があり、設問について 1:賛成 2:中立 3:反対 の中から答えを選ぶことにします。実際のデータは表 3.3 のように与えられるものと思います。この生データをもとに度数を男女別に集計して、表 3.4 のような分割表を作ります。分類を性別と回答と 2 つで行っていますので、2 次元分割表といいます。

表 3.3 アンケートデータ

標本番号	性別	回答
1	1	2
2	1	1
3	2	3
:	:	:
:	:	:
162	1	3

表 3.4 2 次元分割表

	賛成	中立	反対
男	42	18	23
女	34	25	20

ここで項目ごとの合計を加えると、さらに分かり易い表示となるでしょう。2 次元分割表をグラフ化する際に最もよく使われるグラフは積み重ね（積み上げ）棒グラフです。図 3.3 には Excel を用いた例を示しています。積み重ね棒グラフにはこの図のように、棒の上端が度数の合計を表わしているものと、上端を固定して 100%を表わしているものとがありますが、後者は比率を見る場合に使われます。一般には前者のグラフの方がよく利用されていると思います。

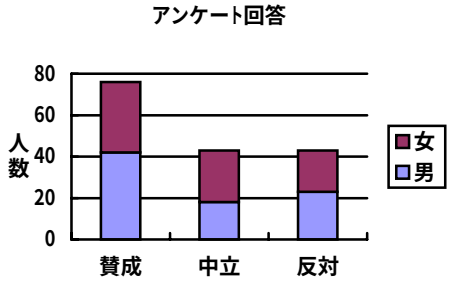


図 3.3 積み重ね棒グラフの例

このグラフは *x* 軸に回答項目を取りましたが、もう 1 つの項目の男女をとってはどうでしょうか。形式としては可能ですが、その場合、棒の高さは男女の人数になります。我々は、第一に回答結果を見たいわけで、*x* 軸にはやはり回答項目を選ぶべきだと思います。

問題

20 人にアンケートを取ったところ、次のような結果を得た。以下の問いに答えよ。

性別	回答	性別	回答	性別	回答	性別	回答
1	1	2	3	1	2	2	1
2	1	1	1	1	2	1	2
1	2	2	2	2	1	1	1
1	2	2	3	2	1	2	3
2	1	1	1	1	3	1	1

注）性別：　1：男　2：女

回答：　1：はい　2：いいえ　3：分からない

1) 回答に関する 1 次元分割表を描け。

2) 性別と回答に関する 2 次元分割表を描け。

3) 1)の分割表を用いて棒グラフと円グラフを描け。

4) 2)の分割表を用いて積み重ね棒グラフを描け。

解答

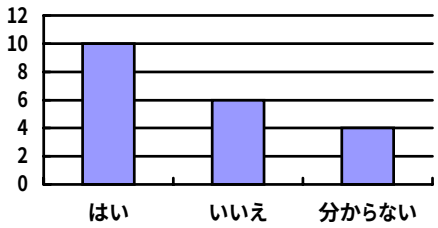
1)

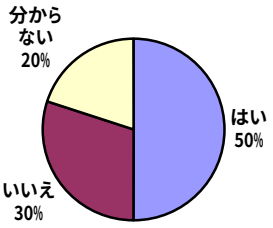
はい	いいえ	分からない	合計
１０	6	4	２０

2)

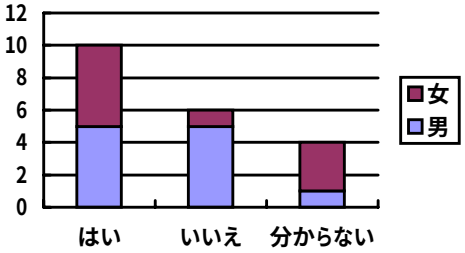
	はい	いいえ	分からない	合計
男	5	5	1	１１
女	5	1	3	9
合計	１０	6	4	２０

3)





4)



2．量的データ

数値が意味を持つ量的データは、どんな値が中心なのか、どの位の広がりなのか、ある値の周辺にどの位の度数があるのかを調べてみたくなります。１変数についてのこれらの単純集計の方法を説明するにはページ数が必要ですので、章を改めて述べようと思います。

ここではクロス集計の方法について簡単に説明しておきます。以下の表 3.5 を見て下さい。20 人の身長と体重のデータです。

表 3.5 身長と体重

身長(cm)	体重(kg)	身長(cm)	体重(kg)
169	71	170	62
175	68	180	75
170	67	177	70
179	72	175	70
176	69	172	62
174	80	166	58
173	75	168	60
181	65	173	58
179	74	169	59
178	71	170	73

このデータから身長と体重の関係を見る１つの方法は、図 3.4 で与えられる散布図と呼ばれるものを描いてみることです。散布図は分布図または相関図とも呼ばれ、１つの変数を横軸に、もう 1 つの変数を縦軸に取って点を打ちます。片方の変数がもう一方の変数の原因となっている場合、横軸に原因となる変数を取ります。図の点のばらつき方によってその 2 つの変数の関係が見えてきます。

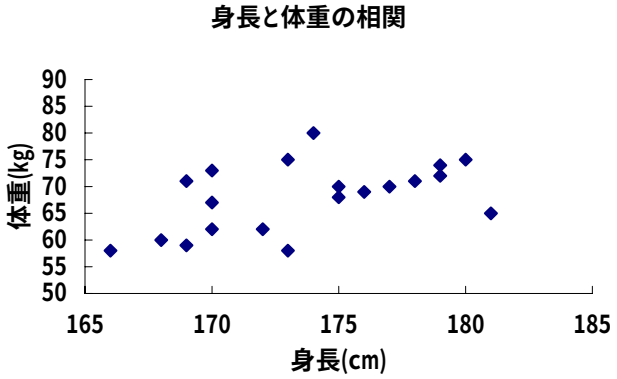


図 3.4 身長と体重の散布図

もし、2 つの変数の間に 1 次式 $y = ax + b$ の関係が成り立つならば、点は直線状に並びます。データがどの程度直線状に並んでいるかを表わす指標としてよく用いられるのが、相関係数です。相関係数は r という記号で表わされることが多く、$-1 \le r \le 1$ の範囲の値を取ります。$|r|$ が 1 に近いほど強い相関を持つと考えます。Excel では、2 つの変数のデータ範囲を、範囲 1、範囲 2 として、相関係数は以下のような関数で表わされます。

相関係数 = correl(範囲 1,範囲 2)

この値がおよそ 0.5 程度以上あれば相関があると考えるべきでしょう。この例では $r = 0.513$ となります。これについての詳細も後ほど紹介します。

問題

上の例題のデータから、実際に図 3.4 のような散布図を描き、相関係数を求めよ。

解答

省略