確率統計II講義ノート (補足)

Yasushi Ishikawa ２０１９

第1章　7章

AA

カレーライスには、生卵？　ゆで卵？

たとえばカレーライスに卵を加えるなら生卵かゆで卵かという問いに対する読者投票の結果は、関西以西が生卵、東日本はゆで卵で、その中間地帯の中部地方は生とゆでが混在していた。明治時代に開店した大阪の洋食店が、カレーに生卵を落として提供したのが始まりのようだ。それが西に向かって広がった。

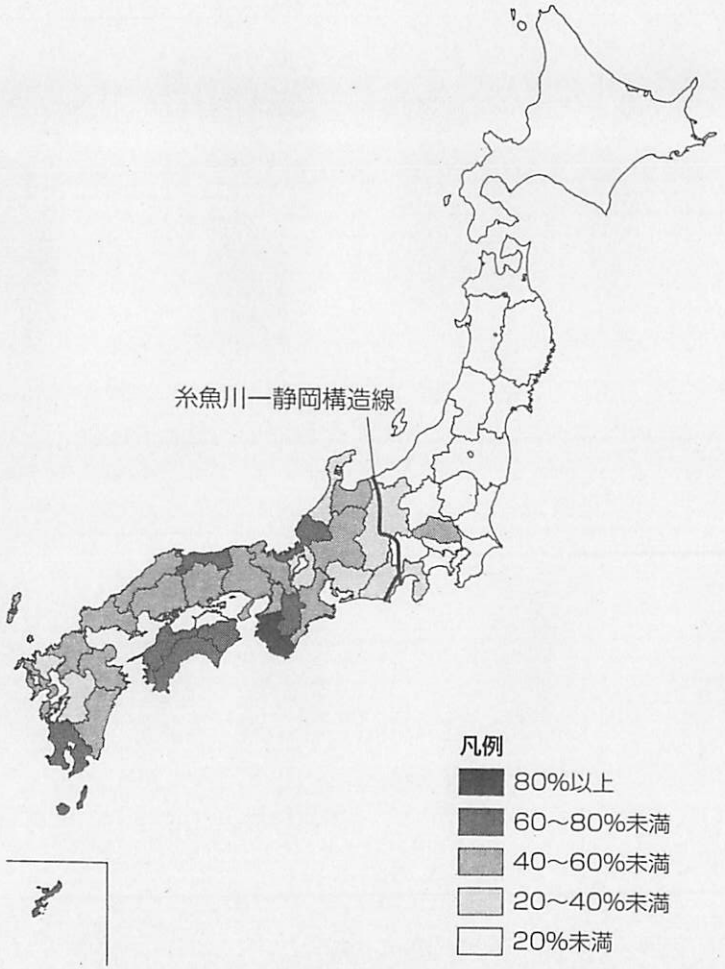
「ぜんざい」と呼ぶのは西日本、「お汁粉」が東日本、「焼き飯」は西、「チャーハン」が東と、さまざまな日本地図が出来上がっていった。

日本経済新聞の電子版発刊に伴って「列島あちこち　食べるぞ！　B級ご当地グルメ」という新企画を立ち上げ、読者とともに各地のご当地グルメの発見・発掘を続けてきた。

「食べ物　新日本奇行」から数えると十三年余にわたる長期連載となり、サイト内と私の頭の中は食文化の分布図でいっぱいになった。

しかしながら読者の投票なり投稿は二次情報である。そこで私は食べ物や分布状態の確認のため現地取材を続けた。東京を除く四十六道府県をくまなく歩くために、ある県

図1　天ぷらにソースをかける人の割合



（日本経済新聞電子版「食べ物　新日本奇行」を基に作成）

のである。近代豚肉食発祥の地ともいうべき地域で、今は牛が優勢なのがおもしろい。

これは、この三都県が出身地混合地帯なのが影響していると考えられる。また、これは「何の肉を思い浮かべるか」というアンケートであって、「普段何の肉を食べているか」という設問ではないため、両者のギャップを多少念頭に置く必要があるだろう。

いずれにせよ、トンカツ料理が東京で生まれたというのが実にキーポイントで、もし豚で生まれていたら、この分布図は逆になっていた可能性もある。

食文化の伝播（でんぱ）はなかなか一筋縄ではいかないおもしろさを秘めているのだ。

「ご馳走と言ったら鶏だよね」の九州人

牛、豚ときたら、次は鶏（とり）である。

そして、鶏といえば圧倒的に九州なのである。

もちろん、九州には薩摩や長崎市の豚肉文化や、熊本県の馬肉文化があるし、次章で述べる通り、労働者の街が生んだ独特のモツ文化があるが、そのほかの地域は圧倒的に普段使いの肉は鶏だ。

図２　「お肉」といえば何の肉？

（日本経済新聞電子版「食べ物　新日本奇行」を基に作成）

私は、これに「環瀬戸内海T字型鶏肉回廊」と名付けた。

またもや名付けたからって何なのだという話だが、その奥には江戸時代の交易路、つまり自動車も鉄道も飛行機もなかった時代の営み、さらに近代に入って工業化した日本のたどってくるのだ。

グルメ、ご当地グルメと侮ることなかれ。つぶさにみていけば、正史には記されることのないリアルな生活史が浮かび上がってくるのである。

肉じゃがに入れる肉と言ったら？

ここまで外食を中心にみてきたが、牛・豚・鶏の地域差は家庭料理でも顕著だ。

わかりやすいのはカレーや肉じゃがといった煮込み料理に使用する肉である。

家庭、店舗を問わず関東で使われるカレーの肉はデフォルトが豚肉だが、関西では牛肉が使われる。関東ではビーフカレーもメニューにあることが多いものの、関西ではポークカレーはほぼ見かけない。

さらに、肉うどんとなると関西で豚を使うことはほぼない。関東ではその逆。牛を使

図４　環瀬戸内海Ｔ字型鶏肉回廊

はり海のすぐそばの街だろう。鮮魚店や地場のスーパーに行けば、大規模流通に乗らない珍しい地魚が並んでいる。それに比べると、ぐるりと海に囲まれている海洋国家ながら、大都市や山間部では手に入る魚の種類がいまだ限られるのは仕方ないことだ。

まず、魚介類全体の年間購入金額ランキングベスト10をみてみよう。

魚介類購入金額ランキング上位10（単位：円）

一位	富山市	9万5512
二位	青森市	9万4572
三位	仙台市	9万1432
四位	静岡市	8万9273
五位	北九州市	8万9140
六位	新潟市	8万8980
七位	盛岡市	8万8100
八位	京都市	8万7735

九位	秋田市	8万7539
十位	奈良市	8万7146

一位から六位まではわかりやすい。どこも市内または近くによい漁港がある土地だ。だが、七位の盛岡市、八位の京都市と十位の奈良市はともに海なし市である。

この謎はあとでふれることにして、まず仮説を立てよう。

日本全国の食が均一化されているとしたら、どの魚介も全体のグラフと同じよ
上位にあがってくるはずだ。
品目別にみていくことにする。

日本で最もポピュラーな魚といえばマグロとタイといったところだろうか。それぞれ赤身と白身の王様であり、寿司（すし）ネタにも欠かせない。

では、マグロからみてみよう。

マグロ購入金額ランキング上位10（単位：円）

一位	静岡市	1万4090
二位	甲府市	9552
三位	横浜市	8882
四位	宇都宮市	8816
五位	東京都区部	8655
六位	相模原市	8241
七位	川崎市	8127
八位	前橋市	7957
九位	さいたま市	7730
十位	浜松市	7601

おっと、いきなりの番狂わせである。

総合ランキングで十位に入ってきていた土地のうち、マグロ購入の上位十位に入って

きているのは静岡だけだ。

静岡には日本一の冷凍マグロの水揚げ港である清水港がある。よいマグロは築地に運ばれていくものの、地元でも安い冷凍マグロが流通している。だから堂々の一位は納得できるものの、甲府市、宇都宮市、相模原市、前橋市、さいたま市はいずれも関東の海なし市。しかも関東および静岡県以外の市が見当たらない。

この数字を見る限り、海辺の魚食いの地域ではマグロはさほど食べられていない、ということになる。

では第二位につけていた青森市は、県内に大間町というマグロの産地を抱えながらランキングに入ってきていない。それどころかマグロに限ればまん中あたりなのだ。

逆に第二位の甲府市は魚介全体で二十九位。つまり、日本の中ではどちらかというと魚の消費が少ない方の地域に入る。

では、なぜそんな場所がマグロに限っては上位に入ってくるのか。

実は、甲府の場合、食べるマグロは種類が決まっている。キハダマグロだ。

キハダマグロは、マグロ類全体の中で最も漁獲量が多い種だ。マグロの中では脂身が

少ないので、身の劣化が遅く、内陸輸送に耐えられるのが特徴だ。

つまり、今ほど冷蔵／冷凍技術が発達していなかった時代にも、刺身用として内陸に運ぶことができる貴重な魚だったのである。山間地の人々は、海のものへの憧れが非常に強い。だから、キハダマグロを昔から食べてきた。その名残で、今でも甲府の居酒屋でマグロぶつを頼むと、ほぼ百パーセント、キハダマグロが出てくる。甲府の魚市場で扱っているマグロのほとんどは、キハダマグロだと市場の人も言っていた。

内陸部もほぼ同じ事情だと考えていいだろう。

昔前まで、山間の温泉地などで供される会席料理ではだいたい冷凍マグロとイカの刺身が出てきたものだった。こっちにしてみれば、なぜ山に来てわざわざたいしてうまくもないマグロやイカを食べさせられるのだ、という気がしたものだが、かの地の人々にしてみれば、それが一番のご馳走であり、おもてなしだったのだろう。

二十一世紀になった今ではもう見かけなくなった、懐かしい昭和の光景である。

さて、念のため下から十位をみてみよう。

マグロ購入金額ランキング下位10（単位：円）

一位	北九州市	1073
二位	長崎市	1225
三位	松江市	1362
四位	大分市	1529
五位	福岡市	1657
六位	鳥取市	1690
七位	岡山市	1897
八位	佐賀市	1963
九位	山口市	1979
十位	高松市	1996

なんと魚介食い地帯として第五位に入っていた北九州市がぶっちぎりで「マグロを食べない」地域ナンバーワンだったのだ。その額も一位の静岡市に比べ、十分の一以下だ。

六位　盛岡市　432
七位　甲府市　450
八位　青森市　453
九位　浜松市　453
十位　福島市　503

見事。見事な西高東低だ。そして、完全にマグロと反比例している。

要するに、もともとマグロが穫れないうえに、先ほどの佐賀県の話ではないが、良質白身魚をはじめ、アジやサバなどが豊富に獲れる九州から中国、近畿地方べる必要がなかったのだ。

寿司屋でマグロを二回注文したら、「白身ば食べんね」と叱られた経験がある。江戸前の寿司はマグロが命といったりするが、九州では寿司ネタとしても白身魚がやっぱりマグロより上位なのだ。現実に、白身魚が圧倒的においしい。

青魚代表、アジ・サバ購入金額トップはどこ?

また関アジ、関サバではないが、アジやサバの脂のりは九州のものが優れている。

アジ購入金額ランキング上位10（単位：円）

一位　長崎市　3602
二位　山口市　2922
三位　佐賀市　2811
四位　松江市　2796
五位　大分市　2503
六位　宮崎市　2351
七位　北九州市　2249
八位　富山市　2064
九位　広島市　1943
十位　熊本市　1865

サバ購入金額ランキング上位10（単位：円）

一位	和歌山市	1715
二位	鹿児島市	1653
三位	北九州市	1646
四位	浜松市	1451
五位	松江市	1432
六位	福岡市	1402
七位	宮崎市	1400
八位	広島市	1365
九位	山口市	1353
十位	長崎市	1330

やはりほとんどを西日本の都市が占めている。青魚文化は西にあり、なのだ。

このように日本がマグロ一辺倒の国でないことは数字でみても明らかだ。たまたま関東圏がマグロ好きというだけで、日本人＝マグロというイメージが過大に刷り込まれているのかもしれない。万事につけ中央の文化が普遍的なわけではないことを、マグロはるのである。

米をよく食べるナンバーワンは？

次は主食をみてみよう。

日本人の主食は米と相場は決まっているが、その消費量の落ち込みはずっと続いている。一方で小麦の消費量はこの十年ほどはさほど変化していない。

総じて、日本人は主食より副食をよく食べるようになったことを示していると思われる。

では、米とパンを比べたとき、消費傾向に何らかの差異があるのだろうか。

まずは米の購入量上位十位と下位十位から。

圧倒的に食パンを食べている市は？

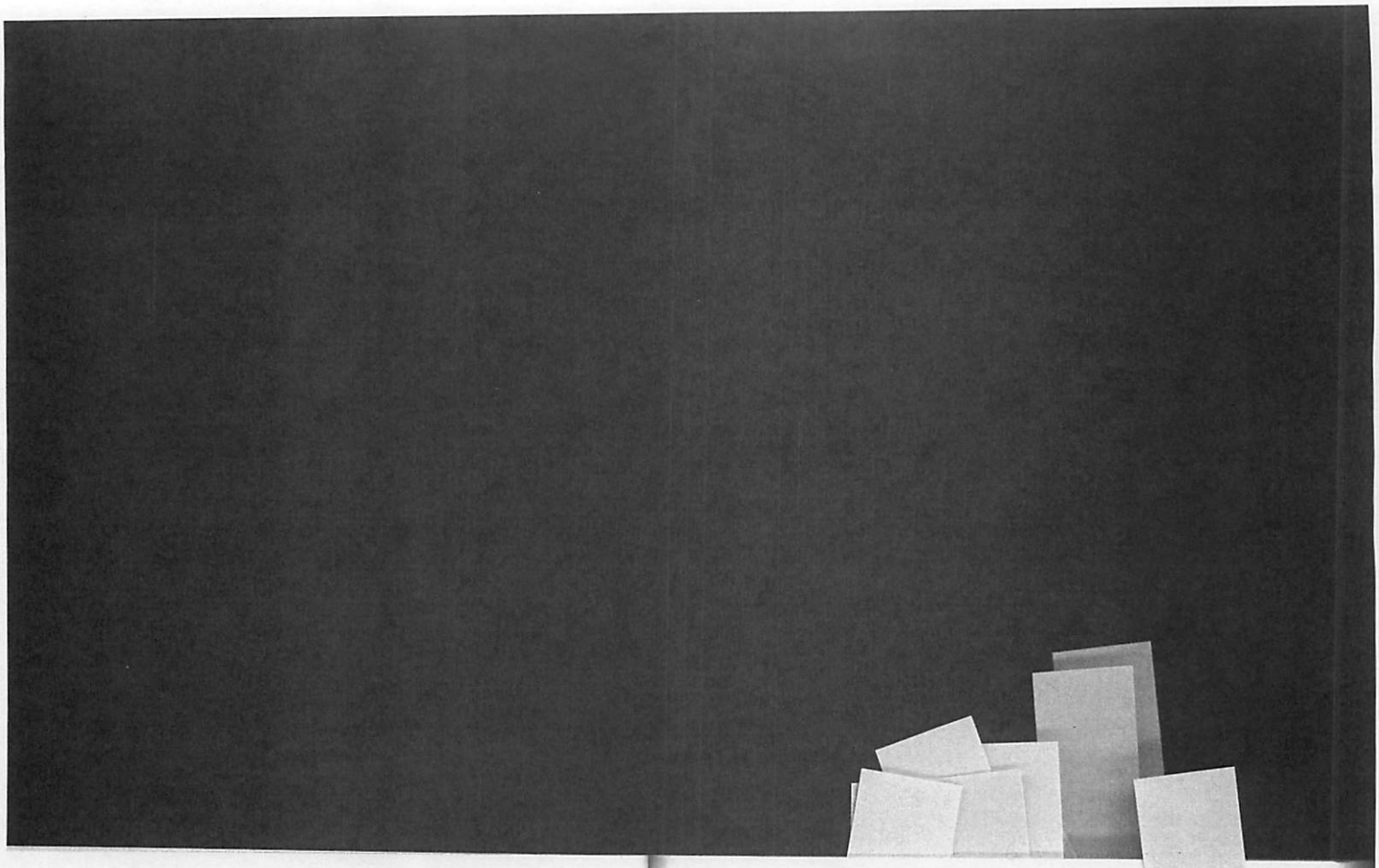
食パンの購入金額ランキング上位10（単位：円）

一位	神戸市	1万3405
二位	奈良市	1万1379
三位	堺市	1万1337
四位	高知市	1万0836
五位	金沢市	1万0787
六位	大阪市	1万0579
七位	松江市	1万0546
八位	京都市	1万0535
九位	鳥取市	1万0445
十位	横浜市	1万0419

食パンの購入金額ランキング下位10（単位：円）

一位	秋田市	5259
二位	熊本市	6347
三位	山形市	6565
四位	鹿児島市	6664
五位	福島市	6807
六位	盛岡市	7043
七位	青森市	7091
八位	前橋市	7092
九位	宮崎市	7142
十位	仙台市	7346

こちらはおもしろいデータが出てきた。



ぐいたのだろう。ちなみに、トンカツソースが生まれたのは戦後の神戸。道具所（現在はオリバーソース株式会社）が発売した。関東で使われる中濃ソースは一九三二年の発売なので、相当後発なのだ。

こうした事情もあってか、関西では今もソースを使う料理が多い。お好み焼きしかり、たこ焼きしかり、串カツしかり、である。余談だが、串カツのソースはウスターソースを昆布だしで割ったものであり、こんなところにも大阪人の昆布だし好きが表れている。そういうわけなので、当然ソースの購入量ナンバーワン市は大阪市と思ったら間違いで、結果はこう。

ソース購入量ランキング上位10（単位：ミリリットル）

一位	岡山市	2739
二位	広島市	2675
三位	徳島市	2457
四位	神戸市	2092
五位	大阪市	2091
六位	奈良市	1994
七位	堺市	1974
八位	高松市	1899
九位	京都市	1850
十位	松山市	1774

ソースも銘柄ごとに価格が違うので量でみてみたが、なんと上位三位を瀬戸内圏の岡山、広島、徳島が占めてしまった。実は、金額でみても広島と岡山の順位が入れ替わるだけで、この三市ががっちりとベスト3を守っている。ソース発祥の地である神戸市と大阪市はその後塵を拝しているのだ。

では、ここから何がみえてくるか。

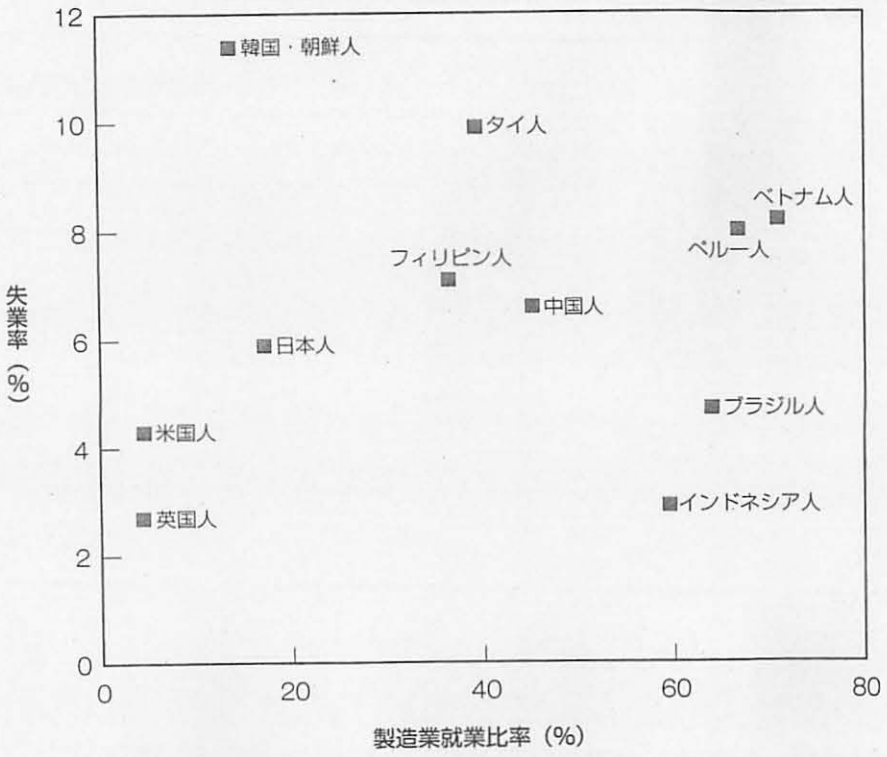
まず考えられるのが、広島を中心とした瀬戸内お好み焼き文化圏の存在である。広島にはオタフクソースという地元の非常に強力なソース会社がある。そして、お好

3-1 工場で真面目に働く外国人、サービス業を転々とする外国人
～在日外国人の失業率と製造業比率～

日本に住む外国人（外国人登録人数）は２００９年末に２１９万人に及んでいます。日本に住む外国人の国籍としては、従来は、韓国・朝鮮人が特別永住者（戦前、日本国籍とされた外国人とその子孫の在留資格）を中心にほとんどを占めていましたが、近年とともに減少を続けています。他方、中国人、ブラジル人、フィリピン人、ペルー人が１９９１年以降の18年間で２・２～４・０倍と大きく増加しています。増加人数では中国人が同期間に50・９万人増と全体の増加数96・７万人の半分以上を占めており特に目立っています。２００７年末以降にはついに中国人が韓国・朝鮮人を上回りました。

韓国・朝鮮人でも特別永住者以外は増加しています。韓国・朝鮮人特別永住者は１９９６年末の55万人から２００９年末の41万人へと14万人の減ですが、特別永住者以外は同時期に11万人から17万人へと６万人の増です。

図3-1　国籍別に見た製造業比率と失業率の相関（2005年）



（資料）総務省、「国勢調査」

3-2 お国なまりが好きな県民、恥ずかしがる県民

～土地の言葉への愛着度と恥ずかしさ～

方言

「お国言葉」、「○○弁」などと呼ばれる各地方の特色をもった

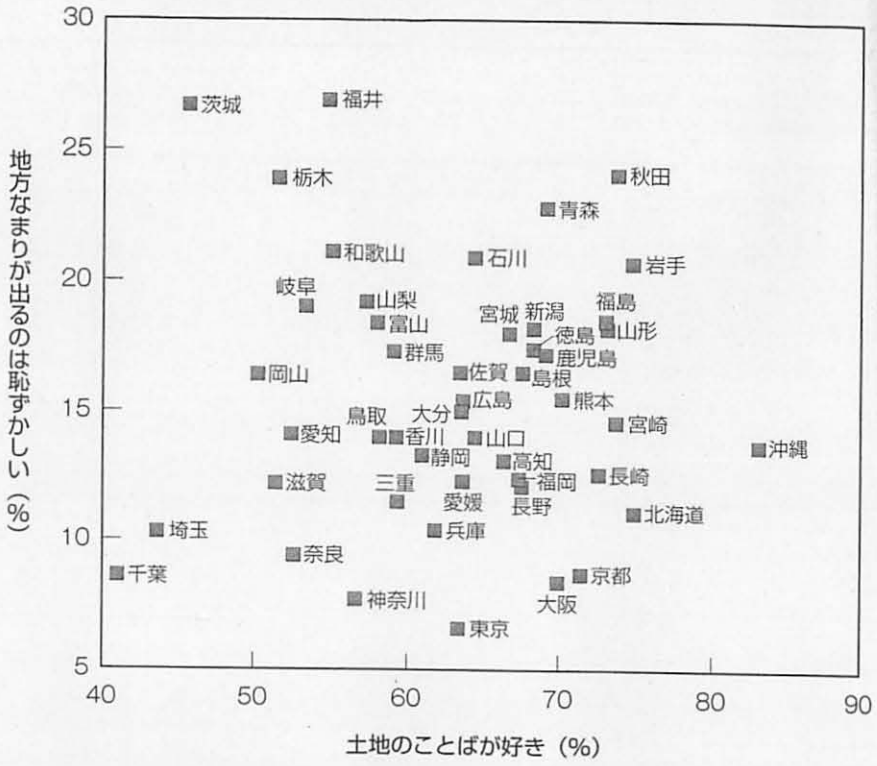
います。ここで取り上げた散布図では、各地域住民にとって、方言に対す

方はさまざまである点を示しています。

明治維新以降の国民統合へ向かう動きのなかで、東京方言が標準語とされたため、長い間、それ以外の方言に対して「遅れたもの」、あるいは「国内のコミュニケーションを阻害するもの」としてマイナスのイメージがつきまとうこととなりました。最近では、学校教育やテレビの普及などで標準語の使用や理解が容易となったため、むしろ、多文化共生の考え方に沿って、我が国の文化をより豊かなものにする方言の役割が見直されています。

図3-2には、ＮＨＫの全国県民意識調査（１９９６年）の結果を使って、方言に対す

図3-2　方言に対する感じ方（都道府県比較）



（注）各県16歳以上900人を対象とした1996年の個人面接調査による（回答率全国平均70%）

（資料）NHK放送文化研究所、「現代の県民気質―全国県民意識調査―」

3-3 あなたは醤油派か、ソース派か？

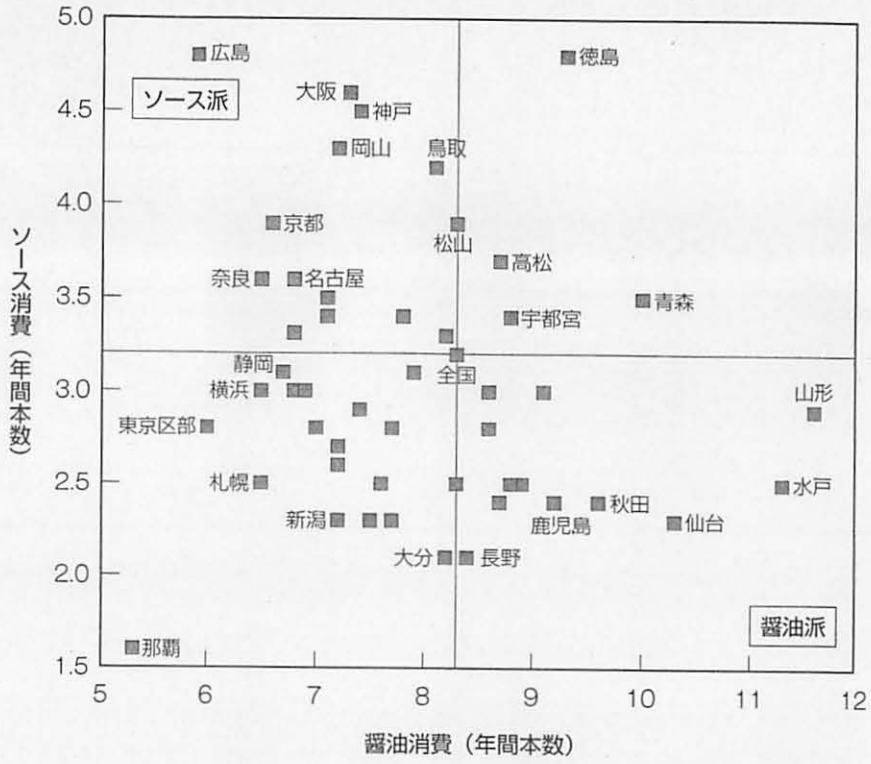
～都道府県別の醤油とソースの消費量～

目玉焼きにかける調味料としては塩のみ、塩・コショーの他、醤油派とソース派に分かれます。トンカツにかけるのはソースが通常ですが醤油をかける者もいます。カレーにかける者もいます。日本のソースは、英国のウスターソースをもとに明治時代に調味料として開発され、洋風化にともない全国に普及したものです。

総務省統計局が実施している家計調査により、県庁所在都市別の醤油の世帯消費量とソースの世帯消費量をグラフにしました（図3-3）。家計調査は、戦前の社会政策的な都市労働者の生計費調査から出発したという経緯から、大都市から全国に調査対象が広がった今でも、最も細かい地域区分は都道府県ではなく都道府県の県庁所在都市となっています。

醤油の消費量が多い都市トップ5は、山形市、水戸市、仙台市、青森

図3-3　醤油消費とソース消費の相関（都道府県庁都市）

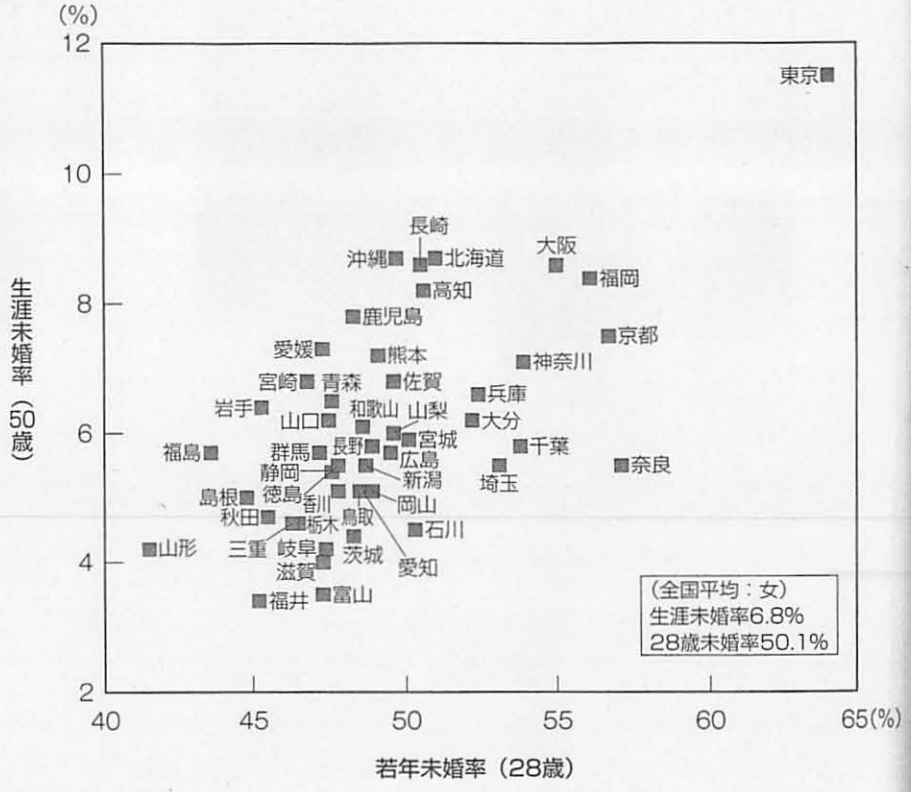


（注）世帯当たりの年間購入量。都道府県庁所在市および川崎市・北九州市（49市）の二人以上の世帯が対象。醤油は1000mlペットボトル換算、ソースは500mlペットボトル換算。

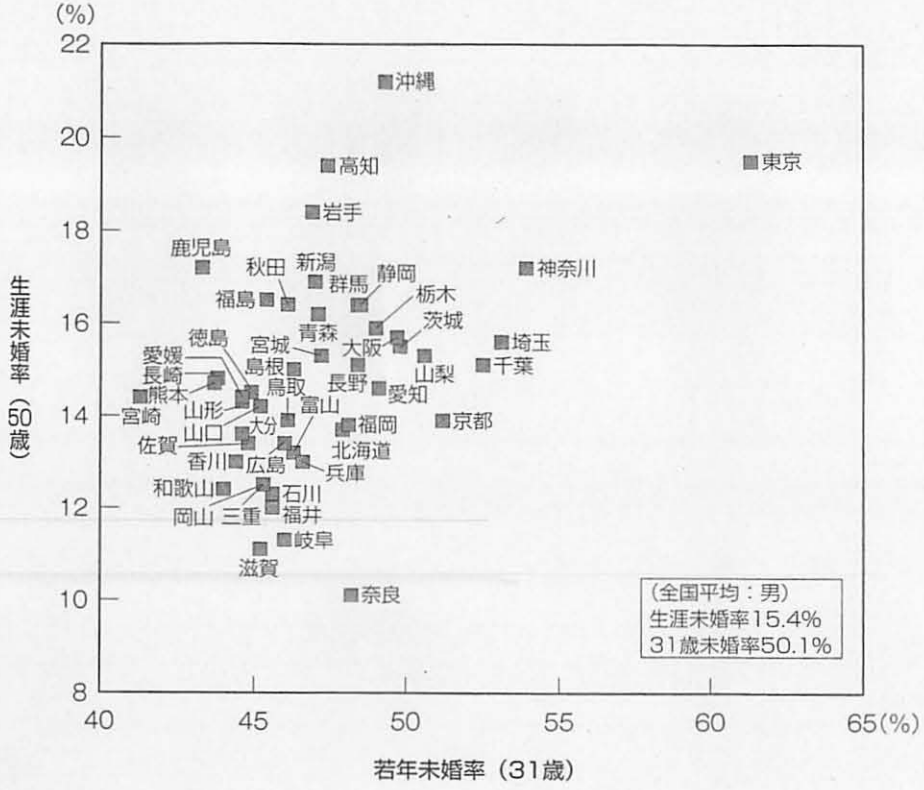
（資料）総務省、「家計調査」

図4-1　都道府県別の未婚率（2005年）

■女性



■男性



（注）直線回帰におけるR^2値は、男は、0.1861、女は、0.4117
（資料）総務省、「国勢調査」

会保障制度全体の設計や財源確保の見通しなく着手したので、２０１０年の参議院選挙においては野党自民党から批判されるに至っています。「子ども手当の財源のほとんどは国の借金だ。子どもたちが将来大きくなって、利息を付けて返さなければならない。いわば長期的な児童虐待だ。」（自民党の河野太郎幹事長代理、東京・東池袋の街頭演説で）（東京新聞２０１０年７月２日）

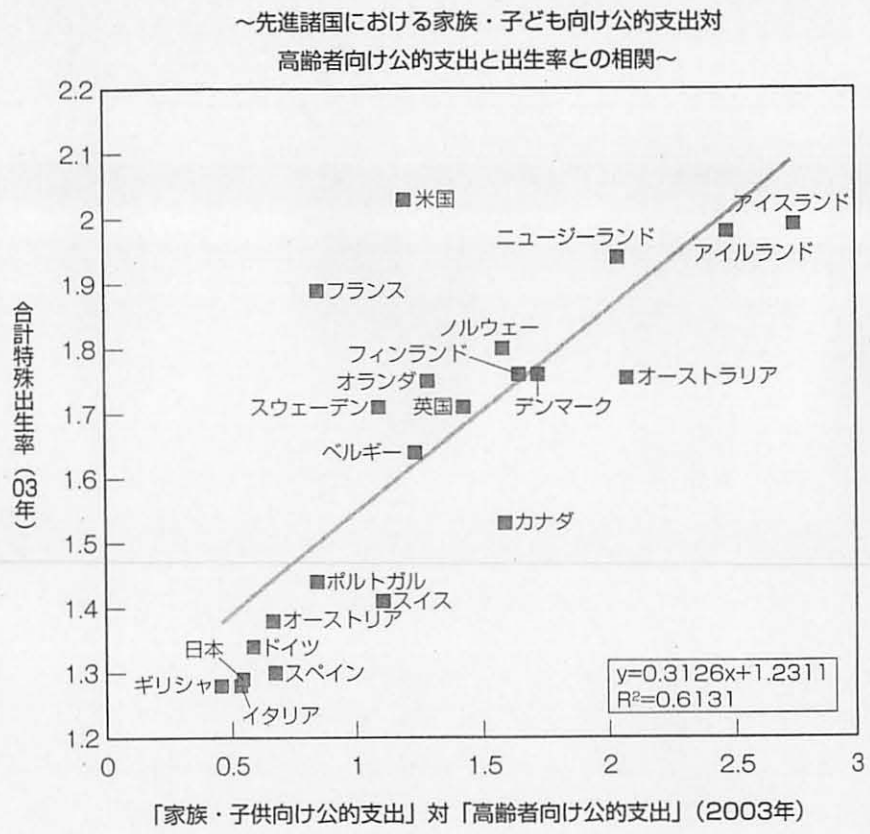
高齢化対策に対する少子化対策の相対ウェイトという指標には、高齢化対策に力を入れすぎて財政的に身動きがとれない程度が暗に含まれているとも考えられます。少子化対策の恩恵をうけても将来の増税を予想して素直には出生行動に国民が踏み切れないというこ

の話しの流れを汲んだ極めつけの相関図、図6-1-3に移ります。子育ての経費は、子どもが小さなうちばかりではありません。むしろ、高校、大学に通わせる経費の方が深刻です。となると、これまでの少子化対策支出で充分ではないので、教育費の公的支出も加えて、合計特殊出生率との相関を調べてみようということになります。

こうして描いた相関図を見ると、相関度はさらに上昇しました（表6-1）。日本がギリシャ、イタリア、スペイン、ドイツなどと並んで出生率が低いのは、高齢化

図6-1-3　高齢化対策に対する少子化対策の相対ウェイトと出生率

（少子化対策に教育費公的負担を含む）



（注）対象は世銀定義によるOECD高所得国（韓国を除く）。公的支出は社会保険や税の支出。家族・子供向け公的支出には児童手当、出産手当、産休給付金などの他、学校教育費の公的負担を含む。ルクセンブルクはデータなし。韓国はx軸値が3.99と異常に高いので除外した。

（資料）世界銀行、WDI／OECD, Social Expenditure Database 2007

7-1 身長の高い国・低い国から見える傾向

～男と女の平均身長～

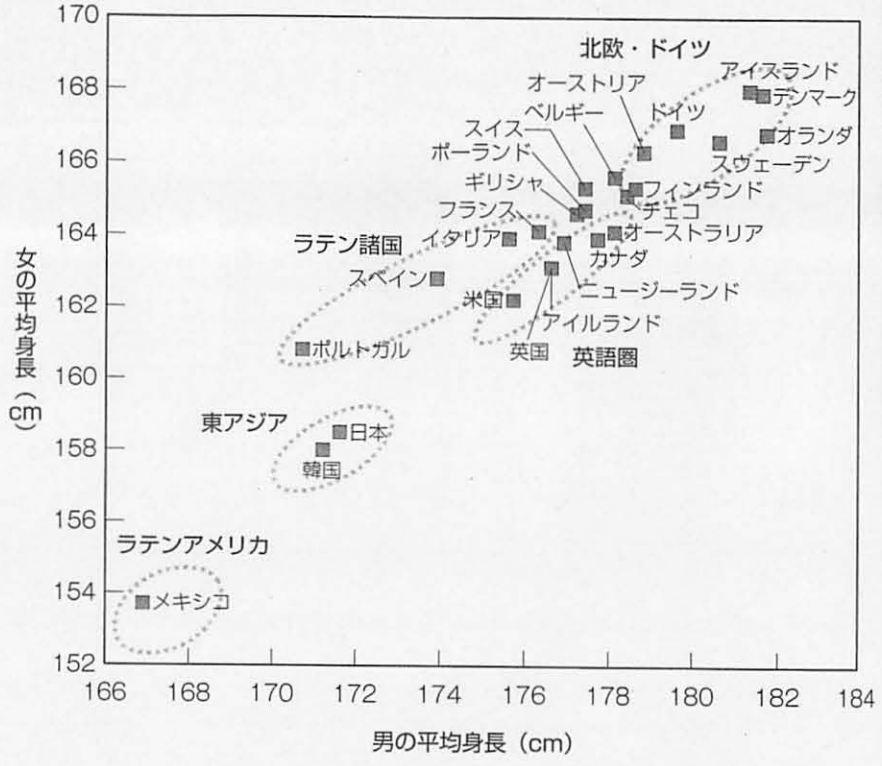
日本人が欧米人に比べ背が低いことは、経済成長に伴い栄養が改善されてかつてよりはずっと背が高くなった今でも、テレビで見る外国人、国内にいる外国人との比較で感じていることです。

図7-1はOECD諸国の男女の身長についての統計データをグラフ化したものです。男女はほぼ平行したパターンなので、男性についてみてみましょう。

日本人男性の平均身長は171・6cmであり、最も背の高いオランダ人男性181・7cmよりちょうど10cm低くなっています。オランダの他、デンマーク、アイスランド、スウェーデンでは男性の平均身長が180cmを越えています。

一方、日本人より身長が低いのは、OECD諸国の中では、韓国、ポルトガル、メキシコの3カ国のみです。この他、欧米の中ではポルトガルの他、スペイン、イタリアなどラ

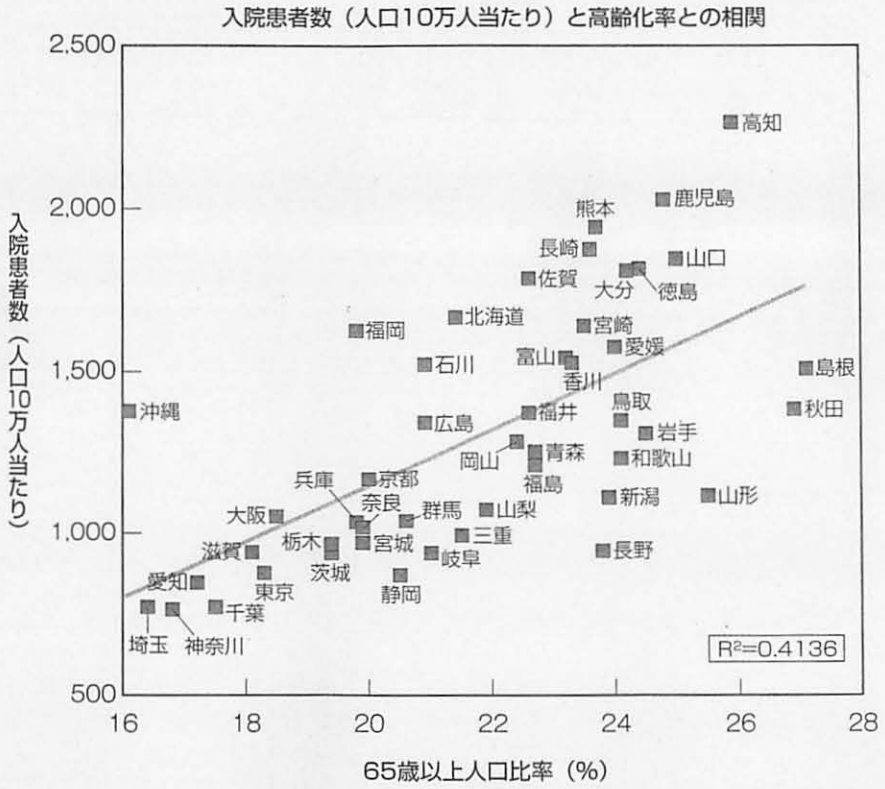
図7-1　世界各国の男女別平均身長



（注）国別ないし国際的な健康調査（2001～2007年）の結果から事務局でまとめた、おおむね20～49歳の成人身長平均（mean）値データ。男女両方のデータのないトルコとノルウェーを除く。

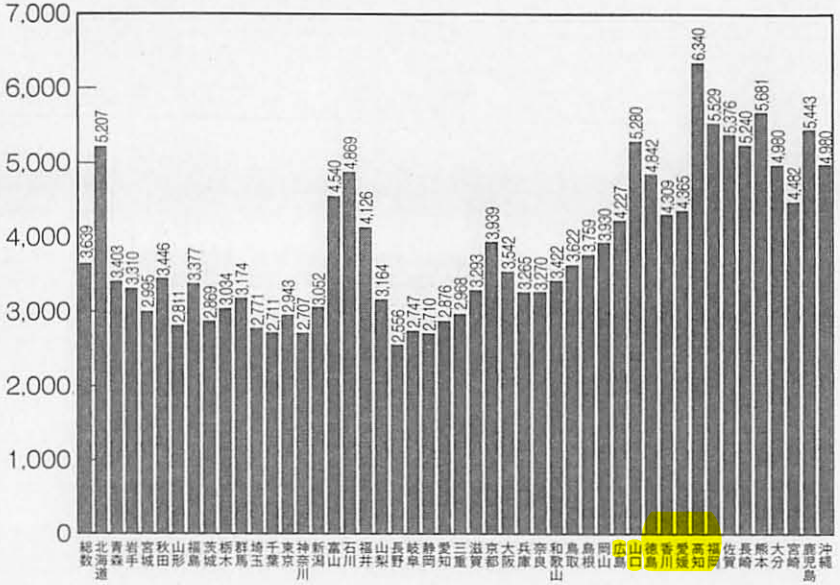
（資料）OECD, Society at a Glance 2009

図11-3-1　都道府県別の入院患者数と高齢化率



（注）2005年10月。都道府県別入院患者数は、患者の住所地別に算出したものである。
（資料）厚生労働省、「患者調査」

図11-3-2　65歳以上入院患者数（人口10万人当たり）



（注）（資料）図11-3-1と同じ

第2章

2 結婚してないカップルの子ども

～フランスでは半分以上、日本では2%～

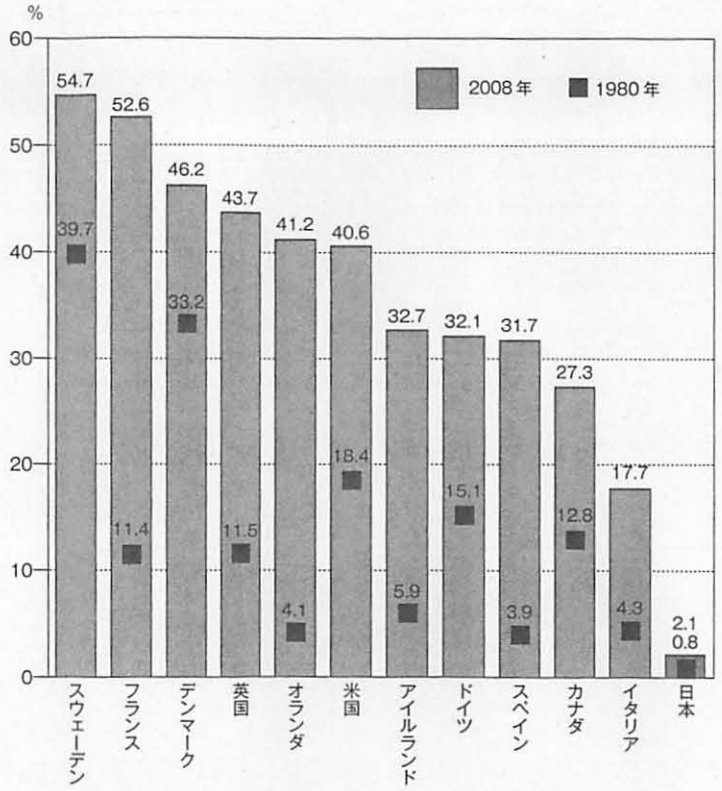
出生率を回復させた国々における出生率回復の要因のひとつとして、結婚しないまま子供を産むことが社会的に認知されている点があげられることが多くなっています。

そこでここでは、婚外子（こんがいし）、ないし非嫡出子（ひちゃくしゅつし）と呼ばれる結婚していない母（未婚の母、離別・死別後再婚していない母）からの出生の割合の各国ランキングを掲げました。

一目瞭然、最も目立っているのは日本の婚外子割合の低さです。一方、スウェーデンが54・7%と5割以上で最も高いのが目立っており、次ぎにフランス、デンマークがそれぞれ52・6%、46・2%で続いています。図には1980年当時の婚外子比率の高さを合わせて示しています。北欧のスウェーデンやデンマークは1980年でも3割を越えており、かなり前から高かったことが分かります。

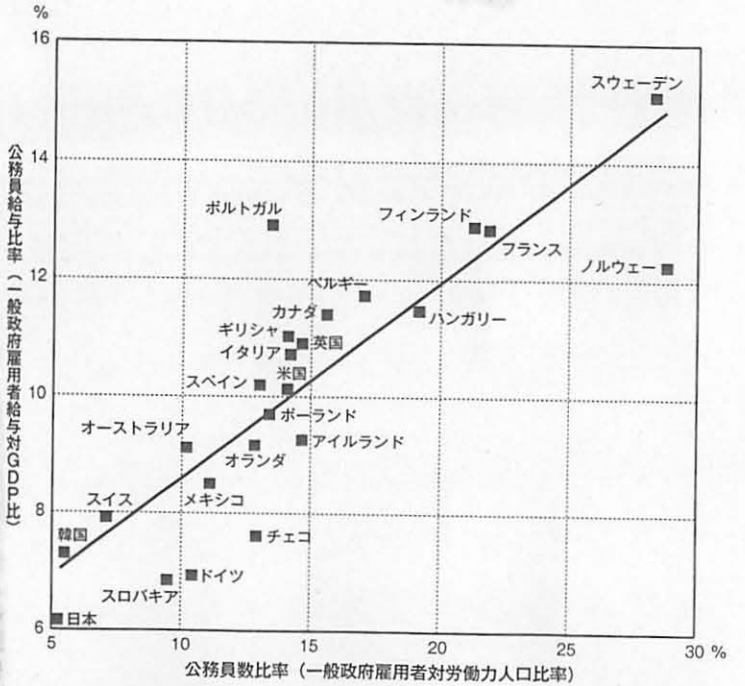
欧米の中でもスペイン、イタリアといった典型的なカトリック国では相対的に婚外子の割合が低くなっています。またフランスやアイルランドといったその他のカトリック国、あるいはオランダ、英国といった国も1980年段階では低かったのですが、その後、大きく上昇しているのが目立っています。

図2-2 世界各国の婚外子の割合



（注）未婚の母など結婚していない母親からの出生数が全出生数に占める割合をあらわす。ドイツの1980年は1991年のデータである。2008年について英国、アイルランドは2006年、カナダ、イタリアは2007年のデータである。
（資料）米国商務省、Statistical Abstract of the United States 2011
日本：厚生労働省「人口動態統計」

図2-9 OECD諸国の公務員給与水準



（注）公務員数比率は基本的に2005年データであり、公共分野の雇用比較（CEPD）調査（OECD 2006）によって収集されたデータに基づいている。公務員の総人数が基本であるが常勤換算の国（オーストリア、オランダ、スウェーデン、スイス、英国）では比較上は過小評価されていると考えねばならない。データはSNAの定義にもとづく一般政府をカバーしている。一般政府はすべてのレベルの政府（中央、州、地方、社会保障）からなり、省庁、独立部局、及び政府のコントロール下にある非営利組織を含んでいる。

公務員給与比率（一般政府雇用者給与対GDP比）は2007年データによる。ただし、メキシコは2004年、日本・韓国・スイスは2006年データ。ここでの給与には、政府による社会保障負担や任意的な手当等を含む。

（資料）OECD Government at a Glance 2009

なお、小さな政府を標榜する米国は、この図では、公務員比率、公務員給与比率ともに中位のレベルにあり、小さな政府とはいえません。これは地方公共団体、地方自治体の公務員が多く、その給与も決して低くはないためです。東京新聞は「役人天国アメリカ」という国際面の連載で、強力な組合を背景に地方公務員の給与が民間水準より高く、老後保障も手厚い場合が目立つこと、また自治独立の精神から小さな自治体が非常に多く（人口、面積とも日本より小規模なニューヨーク州の自治体の数が約３４００）、それだけコストは高いことを報じました（２０１０年12月25～27日）。

他方、一次近似直線より下方の国は、給与水準が比較的低い国と見られます。ノルウェー、チェコ、ドイツ、スロバキアといった国では、相対的に給与水準は低いことが分かります。日本についても、この直線より下であり、給与水準が高いとは言えません。

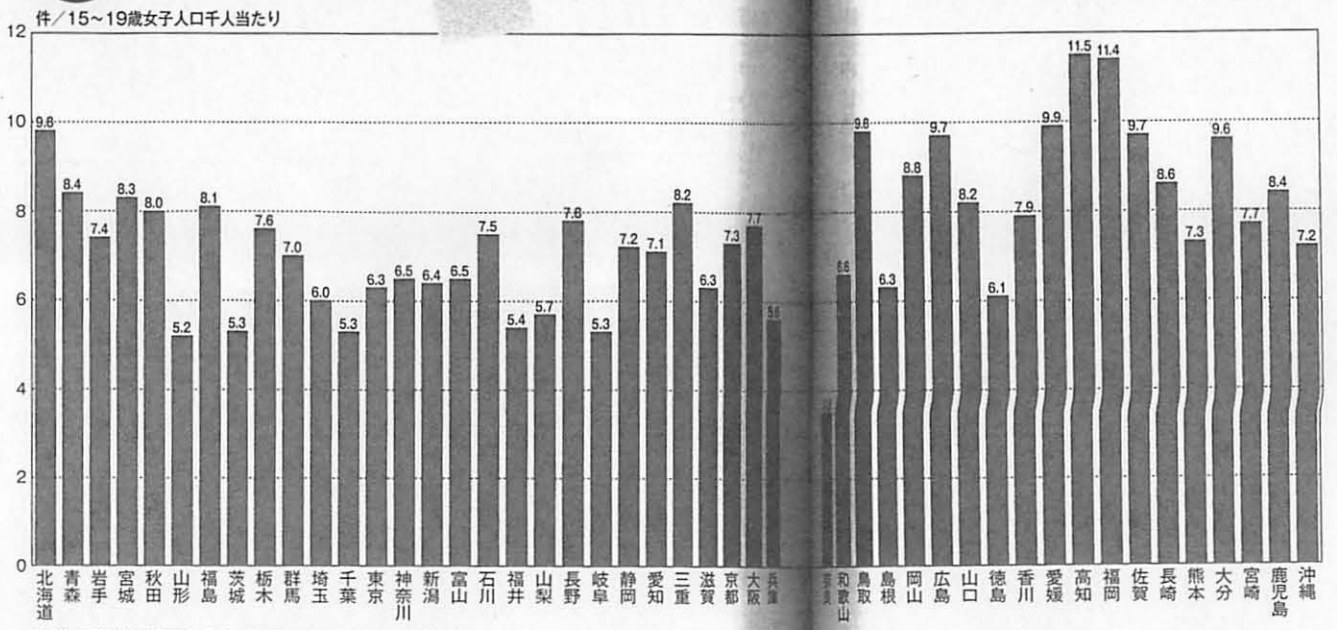
データから見ると、日本の公務員数は労働力人口との対比で最少なので、日本の政府サービスの範囲が他国並みの大きさであるとすると、日本の公務員は「少数精鋭」あるいは「政府サービス実施のための一人当たりの負荷が大きい」と考えることも可能ですが、だからといって以上のように給与水準が世界と比べて高いわけでもなさそうです。

この図は日本の公務員が公務員以外と比較して恵まれているかどうかを示したものではありません。日本の公務員が給与的に恵まれているとしたら、それでも、海外の公務員が恵まれている程度以上ではないことを示しているのです。

高いか、島根は低くて鳥取は高いかといった疑問に回答を与えるものではありません。

地方の方が大都市部より中絶率が高い理由としては、地方圏では、未婚男女にとって、性的な異性関係以外の刺激的な娯楽や気晴らしに乏しいからという点が指摘されることもありますが、未成年だけでなく女性全体の中絶率を見ても、同様の傾向が認められるので、夫婦関係が主となる成年にも同じ理由が当てはまるとするか、あるいは成年・未成年の違いとは関係ない地域的な精神風土や避妊の普及度、公衆衛生上の取り組みの違いなどに理由を求めるべきでしょう。

図3-5 未成年の人工妊娠中絶実施率
(2009年)

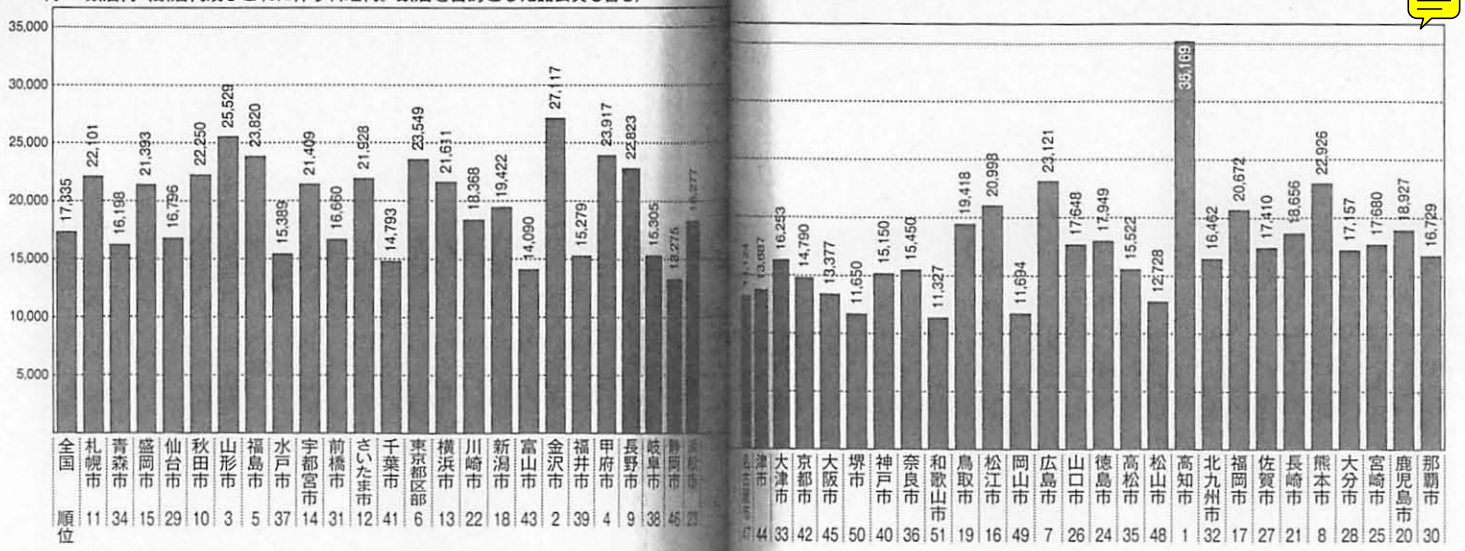


(資料) 厚生労働省「保健・衛生行政業務報告(衛生行政報告例)」

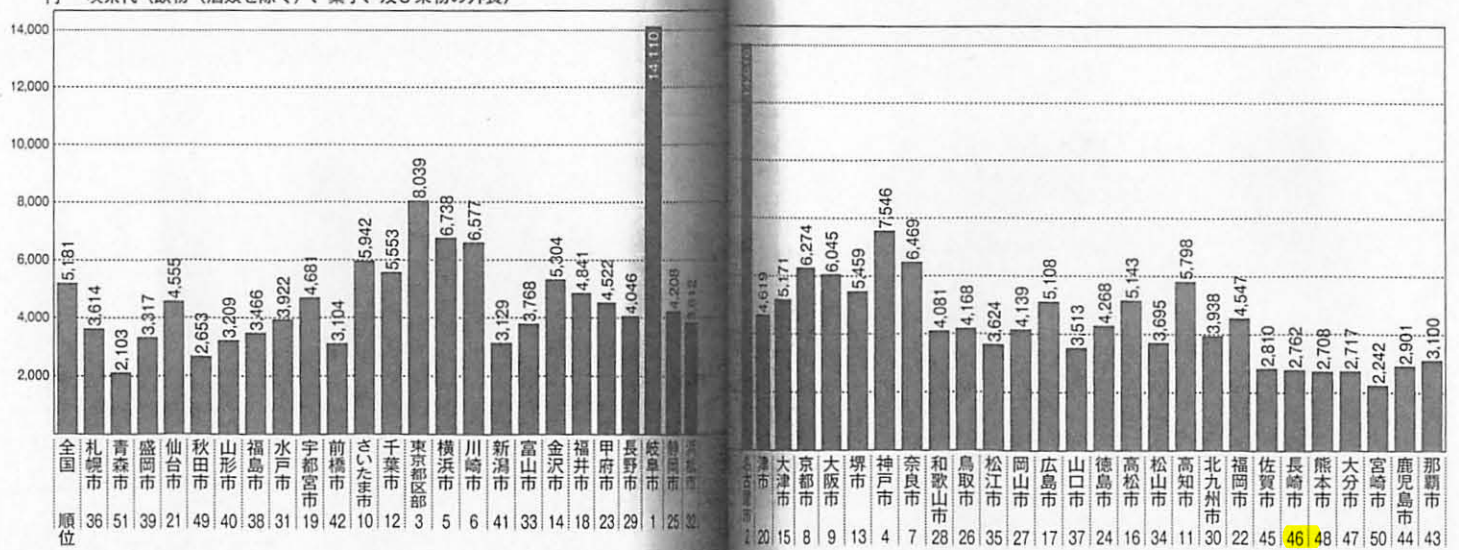
図3-6 県庁所在都市別の飲み屋代及び喫茶店代

（年間支出額、2008～10年平均）

円　飲酒代（飲酒代及びこれに伴う料理代。飲酒を目的とした諸会費も含む）

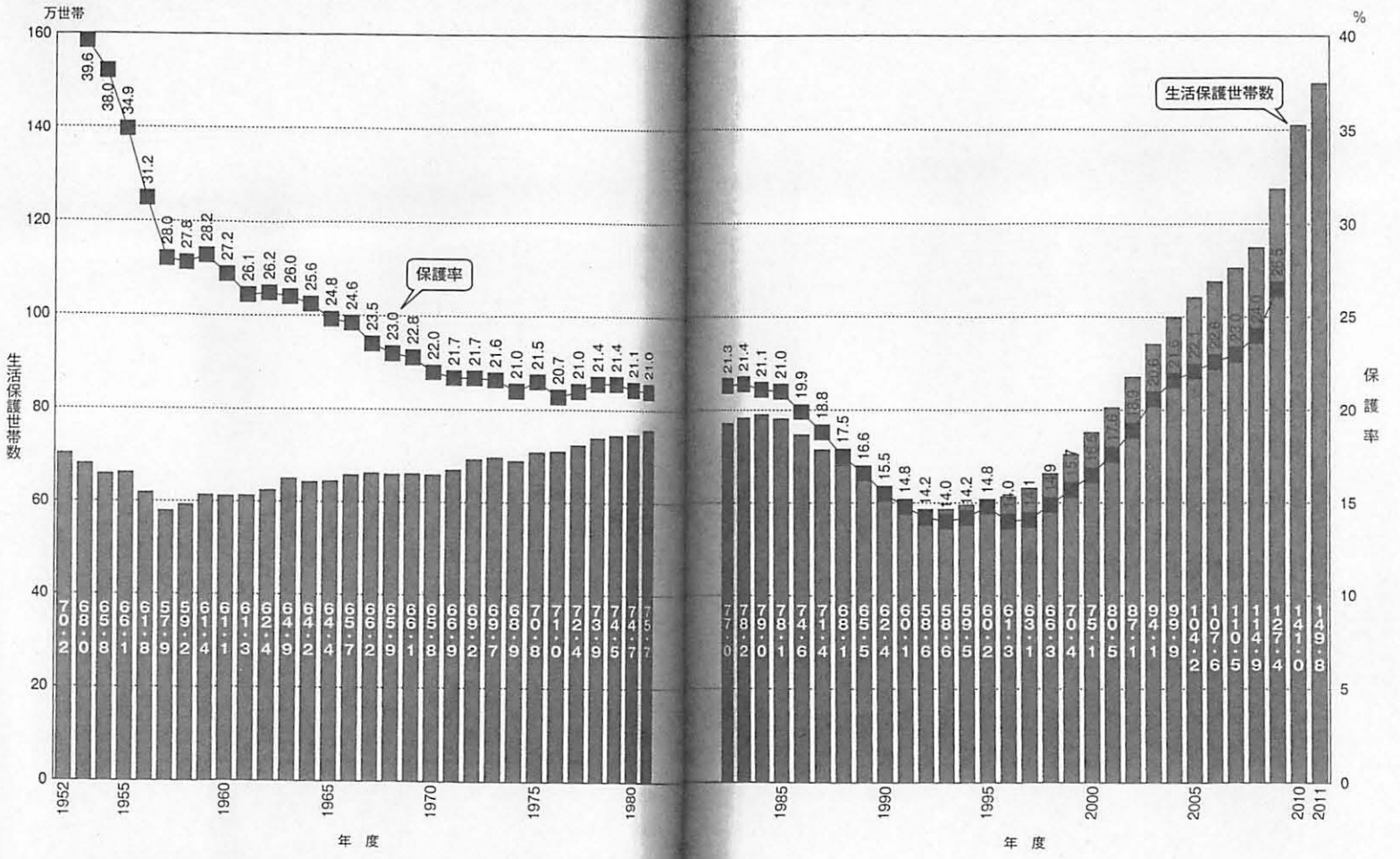


円　喫茶代（飲物（酒類を除く）、菓子、及び果物の外食）



（注）都道府県庁所在市・政令市（全51市）の二人以上の世帯が対象。（資料）総務省統計局「家計調査」

図5-5 生活保護世帯数と保護率の推移



(注) 年度の1か月平均(2011年度は概数)。保護率は社会保障・人口問題研究所「「生活保護」公的統計データー覧」。

(資料) 厚生労働省「社会福祉行政業務報告(福祉行政報告例)」

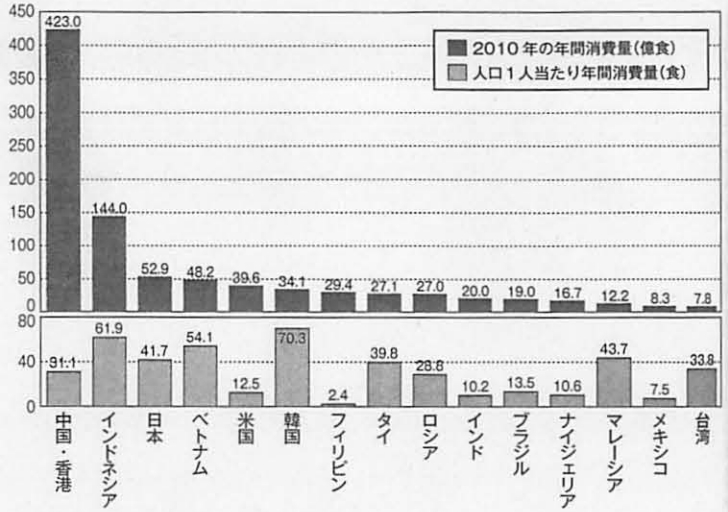
2 インスタントラーメン大国はどこか？
～全体量では中国、一人当たりでは韓国～

インスタントラーメンは日本の発明品です。１９５８年に日清食品の創業者安藤百福（ももふく）が大阪で「チキンラーメン」を開発し、その後、日本ばかりでなく世界に消費が拡がっていきました。日清食品はインスタントラーメンの世界的な普及のため、自らが会長となって世界ラーメン協会（ＷＩＮＡ）をつくっています。この協会が世界の主要消費国における即席麺（インスタントラーメン）の消費量をとりまとめているので、これをグラフにしました。当然、インスタントラーメンの消費は人口規模に左右されるので、人口１人当たりの消費量も算出してグラフに加えました。

２０１０年の集計では、９５４億食と過去最大となりました。２００８～09年には世界不況の影響でインスタントラーメンの消費も落ち込んだので世界ラーメン協会が期待した１千億食の予想はなお達成されていません。２００２年には５８７億食でしたから、消費量は世界的に大きく拡大して来ています。

消費量規模では、中国が４２３億食と世界全体の約半分ちかくを占め最も多く、次にインドネシアの１４４億食が続いています。日本は世界第3位の53億食です。一方、人口１

図7-2 インスタントラーメン（即席麺）消費量の国際比較



（注）世界全体で954億食、13.8食／１人当り。

（資料）世界ラーメン協会 HP：2011年5月10日現在（人口は FAO Online 2011.7.16）

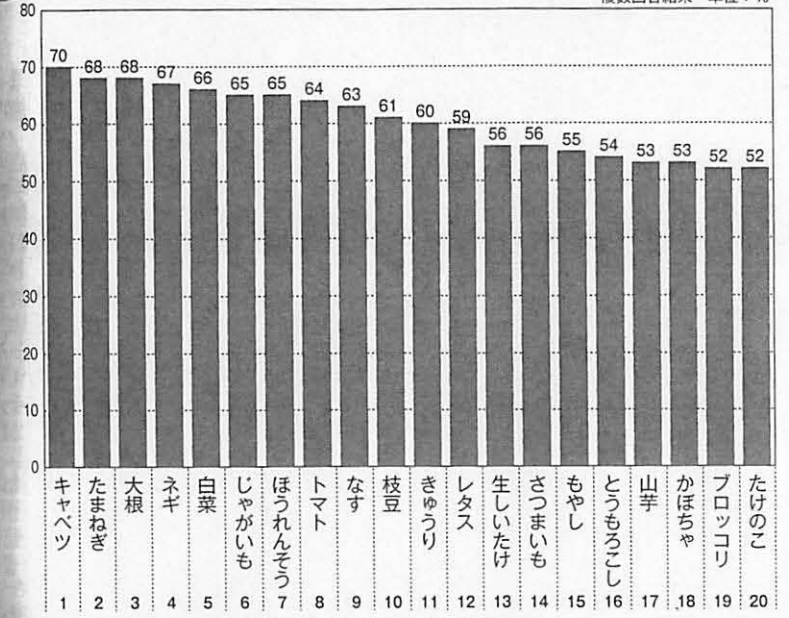
表7-1 世界のインスタントラーメン消費

国　名	特　　徴
中国	世界消費量の半分ちかくを占める。香辛料の効いたビーフ風味が中心
インドネシア	１人当たり消費は世界第２位（図参照）。汁なしのやきそばタイプが半分を占める
日本	カップ麺割合が60%以上と最も高い。しょうゆ、みそ、とんこつ、塩味基本
米国	スープはチキン味、シュリンプ味、ビーフ味が主流
ベトナム	近年消費量が急拡大。スープは酸っぱくて辛いシュリンプ味が第１位
韓国	１人当たり消費は世界最高（図参照）。日本での開発の５年後、1963年に韓国で生産開始。袋麺の割合が75%でビーフ味、キムチ味など辛口風味が好まれる
フィリピン	汁物袋麺は朝食としても食べられ、チキン味、牛骨味が人気
タイ	世界３大スープの一つ、酸っぱくて激辛のトムヤムクン味が中心
ブラジル	日本より柔らかい麺が好まれる。風味のよい地鶏味が第１位
台湾	カップ麺の割合が50%と高い。ポーク味が半分を占め、次いでビーフ味

（資料）東京新聞2008年4月11日ほか

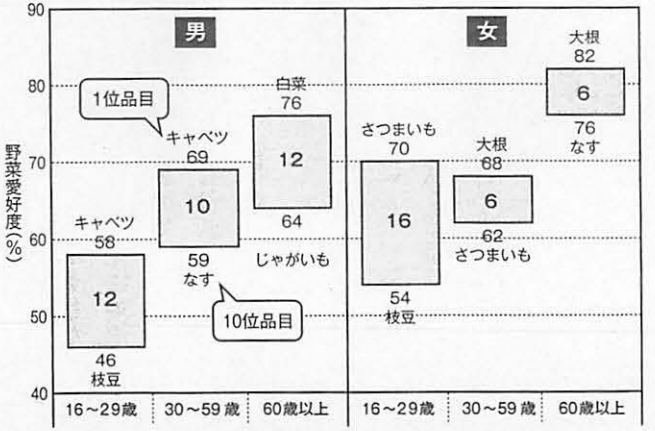
図8-1 日本人の好きな野菜ランキング
(2007年)

複数回答結果　単位：%



(資料) NHK放送文化研究所世論調査部「日本人の好きなもの」2008年

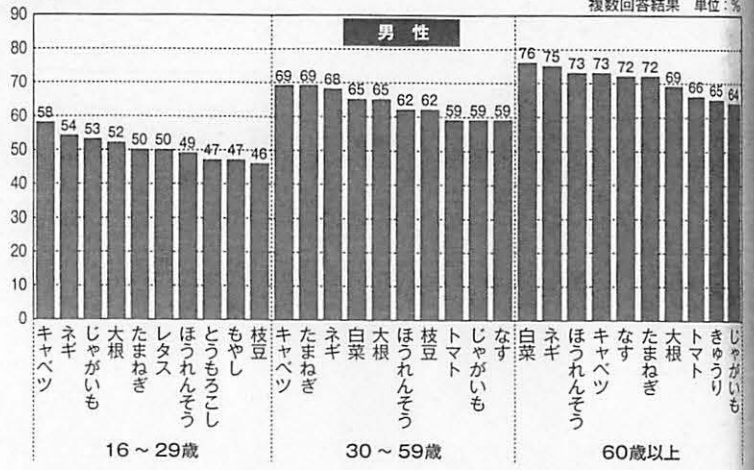
図8-3 野菜愛好度～1位品目と10位品目との差～



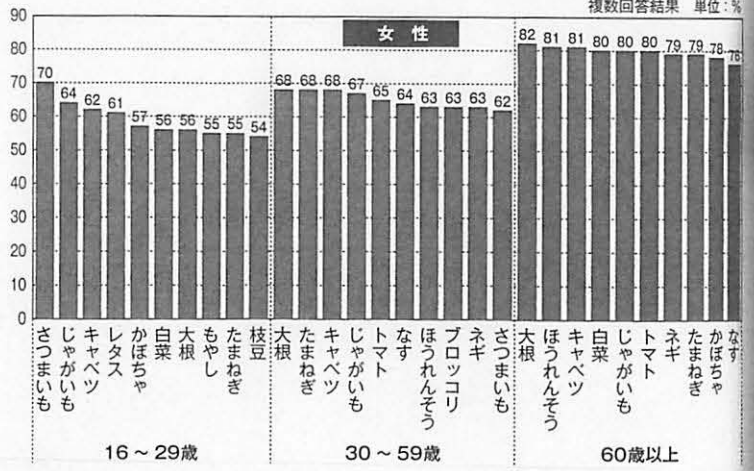
(資料) NHK放送文化研究所世論調査部「日本人の好きなもの」2008年

図8-2 日本人の好きな野菜ランキング・男女年齢別ベストテン
(2007年)

複数回答結果　単位：%



複数回答結果　単位：%



(資料) NHK放送文化研究所世論調査部「日本人の好きなもの」2008年

後に急速な経済成長を遂げた日本における生活に対する満足度は、低下している」という調査結果をもとに、「経済成長だけでは国民の幸せは測れない」という「イースタリンの逆説」を提唱しました。この説は、所得水準と生活満足度（well-being）はある時点の一国内ではゆるく相関しているが、時間を超えた2時点や地域を越えた2地点ではほとんど相関がないとする説です。

ところが、近年、世界各国で行われるようになった国際共同調査によれば、生活満足度でも幸福度でも豊かさと幸福度は認められるという研究成果があらわれ、海外の有力経済誌でも紹介されるようになっています。「これまで命題の証拠がないから命題は当てはまらないと混同されてきた」という訳です（"Comparing countries - The rich, the poor and Bulgaria", The Economist December 18th 2010）。

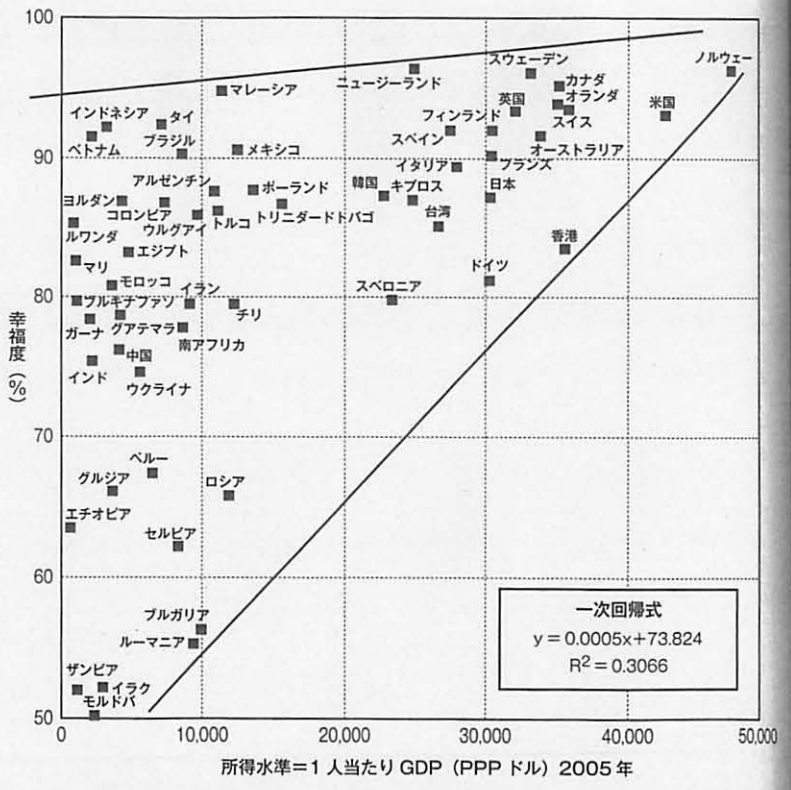
それでは、「幸せはお金で買えるか」ということを幸福度と所得水準の国別相関図で見てみましょう（図8-5）。

相関度をあらわすR^2値は0・3066であり、まあまあの相関が認められます。

しかし、相関図を見て、より印象的なのは、所得水準の高い国では幸福度がある一定水準以上に収斂（しゅうれん）している（不幸と感じている者はそれほど多くない傾向がある）のに対して、所得水準の低い国では、幸福度に大きなばらつきが認められる点です。

こうした相関パターンは「片相関」として理解できるように思います。前著「統計デー

図8-5 幸せはお金で買えるか
（所得水準と幸福度の国別相関）



（注）幸福度は「非常に幸せ」及び「やや幸せ」と回答した比率の計であり、各国の全国18歳以上の男女約1,000～2,000人を対象として実施された世界価値観調査（2005年前後）による。ここでは世界価値観調査実施国57カ国中、アンドラを除いて所得データが得られる56カ国を対象としている。

（資料）World Values Survey HP（2011年1月2日），IMF World Economic Outlook Database, September 2011

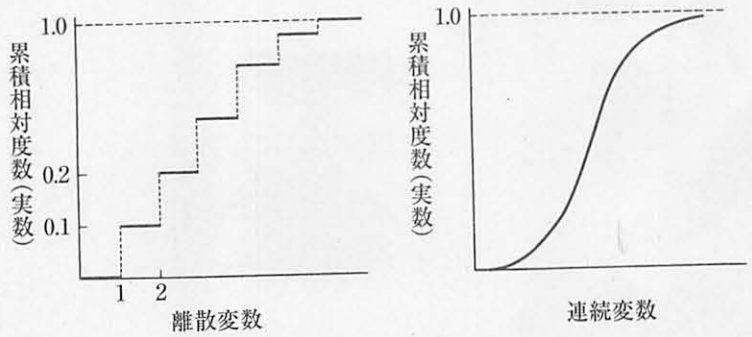


図 3·4　分布曲線

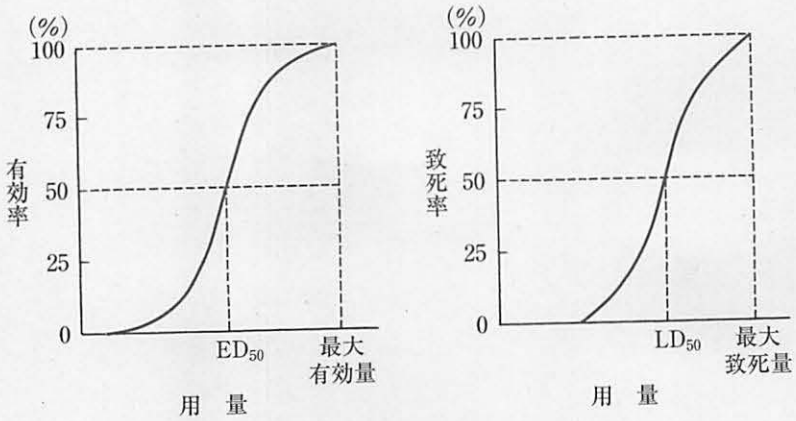


図 3·5　用量-反応曲線

にS字型の累積度数分布曲線を示す．このような曲線を用量-反応曲線という(**図 3·5**)．

観察集団の50%が有効に反応する量を50%有効量(ED_{50})といい，観察集団の50%が死亡する量を50%致死量(LD_{50})という．

度数分布や累積度数分布の他にも，データを積極的に視覚化し，その意味することや解析の方法を探っていこうとする立場から，箱ひげ図(**図 3·6**)，幹葉図(**図 3·7**)など様々な方法が提案されている．たとえば，次のような最高血圧値データに対しては，箱ひげ図(**図 3·8**)，幹葉図(**図 3·7**)のようになる．

134	128	108	124	124	128	132	102	160	136	128
130	114	124	154	114	126	132	136	130	122	

箱ひげの箱やひげの位値には様々な定義がある．この図では，箱の最下端，中央，最上端それぞれデータを小さい順に並べたときの25%，50%，75%を表している．上下のひげの長さについては一般的に箱の長さの1.5倍以内とする．ひげの上(下)端は，箱の長さの1.5倍以内にある最大(小)値である．ひげの端の外側にある値を外れ値とする．この1.5

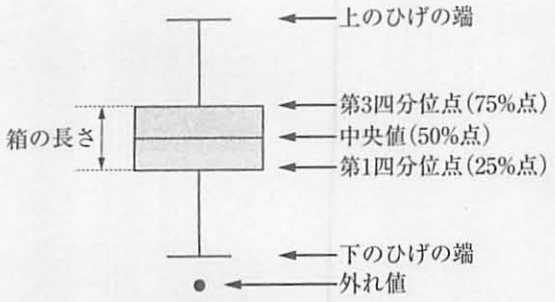


図 3·6　箱ひげ図の構造

幹	葉
160	0
150	4
140	
130	0022466
120	24446888
110	44
100	28

図 3·7　幹葉図

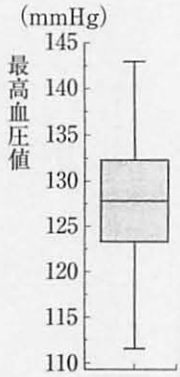


図 3·8　箱ひげ図

倍は絶対的なものではない．

幹葉図は，生データを幹と葉に対応する数に分けて分布を作成する方法で，ヒストグラムと違って生データを再現できる．

質的データにおいても，数量データにおいて度数分布表を作ったときと同様に，各ラベルに属する人数を度数とし，度数分布表を作成できる(**表 3·4**)．順序尺度では各ラベルに順序があるので累積度数分布表にも意味がある．

表 3·4　質的データにおける度数分布表

血液型	インフルエンザ罹患者数(人)
A	84
B	26
AB	16
O	74
計	200

❷ 関連性の表示

2つの変数からなる大きさ n の標本を考えたとき，身長(cm)と体重(kg)，薬剤投与(投与群，非投与群)と最低血圧値，薬剤投与(新薬群，偽薬群)と効果(効果あり，効果なし)のような，それぞれのタイプの2つの変数の組の関連を図示するにはどのようにしたら良いだろうか？ これらの2つの変数の組はデータの形に注目すると，(数量データ，数量データ)，(質的データ，数量データ)，(質的データ，質的データ)の3種類になる．

(1)（数量データ，数量データ)のデータの図示

座標平面 XY を考えて，各個体の X 変数，Y 変数を座標平面上の点 (X, Y) として表すと，座標平面上に n 個の点が図示される．この図を**散布図**という(**図 3·9**)．関連性がある場合は，X の値の変化に応じて Y の値が変化することから，関連性の傾向を示すことができる．

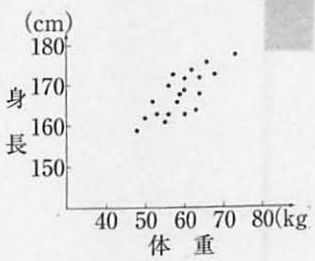


図 3·9　散布図

(2)（質的データ，数量データ)の図示

座標平面 XY を考える，質的データはラベルとして表されるので，X 軸上の任意の点にそのラベルを表す点をとる．ラベルの数だけその点を

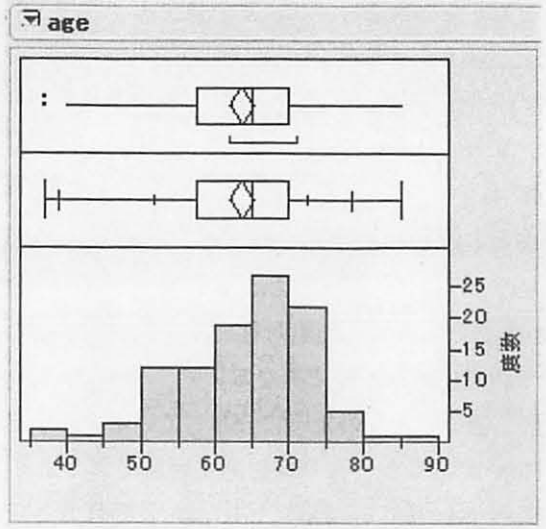


図 2.4　箱ひげ図。

タイル、最大値の値を示している。したがって、ひげは最小値および最大値まで伸びていて、範囲 range を示している。これら 2 つの表示は、いずれも、箱の部分は同じである（p.118 も参照のこと）。

§ 2.4　Microsoft Excel の関数

統計解析ソフトウェアを用いなくても、Microsoft Excel の関数で平均値などの指標を計算することができる。

(セルの範囲) は左上のセルと右下のセルを指定する。

例：`A1:A10`　　カラム A のセル A1 から A10 までの 10 個の値を指定。

`A1:C10`　　セル A1 から C10 までの 30 個の値を指定。

`A1:J1`　　セル A1 から J1 までの 10 個の値を指定。

平均値　＝ `average(セルの範囲)`

標準偏差　＝ `stdev(セルの範囲)`

中央値　＝ `median(セルの範囲)`

最小値　＝ `min(セルの範囲)`

最大値　＝ `max(セルの範囲)`

最頻値　＝ `mode(セルの範囲)`

パーセンタイル　＝ `percentile(範囲,x)` x は 25 パーセンタイルで[illegible]ば、0.25。その他同様。

§ 2.5　臨床医学研究の分類

医学研究には基礎医学研究と臨床医学研究に大別される。臨床医学研究はヒト、すなわち患者と健常者を対象あるいは被験者として研究が行なわれる。さまざまな**危険因子**（リスクファクター）と疾患あるいは病態といった**アウトカム**（転帰）との関係を明らかにする目的で行なわれる観察研究と、治療法などの介入とアウトカムの関係を明らかにする目的で行なわれる実験研究あるいは介入研究がある。

ヒトを対象にした研究は被験者の保護、倫理の遵守、といった点から、厳しい基準が設けられている。少なくともヘルシンキ宣言に従うことが要求される。特に、実験研究はヒトに、薬剤を投与したり、カテーテルを操作したり、放射線を照射したりといった介入を加えるので、ヘルシンキ宣言に則り、厳しい基準が設けられている。特に新薬など新しい治療法の開発には、新 GCP という基準が設けられている。そして、それぞれの、医療機関に設けられている倫理委員会あるいは治験審査委員会の審査を受けることが、必須になっている。

臨床医学研究は研究デザインによって、以下のごとく分類される。

ランダム化比較試験　Randomized controlled trial：介入 Intervention または曝露が、被験者個人のレベルでランダム割付が行われた臨床研究あるいは治験。治療法の有効性を証明するゴールドスタンダード

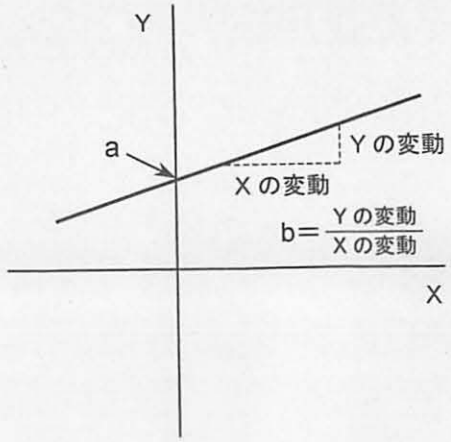


図 4.6　二次元平面における直線。

この場合、2 種類の変数のうちで最初に起きる事象のデータ、あるいは測定がより簡単なデータの方を**独立変数** independent variable または**説明変数** explanatory variable と呼ぶ。一方、独立変数または説明変数によって推測されうる変数は**従属変数** dependent variable または**応答変数** response variable と呼ぶ。

直線回帰 linear regression は、独立変数と従属変数は比例関係にあり、片方の数値 X が大きくなるともう片方 Y も大きくなる。独立変数 X と従属変数 Y の関係は次の式で表される。

$$Y = a + bX$$

二次元の平面でどのような直線でもこの式で表すことができる。b は直線の**傾き** slope を表し、X の値が 1 増加すると、Y がどれだけ増加するかを表している。a は**切片** intercept と呼ばれ、その直線が Y 軸と交差する点の Y の値に相当する (図 4.6)。

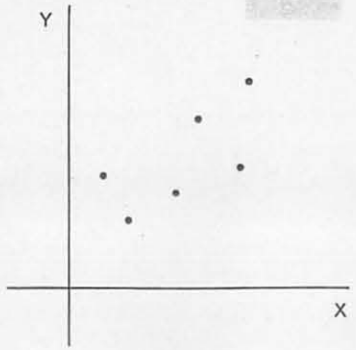


図 4.7　散布図。

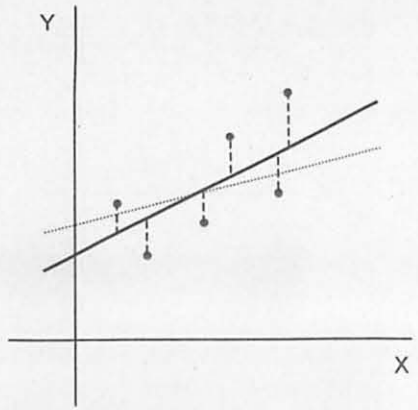


図 4.8　直線回帰：最小二乗法では個々の点から直線までの縦の距離の値を二乗した値の合計が、一番小さくなるように、黒の実線で示す直線を引く。

図 4.7 に示すような散布図で表される 6 症例のデータがあったとしよう。これらの 6 個の点の間を縫って、直線を引いて、X と Y の関係を表そうとする場合、どのように線を引いたらいいか考えてみよう。直線と個々の点の距離が一番小さくなるように線を引いたらいいのではないかと考えられる。たとえば、図 4.8 に示す、黒の実線とグレーの点線では、グレーの点線は 6 個の点の間に引かれているが、少しずれているように見える。一方、黒の実線は 6 個の点からの縦の距離の合計が一番小さくなるように引いてある。縦の距離といっても、個々の点から直線までの縦の距離は、直線の上に点がある場合には、正の値になるが、直線の下にある場合には、負の値になる。そこで、個々の点から直線までの縦の距離の値を二乗した値の合計が、一番小さくなるような直線を引く。これが、**最小二乗法**である。合計が一番小さいということがポイントである。

そのような直線の傾き b と切片 a は、X の値の平均値を $\bar{X}$、Y の値の平

4.0	2.4	3.8
3.1	3.2	3.4
2.9	2.3	2.9
1.4	2.9	3.3
3.7	3.7	3.5
3.5	3.3	3.4
2.6	2.8	2.9
2.9	3.3	2.9
3.0	3.7	1.9
3.5	3.9	2.8
3.4	3.6	2.9
4.2	4.5	3.4
3.1	2.8	3.2
3.3	2.5	3.3
3.0	3.3	3.2
1.4	3.3	3.4
3.3	3.5	3.4
3.5	3.4	3.4
3.4	3.3	3.4
2.6	3.4	2.8

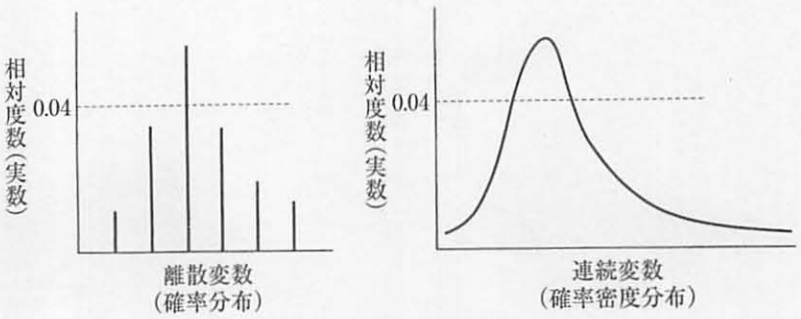


図 3·2　離散変数と連続変数

表 3·3　新生児の体重の累積度数分布表

体重 (kg)　(階級)	階級値 (kg)	累積度数 (人)	累積相対度数 (%)
0.5～1.0(未満)	0.75	1	0.3
1.0～1.5	1.25	3	1.0
1.5～2.0	1.75	6	2.0
2.0～2.5	2.25	17	5.7
2.5～3.0	2.75	91	30.3
3.0～3.5	3.25	229	76.3
3.5～4.0	3.75	288	96.0
4.0～4.5	4.25	297	99.0
4.5～5.0	4.75	299	99.7
5.0～5.5	5.25	300	100.0

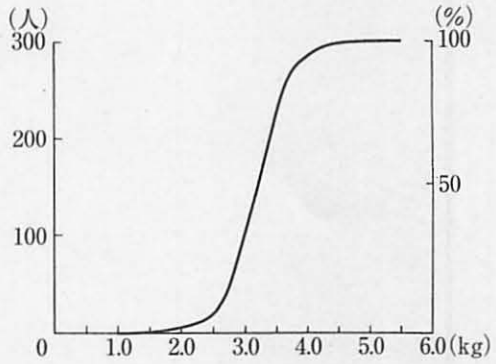


図 3·3　新生児の体重の累積度数分布曲線

未知であるが，自然法則を基に理論的に導かれる分布(正規分布，二項分布，ポアソン分布など)に従うと仮定する場合が多い．

ヒストグラムにおける柱の面積をそれぞれの階級における相対度数に等しくとっておけば，ヒストグラムの全体の面積は1(100%)となる．また同様に度数分布曲線と横軸との間の面積も1(100%)になる．

度数分布表においてその階級以下の度数，相対度数の和をそれぞれ，その階級の**累積度数**，**累積相対度数**といい，このようにして得られた表を**累積度数分布表**という(**表 3·3**)．

累積度数分布において度数分布から度数分布曲線を作ったような方法で，累積度数分布曲線を作ることができる(**図 3·3**)．

母集団では縦軸に累積相対度数(実数)に対応する値をとって，離散変数の場合は階段状のグラフで，連続変数の場合は右上がりの曲線のグラフとして考える．このグラフを**分布曲線**という(**図 3·4**)．

母集団が正規分布を示す場合は，累積度数分布曲線は特徴的な**シグモイド曲線**[*3]を示す．

累積度数分布曲線は薬の効力の記述や毒性の評価などにも利用される．投与量を対数目盛上にとるか，あるいは投与量の対数値を算術目盛上にとり，投与量を少しずつ増加したとき投与量に対する反応率は一般

*3　S字型の曲線で中央付近でほぼ直線になり，両端で曲線となる．

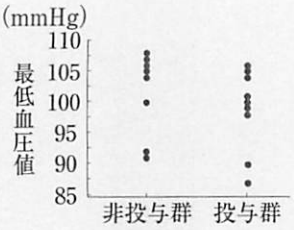


図 3·10　ドットプロット(プロット図)

表 3·5　2×2 分割表

効果 / 薬品	効果あり	効果なし
新薬	68	22
偽薬	51	39

(人)

例題

A 県の心疾患による年齢階級別死亡率，および B 県の年齢階級別人口は次のとおりである．

年齢(歳)	A 県の心疾患死亡率(人口 10 万対)	B 県の人口
0～4	6.1	56,326
5～9	1.6	53,177
10～14	1.1	58,163
15～19	2.9	53,186
20～24	5.5	42,548
25～29	7.6	56,810
30～34	7.1	47,159
35～39	21.9	46,857
40～44	25.7	60,419
45～49	38.3	60,210
50～54	58.5	51,633
55～59	102.2	42,972
60～64	190.2	39,190
65～69	375.4	33,060
70～74	619.9	27,134
75～79	1,099.9	19,320
80～	2,570.4	16,193

A 県の年齢階級別死亡率に B 県の年齢階級別人口を重みづけした加重平均を求めよ．

解答

求める加重平均を $\overline{X}_w$ とすると

$$\overline{X}_w=\frac{6.1\times 56{,}326+1.6\times 53{,}177+\cdots+2{,}570.4\times 16{,}193}{56{,}326+53{,}177+\cdots+16{,}193}=148.5$$

心疾患死亡率は加齢と関係があり，高齢者の多い人口では一般的に高くなるので，人口構成の異なる地域の死亡状況の比較では，死亡における年齢の影響を除く(調整する)ために，このような操作を行うことがある．

観察数が多く，標本を階級に分類し，度数分布をつくった場合には階級値を $X_1, \cdots, X_n$，その度数をそれぞれ $f_1, \cdots, f_n$ とすれば，度数分布から標本平均 $\overline{X}$ は次のように書ける．

$$\overline{X}=\frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

*4　$\sum$ はすべての項を合計する記号．$\sum_{i=1}^{n} X_i$ は X_i について i が 1 から n まで変わるときすべての X_i を合計する．
$\sum_{i=1}^{5} X_i=X_1+X_2+\cdots+X_5$．
$\sum$ の上下の添字を略すこともある．

❷ 中央値*5(メジアン)

標本の各個体を小さい順に並べたとき，中央に位置する値を**中央値**という．標本の大きさ n が偶数のときは中央にもっとも近い 2 つの値の算術平均で，標本の大きさ n が奇数のときは，小さい順に $\frac{n+1}{2}$ 番目の値で定められる．中央値はデータを小さい順に並べたときの 50% を表す点である(50%点)．

*5　母集団で考えたとき母中央値(分布の中央値，母メジアン)，標本で考えたとき標本中央値(メジアン)という．単に中央値(メジアン)という場合は標本中央値を指す場合が一般的である．

例題

10 人の新生児の体重は 3.2 kg, 3.8 kg, 3.1 kg, 2.8 kg, 4.2 kg, 2.9 kg, 2.9 kg, 3.4 kg, 3.5 kg, 3.0 kg であったとする．中央値を求めよ．

解答

体重の小さい順に並べると次のようになる．

2.8 kg, 2.9 kg, 2.9 kg, 3.0 kg, 3.1 kg, 3.2 kg, 3.4 kg, 3.5 kg, 3.8 kg, 4.2 kg

$n=10$ は偶数である．中央にもっとも近い位置にある 2 つの値は 3.1 kg と 3.2 kg である．したがって

$$\text{中央値}=\frac{3.1+3.2}{2}=\frac{6.3}{2}=3.15(\text{kg})$$

分布がほぼ左右対称の場合は，平均値・中央値は分布の中心をよく表し，どちらを用いても似たような値を示すが，分布に強いゆがみがある場合や，極端に飛び離れた値がある場合には**図 3·11** に示すように中央値の方がより中心的傾向を表している．完全に左右対称の分布においては平均値・中央値は一致する．

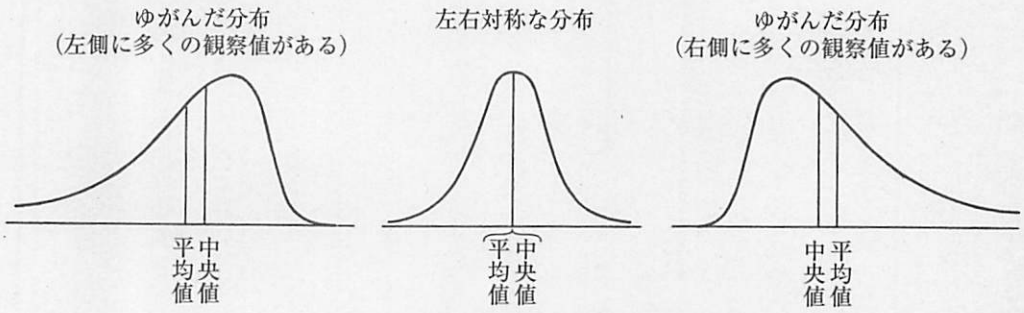


図 3·11　分布の平均値・中央値の位置

C．散らばりを示す指標(散布度)

平均は母集団および標本の代表値として用いられる場合が多いが，同じ値であっても**図 3·12** にみられるように，平均の近くに分布している場合とそうでない場合とがある．そこで母集団や標本の特性をみるには代表値だけではなく，代表値のまわりにどのように分布しているか，その散らばり具合をみることも必要である．このような散らばりの程度(散布度)を表すには一般に範囲，分散，標準偏差などが用いられる．

ような結果が得られたとする．

横軸(x)に体重をとり，縦軸(y)に身長をとり，これらの数値を座標平面上にプロットし散布図を作成したところ，**図3·13**のような関連性の傾向が図示された．この関連性を要約するため，**図3·14**のように2つの変数X, Yのそれぞれの平均$\overline{X}$, $\overline{Y}$を新しい原点になるように座標軸を移動する．

このとき新しい軸x', y'で測った点の座標(x', y')は

$$x'=x-\overline{X} \qquad y'=y-\overline{Y}$$

となる．したがって，x'が増加すればy'も増加するような直線的な関連傾向があるときは第1と第3象限に点が多く，第2と第4象限は点が少なくなる．この傾向はx'とy'の直線的関連が強いほど著しくなる．逆にx'が増加するとy'が減少するような直線的な関連傾向にあるときは

表 3·8　体重と身長

番　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
体重 X(kg)	63	57	48	53	68	64	73	60	59	50	55	58	66	64	52	60	60	62	56	56
身長 Y(cm)	164	173	159	163	173	168	178	169	168	162	161	166	176	172	166	163	172	174	170	163

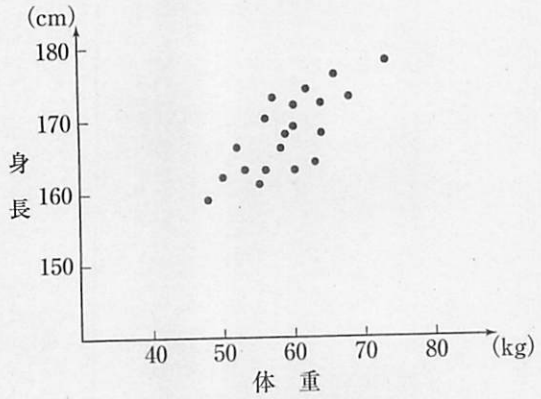


図 3·13　散布図

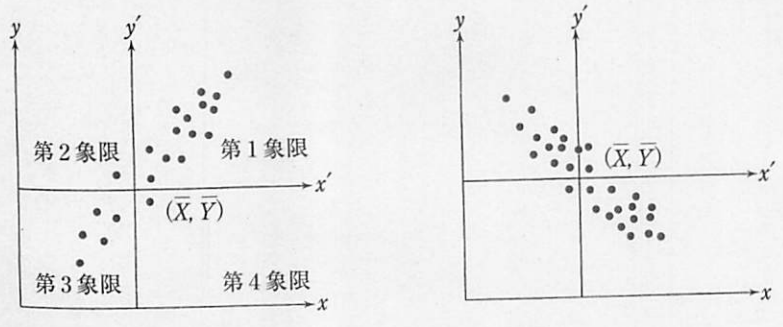


図 3·14　座標軸の位置と観察値

関が存在するという．rの絶対値が大きいほど相関傾向は強く，$r=+1$または$r=-1$のときは完全相関といい，XとYには直線関係が存在する．$r=0$のときは無相関という(**図3·15**)．rはXとYの直線関係の関連の強さを表す尺度である．

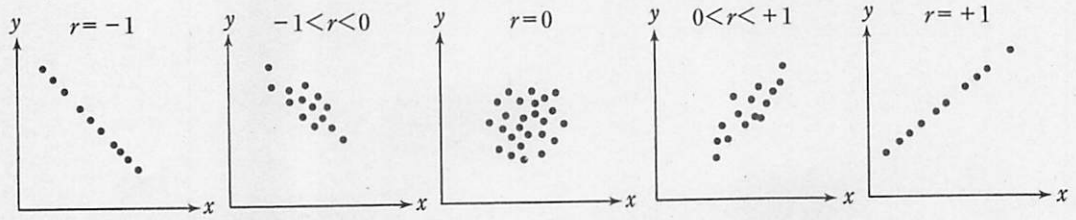


図 3·15　いろいろな相関

前述の20人の身長と体重の相関係数を求めると次のようになる．身長と体重の観察(測)値から**表3·9**が得られる．

したがって

$$r=\frac{\sum_{i=1}^{20}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{20}(X_i-\overline{X})^2\times\sum_{i=1}^{20}(Y_i-\overline{Y})^2}}=\frac{489}{\sqrt{729.2\times552}}=0.77$$

2変数間に統計的に相関が認められること(統計学的な関連)と因果関係(因果関連)が存在することは違う概念なので，注意が必要である．すなわち因果関係は原因と結果の関係なので，因果関係があれば相関は認められるが，逆は必ずしも認められない[*11]．

❸ Spearman(スピアマン)の順位相関係数

Pearson(ピアソン)の相関係数rはデータが数量的に測れるものでないと計算できない．しかし，それぞれのデータについて順位だけわかれば，この両者の相関関係を調べることができる．このような順位に対して計算された相関係数r_Sを**Spearmanの順位相関係数**という．

いま，$X_1,\cdots,X_n$の順位(ランク)[*12]を$R_1,\cdots,R_n$とし，$Y_1,\cdots,Y_n$の順位(ランク)を$S_1,\cdots,S_n$とすると順位相関係数r_Sは

$$r_S=\frac{\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\overline{R})^2\sum_{i=1}^{n}(S_i-\overline{S})^2}}=1-\frac{6\sum_{i=1}^{n}(R_i-S_i)^2}{n^3-n}$$

上式の左辺と右辺が等しいことは

$$\sum_{i=1}^{n}R_i=\sum_{i=1}^{n}S_i=\frac{n(n+1)}{2}$$

*11 たとえば「交通事故が増えたことが原因で，救急車の出動回数が増えた」という自然な因果関係を仮定する．このとき「交通事故の頻度」と「救急車の出動頻度」には，統計学的な関連が認められるであろう．すなわちこれらの散布図を描いたときに右上がりの傾向(相関)が見られる．しかしこの統計学的な関連があるからと言って，「救急車の出動回数が増えたことが原因で交通事故が増えた」という因果関係は，普通は考えない(因果関係は認められない)．

*12　$X_1,\cdots,X_n$を小さい順に並び替えたとき，X_iが小さい方から数えてs番目であるとき，X_iの順位(ランク)はSである．

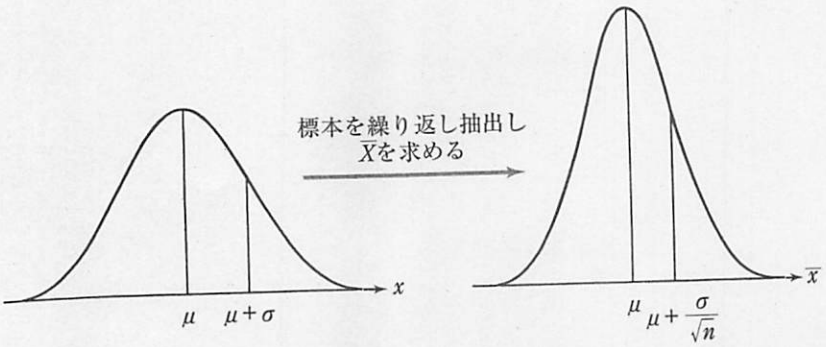


図 4·12　標本平均の分布

❼ χ^2(カイ二乗)分布

正規分布する母集団から無作為に大きさ n の標本を抽出し，その標本から求められた偏差平方和 SS を母分散 σ^2 で割った

$$\chi^2=\frac{SS}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$$

は，自由度 $df=n-1$ の χ^2 分布に従うことがわかっている(**図 4·13**).

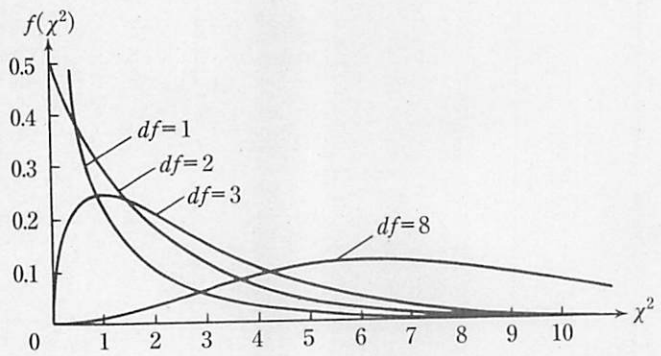


図 4·13　種々の自由度の χ^2 分布

自由度 df の χ^2 分布(χ^2_{df} 分布)の確率密度関数は

$$f(\chi^2)=\frac{1}{2\Gamma\left(\frac{df}{2}\right)}\times e^{-\chi^2/2}\left(\frac{\chi^2}{2}\right)^{df/2-1}\qquad(0<\chi^2<\infty)$$

である[*8].

*8 Γ はガンマ関数で正の実数 x について
$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$
で定義される.

6 2群の比較

本章のねらい

- 2群を比較するときに用いられる検定手法を学ぶ.
- データの形によって検定手法が異なる点を理解する.

2群を比較するとき，この2群間には「対応がない」場合と「対応がある」場合がある.「**対応がある2群**」とは，2群のデータの間に薬の投与前，投与後などのように，同一対象の異なる2時点の観測のような対応が成立する場合である．これに対して，そのような対応が存在しない2群を「**対応のない2群**」という.

A. 対応のない2群の差の検定

❶ t 検定

平均に関する検定には母平均の差の検定がある．2つの母集団の分散の状況によって，用いる検定方法が異なるので，一般に平均の検定の前に分散に関する検定を行い，比較しようとする2群の母分散が等しいかどうかを判断する(56ページ(3)参照).

一般に，正規分布に従う2つの独立な確率変数 X, Y について，X が期待値 μ_X，分散 σ_X^2 に従い($X \sim N(\mu_X, \sigma_X^2)$)，$Y$ が期待値 μ_Y，分散 σ_Y^2 に従う($Y \sim N(\mu_Y, \sigma_Y^2)$)とき，

$$aX+bY \sim N(a\mu_X+b\mu_Y,\ a^2\sigma_X^2+b^2\sigma_Y^2)$$

が成立する．このことから，正規分布 $N(\mu, \sigma^2)$ に従う標本 $X_1, \cdots, X_n$ の標本平均 $\overline{X}$ は正規分布 $N(\mu, \dfrac{\sigma^2}{n})$ に従うことがわかる[*1]．また，正規分布 $N(\mu_1, \sigma_1^2)$ に従う大きさ n_1 の標本からの標本平均 $\overline{X}_1$(分布は $N(\mu_1, \dfrac{\sigma_1^2}{n_1})$)と，正規分布 $N(\mu_2, \sigma_2^2)$ に従う大きさ n_2 の標本からの標本平均 $\overline{X}_2$(分布は $N(\mu_2, \dfrac{\sigma_2^2}{n_2})$)について，その差 $\overline{X}_1-\overline{X}_2$ は正規分布 $N(\mu_1-\mu_2, \dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2})$ に従うことがわかる[*2]．このことから $\overline{X}_1-\overline{X}_2$ の分布を標準化すると，

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

となることがわかる．この関係が正規分布を用いた検定では重要になる.

*1 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$
$\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

*2 $\overline{X}_1 \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right)$
$\overline{X}_2 \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right)$
$\overline{X}_1-\overline{X}_2 \sim N\left(\mu_1-\mu_2, \dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$

例題

集団 A(男性)14 人，集団 B(男性)12 人について血色素量(g/dL)を調べたところ，次のような結果を得た．両集団に差が認められるか．有意水準 1%で検定せよ．なお，A と B の分散は未知だが等しいとする．

番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	15.3	14.0	14.5	17.0	15.2	15.3	14.6	16.2	17.0	15.1	16.2	17.1	16.8	15.9
B	14.1	16.2	13.2	15.0	16.1	13.1	14.4	16.3	14.1	14.7	16.3	16.1		

解答

① 帰無仮説，対立仮説を立てる．

H_0：集団 A の血色素量の平均と集団 B の血色素量の平均には差がない($\mu_1=\mu_2$)．

H_1：集団 A の血色素量の平均と集団 B の血色素量の平均に差がある($\mu_1\neq\mu_2$)(両側検定)

② 有意水準を定める．

$\alpha=0.01$(1%)

③ 検定統計量 $T=\dfrac{\overline{X}_1-\overline{X}_2}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ にデータを代入する．

ここで

$$\overline{X}_1=15.73 \qquad \overline{X}_2=14.97$$

$$S_1^2=\frac{\Sigma(X_1-\overline{X}_1)^2}{n_1-1}=\frac{13.55}{14-1}=\frac{13.55}{13}=1.04$$

$$S_2^2=\frac{\Sigma(X_2-\overline{X}_2)^2}{n_2-1}=\frac{16.15}{12-1}=\frac{16.15}{11}=1.47$$

$$n_1=14 \qquad n_2=12$$

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}=\frac{(14-1)\times1.04+(12-1)\times1.47}{14+12-2}=\frac{29.69}{24}=1.24$$

よって

$$S_p=\sqrt{1.24}=1.11$$

したがって検定統計量 T にデータを代入した値は

$$T=\frac{\overline{X}_1-\overline{X}_2}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}=\frac{15.73-14.97}{1.11\sqrt{\dfrac{1}{14}+\dfrac{1}{12}}}=\frac{0.76}{1.11\times0.39}=1.76$$

7 分割表の解析

本章のねらい

▶分割表を用いた検定手法を学ぶ.

A. 独立性の検定

❶ χ^2 検定

ある大きさの観察標本が行と列に分類され，組み分けされたところに数値が記入されているような表を**分割表**という(**表 7·1**)．この分割表において行と列が互いに独立であるか(関連の有無)を統計的に検定することができる．これを**独立性の検定**といい，対応のない質的データの差の検定と理解できる.

表 7·1　2×2 分割表

	+	−	計
Ⅰ	N_{11}	N_{12}	$N_{1\cdot}(=N_{11}+N_{12})$
Ⅱ	N_{21}	N_{22}	$N_{2\cdot}(=N_{21}+N_{22})$
計	$N_{\cdot 1}(=N_{11}+N_{21})$	$N_{\cdot 2}(=N_{12}+N_{22})$	$n(=N_{11}+N_{12}+N_{21}+N_{22})$

いま，**表 7·1** のような 2×2 分割表が得られたとする．このとき，各セル(Ⅰ, +)，(Ⅰ, −)，(Ⅱ, +)，(Ⅱ, −)は**表 7·2** で表される確率に従って起こりやすさが表されているとする．これは母集団の状態が確率分布によって定まり，標本の状況が 2×2 分割表によって表されていると考えることができる.

表 7·2　確率分布

	+	−	計
Ⅰ	p_{11}	p_{12}	$p_{1\cdot}(=p_{11}+p_{12})$
Ⅱ	p_{21}	p_{22}	$p_{2\cdot}=(p_{21}+p_{22})$
計	$p_{\cdot 1}(=p_{11}+p_{21})$	$p_{\cdot 2}(=p_{12}+p_{22})$	1

行と列が独立であるとは，$p_{ij}=p_{i\cdot}\times p_{\cdot j}(i=1, 2, j=1, 2)$ が成り立つことである．ここで,

C. 線形回帰分析

❶ 回帰直線

2変数 X と Y の関連が強く，$Y=\beta_0+\beta_1 X$ という直線的な関係が想定されるとき，このような直線を**回帰直線**という．母集団上で考えられた回帰直線を**母回帰直線**，標本上で考えられた回帰直線を**標本回帰直線**という．**y 切片**，**傾きを回帰係数**といい，母数のときはギリシア文字 β_0, β_1 を，統計量のときはアルファベット b_0, b_1 を用いることにする．β_0, β_1 は**母回帰係数**であり，b_0, b_1 は**標本回帰係数**である．数学的には標本上で

$$Y_i=b_0+b_1X_i+\varepsilon_i \quad (i=1, \cdots, n)$$ *1

となるモデルを考える．ここで $\varepsilon_i(i=1, \cdots, n)$ は互いに独立に，平均 0，分散 σ^2 の分布に従う．

$b_0+b_1X_i$ は，X_i を用いて Y_i を予測する式になっているので，予測した値を Y_i の**予測値**といい，$\hat{Y}_i(=b_0+b_1X_i)$ と書く．Y_i と $\hat{Y}_i$ の差を**残差**といい $e_i(=Y_i-(b_0+b_1X_i))$ で表す(**図 10·1**)．

*1　回帰，重回帰を考えるとき，従属変数は確率変数，説明変数は確率変動しない普通の変数を考えるので，本来は X_i を x_i と小文字で書くべきところであるが，この本では慣例に従って，大文字で書くことにする．

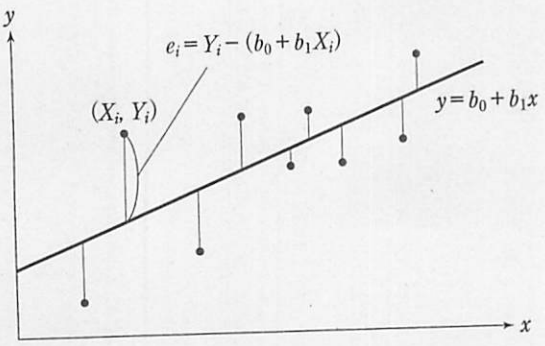


図 10·1　残差 e_i

標本回帰係数は次のようにして求めることができる．

n 個の観察値 (X_i, Y_i) が**図 10·1** のように得られたとき，各観察(測)値から直線 $y=b_0+b_1x$ までの y 軸に平行な距離(残差)の平方の和が最小になるように b_0, b_1 を決定する．このような方法を最小2乗法という．このとき，距離(残差)の平方和 D は次のようになる．

$$D=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}[Y_i-(b_0+b_1X_i)]^2$$

D を最小にするように b_0, b_1 を決めるには，b_0^2, b_1^2 の係数が正なので，…て偏微分*2して0とおけばよい*3(これは1変数の

*2　x, y の関数 $f(x, y)$ において，仮に y は定数と考え，x だけの関数とみて，これを微分することを $f(x, y)$ を x について偏微分するという．たとえば，$z=f(x, y)=ax^2+2bxy+cy^2$ を x について偏微分すると

$\dfrac{\partial z}{\partial x}=2ax+2by$ となる．

*3　このようなことから，β_0, β_1 (b_0, b_1) を母(標本)偏回帰係数ということもある．

§1.4　ベイズの定理 Bayes theorem

さて、枠の色がわかる前の時点では、疾患1である疾患確率は

$$\frac{5}{25}=0.2$$

であった。それが、赤であることがわかった後では0.6に上昇した。この関係を表すのが**ベイズの定理** Bayes theorem である。この例をベイズの定理にしたがって式で表すと次のようになる。

$$\frac{\frac{3}{5}\times\frac{5}{25}}{\frac{3}{5}\times\frac{5}{25}+\frac{1}{7}\times\frac{7}{25}+\frac{1}{13}\times\frac{13}{25}}=0.6=\frac{3}{3+1+1}$$

この式の左辺をよく見ると、分母はこの集団全体の中で症状を認める確率、分子はその中で、疾患1の患者の分を示し、計算結果は症状を認めた場合の疾患1である確率を示していることがわかる。さらに 赤 の症状を認める者全体のなかで、疾患1にかかっている者の割合に相当することがわかる。

それでは、ベイズの定理とはどのようなものか、解説する。

◎　1.4.1 ベイズの定理　◎

ベイズの定理は、ある仮説の正しい確率がデータを得た後にどのように変化するかを示すものである。

たとえば人類の男性と女性の割合はそれぞれ50％だという仮説を立てたとする。その仮説が正しい確率を最初、0.5、すなわち五分五分だと考えたとする。そこである街のある通りで道行く人を観察し、男性と女性の人数を調査したとしよう。

もし、20人の人を観察してそのうち9人が男性、11人が女性だったとしよう。この場合、男性と女性の割合がそれぞれ50％だという仮説が正しいとすると、20人のうち9人が男性、11人が女性だというデータを得る確率はかなり高いと推測される。したがって人類の男性と女性の割合がそれぞれ50％という仮説の正しい確率は、当初考えた0.5よりももっと高くした方が正しいと考えられる。

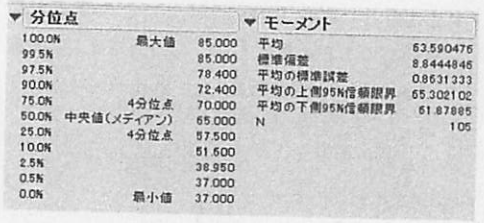
逆に、20人の人を観察してそのうち2人が男性、18人が女性だったとしよう。もし人類の男性と女性の割合がそれぞれ50％という仮説が正しいとするとそのような観察データを得る確率は低いと考えられる。したがってこのようなデータを得た場合には当初の仮説が正しい確率はもっと低く考えた方が正しいと言えるであろう。

以上述べたことは直感的に正しいであろうと推定することができるが、それを数式で表したものがベイズの定理である。データを得る前の仮説が正しい確率を**事前確率** prior probability、データを得た後のその確率を**事後確率** posterior probability と呼ぶ。

◆ベイズの定理◆

Thomas Bayes（トーマス・ベイズ、1702-1761）は英国Tunbridge WellsのPresbyterian Chapelの聖職者で、彼がこの定理について書いたエッセー、“An essay towards solving a problem in the doctrine of chances.”は彼の死後、彼の友人Richard PriceによってPhilosophical Transactions of the Royal Society of Londonに送られ、1764年に53巻370-418頁に発表された。

ベイズの定理は、母集団の母数の分布を推定する場合にも適用される。すなわち、平均値と標準偏差（あるいは信頼区間）、割合の場合であれば、割合の値とばらつきの指標である信頼区間など。この定理は、事前分布が、あるデータを得た際に変化して、事後分布が得られる際の原理を表している。さらに、もっと一般化して、ある仮説の正しさの度合いが、あるデータを得た際に、どのように変わるかを示しているということができ、科学の方法論を示しているともいえる。



分位点			モーメント	
100.0%	最大値	85.000	平均	63.590476
99.5%		85.000	標準偏差	8.8444846
97.5%		78.400	平均の標準誤差	0.8631333
90.0%		72.400	平均の上側95%信頼限界	65.302102
75.0%	4分位点	70.000	平均の下側95%信頼限界	61.87885
50.0%	中央値(メディアン)	65.000	N	105
25.0%	4分位点	57.500		
10.0%		51.600		
2.5%		38.950		
0.5%		37.000		
0.0%	最小値	37.000		

図 2.3　分位点とパーセンタイル。

ができる。$n-1$ 個目までは自由になる。そこで、サンプル数 n から 1 を引き算した値を自由度と呼ぶ。n が大きくなると、n でわり算するのと、$n-1$、すなわち、自由度でわり算するのでは、ほとんど同じ値になる。しかし、n が小さい場合には、自由度でわり算する方が大きな値になる。

それでは、先ほどの例で計算してみよう。図 2.3 に示すように平均値 63.6 歳に対して、標準偏差は 8.8 歳であった。図 2.1 の分布を見ると、平均値から平均値＋標準偏差＝72.4 歳までの間に全症例のうち、約 3 分の 1 位が入っているように見える。(正規分布であれば、正確には 34.1 %が平均値と平均値＋標準偏差の間に分布する)。このように、平均値と標準偏差がわかると、その測定値の分布が、おおよそ想像が付く。この例では、正規分布とは少しずれているので、平均値と標準偏差から想像される分布（図 2.1 の少し薄い（実際は赤）色の曲線で示されている）から、実際の分布を想像することはやや難しくなることがわかる。

それでは、正規分布に従わない場合に、中央傾向を表す指標は中央値を用いるとして、ばらつきの指標としては何を用いたらいいであろうか。その値が与えられると、分布が想像できるような指標が必要である。それは、**最小値**、**最大値**と**分位点**あるいは**パーセンタイル** (percentile) である。JMP で上記データを解析した結果を見てみよう（図 2.3)。

分位点の結果を見ると、中央値（メディアン）と 4 分位点が真ん中あたり表示されている。4 分位点 quartile は 75 %と 25 %、中央値は 50 %に

対応していることからわかるように、測定値を小さいほうから大きいほうに順に並べ、最小値から、観察値を見ていき、測定値の 25 %までが含まれる点が、**25 パーセンタイル** (25 percentile) で、小さい方の 4 分位点である。さらに、50 %までが含まれる点が 50 パーセンタイルで中央値であり、さらに、75 %までが含まれる点が 75 パーセンタイルで、大きいほうの 4 分位点である。2 つの 4 分位点にはさまれる範囲が**4 分位範囲**と呼ばれ、全サンプルの 50 %が含まれる測定値の範囲になる。4 分位点は**ヒンジ** hinge と呼ばれることもある。左右対称の分布でない場合には、中央値から 4 分位点までの間隔は、上下で異なる。

分位点をみて、測定値がどのように分布しているかをすぐに想像できる人はそうはいないであろう。より直感的に理解できるように、**箱ひげ図** (box and whisker plot あるいは単に box plot と呼ばれる）というものが用いられる。図 2.4 に、ここで取り上げた例での、箱ひげ図を示す。箱は 4 分位範囲に相当し、測定値の 50 %がこの範囲に含まれる。4
中央値である。すなわち、4 分位範囲の中に、中央値があるが
べたように必ずしも箱の真ん中というわけではない。さて、ひげはいろいろな使われ方がある。

たとえば、この例では、2 種類の箱ひげ図が表示されているが、上段は、[はずれ値の箱ひげ図] と呼ばれるもので、ひげは、次の式で計算された範囲内で最も遠くにある測定値まで伸びている。すなわち、箱の端から、箱の長さの 1.5 倍離れた点である。この範囲から外れる値は、**はずれ値** outlier と呼ばれる。この例でも、若年の方に，2 例はずれ値が認められる。

$$\text{上側 4 分位点} + 1.5 \times (\text{4 分位範囲})$$

$$\text{下側 4 分位点} - 1.5 \times (\text{4 分位範囲})$$

下段の箱ひげ図は [分位点の箱ひげ図] と呼ばれるもので、最小値、2.5 パーセンタイル、10 パーセンタイル、90 パーセンタイル、97.5 パーセン

うる。比較される群の数も2つの場合、3つ以上の場合などさまざまである。群分けの変数は名義変数であることが多い。それぞれの測定値はどの群に属するかという値（名義変数）と実際の測定値の2つのデータを保持している。したがって、データを準備する場合には、表計算ソフトを用いて、一つの行に1人分のデータを入力し、一つのカラムに群分けの変数、もう一つのカラムに測定値を入力したものを用意する。一例を図3.1に示す。

§3.2　比較の例

たとえば、ランダム化比較試験で、新しい薬剤とプラセボの有効性を比較する場合には、症状の有無など名義変数をアウトカムとして、症状を有する症例の割合が、実薬群の方がプラセボ群より小さいかどうかが問題になる。たとえば、実薬群で10例中2例が症状（+）、プラセボ群で10例中5例が症状（+）というデータを得たとしよう。それぞれ、アウトカム（+）の割合は、20％と50％ということになる。有効性という点から見れば、有効率は、80％と50％である。リスク比で表せば、20％÷50％で、0.4である。さて、このデータからこの薬剤は有効であるといえるであろうか？

次のうち、どの考えが正しいであろうか？

1. 症状を認める者の割合が半分以下に大きく減少しているから、この薬剤は効果がある。
2. 実際に被験者となって調べられた患者は、もっと大きな集団、すなわち母集団からのサンプルであり、それぞれ母集団の有効率は80％と50％である可能性が高いから、この薬剤は有効である。
3. 薬剤の有効率が本当は50％であった場合でも、10人しか調べなかった場合には、たまたま80％で有効ということが起こりえるから、この薬剤は有効とはいえない。

4. 総症例20例を1つにまとめて考えると、症状（+）の者は7例であり、7÷20＝0.35すなわち、35％の者で症状（+）ということになる。したがって、有効率は65％である。もし、実薬群とプラセボ群が本当は、有効率65％で同じ母集団からのサンプルだと仮定したら、10人ずつ調べた場合に、それぞれの有効率が80％と50％になる可能性はかなり高いと考えられるので、この薬剤は有効とはいえない。
5. それぞれの母集団の有効率が80％と50％と考えた場合、10人ずつのサンプルで、80％と50％の有効率になるデータを得る可能性が一番高いから、この薬剤は有効と考えるべきである。
6. 5.の考え方をさらに拡張して、それぞれの母集団の有効率が80％と50％と考えた場合、リスク比が0.4となる確率が一番高いと考えられるが、それでもなお、リスク比が1以下になる確率が低ければ、この薬剤は有効と考えていいのではないか。

さて、あなたならどう考えますか？

ここには、2つの考え方のあることが見える。一つは、実際のデータのことしか考えない考え方。もう一つは、これら20症例がもっと大きな集団、つまり母集団からのサンプルであって、本当に問題なのは、母集団の値（母数）、この場合、有効率という割合であるという考え方である。

このランダム化比較試験の結果が、もしこの薬剤のほうがプラセボより症状の消失する率が高いということを確実に証明しているのであれば、同じ疾患の患者にこの薬剤を投与することが、これから行なわれることになる。その場合、これから新たに投与されるであろう患者は試験での被験者と同じ属性を有している、すなわち、同じ母集団に属していないと、同じ効果は期待できない。ランダム化比較試験を含めた臨床研究の結果を、個々の患者に当てはめていいかどうかは、**外的妥当性**という言葉で表現される。

に1回しか起きない。ここで、2つの考え方ができる。一つは、「このさいころは普通のさいころで、たまたま0.054の確率でしか起きないめずらしいことが起きて、偶然6回中3回も1の目が出た」と考えることである。もう一つは、「このさいころは普通のさいころではなく、何か細工がしてあるため、1の目が出る確率がおそらく0.5位（3÷6）であって、1の目の出る確率が$\frac{1}{6}$であるというのは間違いである」と考えることである。

これは言い換えると、「1の目の出る確率が$\frac{1}{6}$であるという仮説を立てて、その仮説が正しいとしたら、6回転がして、1の目が3回出る確率を求めて、その確率が高ければ、仮説が正しいとして受け入れ、その確率が低ければ、仮説は間違っているとして却下する」ということである。これはまさに統計学的検定の考え方であり、仮説が正しいとしたときに、そのようなデータを得る確率として、算出されるのが、P値である。本当の母集団の値を正確に知ることができなくても、ある仮説の下で、そのようなデータを得る確率を知ることは可能であり、その確率＝P値が小さな値の場合には、その仮説が正しい可能性が低いので、その仮説に対立する仮説が正しい可能性が高いと判断するということである。

さてランダム化比較試験に戻るが、上記3. 4. 5. 以外にも、「母集団はこのようなものである」、という仮説は無数に作ることができる。それぞれについて、その仮説が正しい場合にこのようなデータを得る確率を計算することは可能であろう。しかし、それでは、きりがない。そこで、一番重要な仮説は何かを考えてみよう。それは、「母集団の有効率には差がない」という仮説である。すなわち、上記4. の考え方である。これを**帰無仮説** null hypothesis と呼ぶ。これに対立する仮説は**対立仮説** alternative hypothesis と呼ばれる。統計学的検定では多くの場合、この帰無仮説を設定して、その仮説の下で、そのようなデータを得る確率を計算し、その値がある一定の値以下であれば、帰無仮説を棄却し、対立仮説を採用するということが行なわれる。

§3.3　割合の比較：カイ二乗検定

上記の例で実薬とプラセボで有効率に本当に差があるといえるのか、解析してみよう。

表3.1　2群でのランダム化比較試験の結果

	実薬群	プラセボ群	計
有効	8	5	13
無効	2	5	7
計	10	10	20

さて、比較する2つの群は別の患者であり、実薬の
ことが、プラセボを投与されるということに何の影響も及ぼさないから、これら2つの群は、互いに独立しているといえる。ここで、比較しようとしているのは、

$$\frac{8}{10}(80\,\%\text{あるいは}\,0.8)\ \text{と}\ \frac{5}{10}(50\,\%\text{あるいは}\,0.5)$$

という割合である。あるいは、率 rate といってもよい。2つの独立した群の割合の比較をしようとしているのである。

さて、表3.1の外側の欄、すなわち計のセルだけを書き出したのが、表3.2である。もし、有効率が$13\div 20=0.65$、すなわち65％の母集団を想定すると、中央の4つのセルにはどのような症例数が入る可能性が一番高いかを考えてみよう。この0.65という値は、それぞれの群の症例数で重み付けした、有効率の平均値に相当する。

$$0.65=\frac{0.8\times 10+0.5\times 10}{10+10}$$
$$=\frac{\text{有効率}_{\text{実薬群}}\times\text{症例数}_{\text{実薬群}}+\text{有効率}_{\text{プラセボ群}}\times\text{症例数}_{\text{プラセボ群}}}{\text{症例数}_{\text{実薬群}}+\text{症例数}_{\text{プラセボ群}}}$$

表 4.2　相関係数の計算：相関が全くない場合

症例	X	Y	X−X̄	Y−Ȳ	(X−X̄)(Y−Ȳ)	(X−X̄)²	(Y−Ȳ)²
1	8	8	−2	−2	4	4	16
2	8	12	−2	2	−4	4	4
3	10	10	0	0	0	0	0
4	12	8	2	−2	−4	4	4
5	12	12	2	2	4	4	16
	X̄=10	Ȳ=10		合計	0	16	16

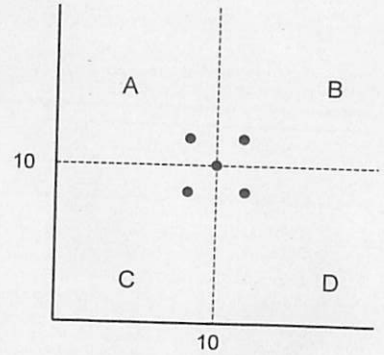


図 4.4　全く相関がない場合の散布図の例。

このデータを散布図として、図示すると、図 4.4 のようになる。

表 4.2 の値から相関係数を算出すると次のようになる。

$$r = \frac{0}{\sqrt{16} \times \sqrt{16}} = 0$$

分子が 0 であるから、相関係数は 0 となる。

なお、この場合の、分母の値は、上記の図 4.3 の例と比べると、より小さな値である。すなわち、点は図 4.3 よりも、より狭い範囲にあるため、より小さな値になる。したがって、散布図で、点が広い範囲に広がっていると、分母は大きな値になるといえる。

このように、相関係数は、2 種類の測定値の間に相関があるかどうかを表すのに適した指標である。

§ 4.3　相関係数に対する t 検定

母集団からランダムに採取したサンプルの 2 種類の数値変数の間の相関係数 r は t 分布に従うことを利用して有意差検定を行い P 値を求めることができる。他の有意差検定と同じ考え方である。相関に対する t 統計値は下記の式で算出されるが、この値は、n をサンプル数とすると、自由度 $n-2$ の t 分布に従う。式の r が相関係数である。

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

母集団の相関係数を rho（ロー）、$\overset{ロー}{\rho}$ と呼ぶ。まず、母集団では 2 つの数値変数の間には相関が全くない、すなわち $\rho = 0$ という帰無仮説を立てる。サンプル数 n から計算された相関係数 r を算出し、さらに r から上記の式で、t 統計値をもとめ、自由度 $n-2$ の t 分布から、設定した α 水準に対応する t 統計値を越えれば棄却し、以下であれば受け入れる。

身長と靴のサイズはある程度相関があると推測されるが、たとえば、「人間では、身長と靴のサイズの間に全く相関がない」という仮説を立てる。すなわち、「$\rho = 0$ である」とする帰無仮説を立てる。そして、n 人の人を集めて、身長と靴のサイズを測定し、上記の式で、相関係数 r を計算する。帰無仮説の下では、母集団では身長と靴のサイズに相関がないはずであるが、偶然サンプリングの偏りで、ある程度の相関が認められたとしよう。たとえば、r が 0.75 になったとする。

次に t 統計値を計算する。t 統計値が大きな値になった場合、そのよう

均値を $\bar{Y}$ とすると、次の式で算出される：

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}}$$
$$a = \bar{Y} - b\bar{X}$$

なお、サンプルではなく、母集団について、直線回帰を論ずる場合には、傾きを β_1、切片を β_0 で表す。そして、個々の点は、ほとんどの場合、直線上に乗るわけではないので、直線からの誤差を伴う。誤差は ε（イプシロン）で表す。したがって、母集団の X と Y の 2 つの数値変数の関係は次の式で表される：

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

傾きの β_1 のことを、**回帰係数** regression coefficient と呼ぶ。

さて、相関係数 r は X と Y の間に何の関係もない場合には 0 となり、X が増加すると Y が増加するような場合、すなわち、散布図で左下から右上に向かって点が分布しているような場合には、正の値となり、散布図で左上から右下に向かって点が分布しているような場合には、負の値となり、-1 から $+1$ までの値をとる。そして、サンプルについての回帰係数、すなわち b と、相関係数 r の間には次のような関係がある。s_Y は Y の値の標準偏差、s_X は X の値の標準偏差である：

$$b = r\frac{s_Y}{s_X}$$
$$r = b\frac{s_X}{s_Y}$$

Y の値がばらつきが大きいということは、散布図で縦に広い範囲に点が広がっていることになり、s_Y は大きな値になる。X の値のばらつきが小さいということは、散布図で横に狭い範囲に点が広がっていることになり、s_X は小さな値になる。このような場合には、図 4.9 に示すように、当然傾

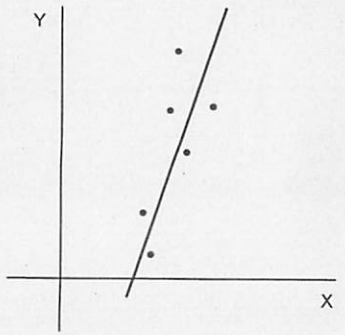


図 4.9　回帰係数が大きな値の場合。

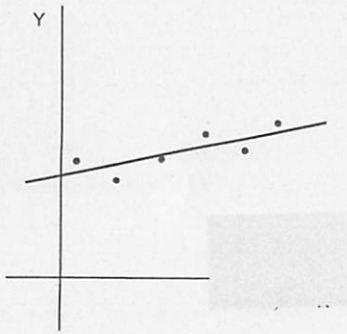


図 4.10　回帰係数が小さな場合。

	A	B	C	D	E	F	G	H	I
1	Case No	Height	Weight						
2	1	156	48	Height Mean	170.6	=AVERAGE(B2:B21)			
3	2	159	54	Height SD	7.293833	=STDEV(B2:B21)			
4	3	160	52	Weight Mean	65.7	=AVERAGE(C2:C21)			
5	4	163	60	Weight SD	11.58538	=STDEV(C2:C21)			
6	5	165	50						
7	6	166	46	Correlation coefficient	0.852426	=CORREL(B2:B21,C2:C21)			
8	7	168	56	Regression coefficient	1.353977	=INDEX(LINEST(C2:C21,B2:B21,TRUE,FALSE),1)			
9	8	169	66	Intercept	-165.288	=INDEX(LINEST(C2:C21,B2:B21,TRUE,FALSE),2)			
10	9	170	65	R-square	0.72663	=INDEX(LINEST(C2:C21,B2:B21,TRUE,TRUE),3)			
11	10	171	67						
12	11	172	75						

図 4.11　Microsoft Excel の関数を用いた相関、直線回帰の解析。カラム E の各セルにはそれぞれ右隣のカラム F に表示してある式が入力されている。

きは急で b は大きな値になる。上の式からわかるように、この場合、b は大きな値になる。

逆に、Y の値のばらつきが小さく、X の値のばらつきが大きい場合には、図 4.10 に示すように、傾きはゆるやかで、b は小さな値になる。

相関や直線回帰に関する関数は Microsoft Excel でも用意されている。単純なケースでは Microsoft Excel を用いて相関係数、回帰係数、切片などの値を算出することもできる。

図 4.11 に示すように、`CORREL( )` が相関係数を求める関数である。カッコ内には身長 Height のデータが入っているセルの範囲 `B2:B21` と体重 Weight

43 (1) $\frac{1}{5}(15+21+13+19+20)=\frac{88}{5}$
$=17.6$

(2) $\frac{1}{8}(45+38+52+54+73+27+25+42)$
$=\frac{356}{8}=44.5$

(3) $\frac{1}{10}\{2+9+6+(-9)+1$
$+(-5)+6+1+2+(-3)\}$
$=\frac{10}{10}=1$

44 A店のデータの最頻値は 11号,
B店のデータの最頻値は 9号 である。

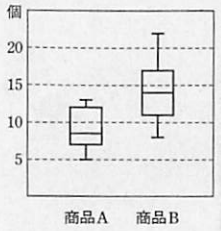
45 (1) データを小さい順に並べると
8, 14, 22, 48, 97
データの大きさは5であるから，中央値は3番目の値である。
よって，中央値は 22
(2) データを小さい順に並べると
11, 20, 20, 38, 39, 50, 51
データの大きさは7であるから，中央値は4番目の値である。
よって，中央値は 38
(3) データを小さい順に並べると
20, 33, 40, 59, 60, 62, 64, 91
データの大きさは8であるから，中央値は4番目の値と5番目の値の平均値である。
よって，中央値は $\frac{1}{2}(59+60)=59.5$

46 札幌のデータの範囲は
$17-1=16$(日)
那覇のデータの範囲は $12-3=9$(日)
札幌の方が範囲が大きいから，データの散らばりの度合いが大きいと考えられる。

47 データを大きさの[illegible]と
5, 9, 10, 11, [illegible]5, 17, 20
第2四分位数は [illegible]
第1四分位数は [illegible]$=9.5$,
第3四分位数は $Q_3=\frac{15+17}{2}=16$

48 データを大きさの順に並べると
46, 48, 49, 50, 50, 51, 51, 52,
52, 53, 54, 54, 55, 57, 58
このデータの最小値，第1四分位数，中央値，第3四分位数，最大値は，順に
46, 50, 52, 54, 58
これらの値をとっている箱ひげ図は ③

49 それぞれのデータを大きさの順に並べると
商品A 5, 6, 7, 7, 8,
9, 10, 12, 12, 13
商品B 8, 10, 11, 12, 13,
15, 16, 17, 20, 22
よって，それぞれのデータの最小値，第1四分位数，中央値，第3四分位数，最大値は，順に
商品A 5, 7, $\frac{8+9}{2}=8.5$, 12, 13
商品B 8, 11, $\frac{13+15}{2}=14$, 17, 22
よって，箱ひげ図は下の図のようになる。



50 平均値 $\overline{x}$ は
$\overline{x}=\frac{1}{5}(1+3+4+10+12)=\frac{30}{5}=6$
分散 s^2 は
$s^2=\frac{1}{5}\{(1-6)^2+(3-6)^2+(4-6)^2$
$+(10-6)^2+(12-6)^2\}=\frac{90}{5}=18$
よって，標準偏差 s は $s=\sqrt{18}\fallingdotseq 4.24$

別解 分散 s^2 は
$s^2=\frac{1}{5}(1^2+3^2+4^2+10^2+12^2)-6^2=18$

51 ①は番号4の $(x, y)=(8, 1)$ がない。
②は番号3の $(x, y)=(4, 6)$ がない。
よって，正しい散布図は ③

第2章　8章

第3章　９章

関連図書