基本的なデータ処理と SPSS の利用法 ver.2.4

最終更新 5/26/2008　中山和弘

１）データの種類

問題となっていることはどんなことか。多くは、
・ある人の数値が平均よりも大きく離れて大き過ぎる、または小さ過ぎる
・ある人が大勢の人と違ってある問題となるグループ（少数派のことが多い）に所属している
このように、ある人がちらばりのある特別なところにいることを把握しなくてはならない。
そのために
まず、データの種類が２つあることを知ることが出発点

■量的データ＝連続変数＝**間隔尺度、比尺度、（順序尺度）**

身長、体重、年齢、血圧やｺﾚｽﾃﾛｰﾙなど生理学的データ、自尊心や抑うつなど多くの心理尺度など

■質的データ＝離散変数＝**名義尺度、順序尺度**＝カテゴリー、分類、グループ、リカートスケール

性別、出生地、疾患名、既往歴、家族歴、配偶関係、職業、学歴、～の有無、はい・いいえ、
とてもそう思う～あまりそう思わない、など

＊「とてもそう思う～あまりそう思わない」などの**順序尺度**は質的データに入れられる場合が多い。
しかし、量的に処理することも多い。**量的にも質的にも**扱ってみるとよい。どちらの性質が強いか？
そもそも、あるものを測定する方法として何が正しいかは考え方の問題（**信頼性と妥当性**）
たとえば、痛みの強さは、どのように測定するべき？量的？質的？hanage の話
→http://www.geocities.jp/kazu_hiro/nurse/hanage.htm

■量と質はものの見方の違い→<u>統計的な扱いの違い</u>が生じるから注目、

・量的データは質的データに変換も可能
年齢（歳）→年齢階級、年齢層別に見ることも多い。青年と高齢者で特徴的な場合もある。
U 字型、 J 字型など、さまざまな質的な関係がありうる＝発見的作業
・質的データも量的（順序尺度として）に見ることができないこともない（無理やり？）
生まれた国→北緯、GNP、人口、血液型→多数派順、日本なら A>B>O>AB で 1,2,3,4 におきかえるとか…
・どちらにして見るかは、あなた次第、説明しやすいよう、説得しやすいよう・・
せっかく量的に測定したものは、まず、量的に扱かってみてから考える
質的に変換すると情報量が減る一方、新しい情報（ある値以上で急激に変化するなど、質的に差が見られる場合もある。しかし、65 歳以上って質的に違うか？）を作り出す場合もある、データの分布をよく見る！

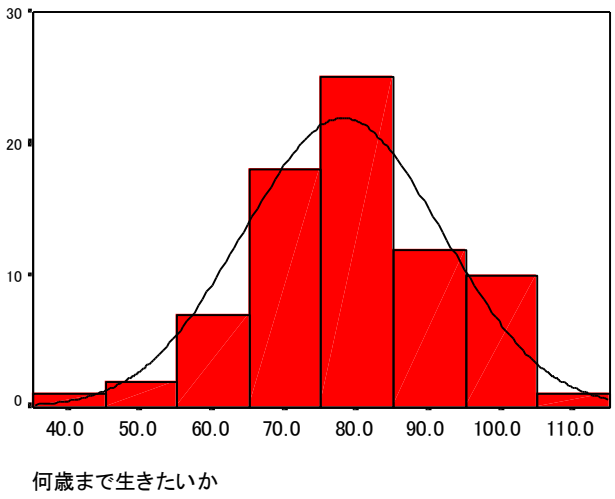
☆簡単に言えば、量的データは<u>数字であらわす意味</u>があり、質的データはそれがない

２）データの分布、ちらばりを説明する＝単純集計

■量的データの測定結果

□視覚的に説明する

分布を視覚的に見るには、ヒストグラムを書いてみる。SPSS　[ｸﾞﾗﾌ]－[ﾋｽﾄｸﾞﾗﾑ]



縦軸は人数で数値を等間隔で階級分けして分布の形をわかりやすいように表現する。分け方を工夫するといろいろなことが見えてくる。正規分布、2 峰性の分布など・・

□数値で説明する

・　どういう値が多いのか→代表値（平均値、中央値、最頻値など）

SPSS　[分析]－[記述統計]－[度数分布表]または[記述統計]

対象者の年齢はだいたいどのくらいだったのか？簡単に説明するには？

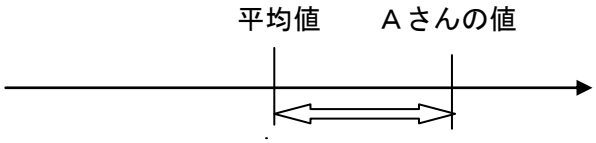
・どのくらいちらばっていたのか→標準偏差、分散、最小値、最大値、範囲など

「こないだ○○のコンサートに行ったよ」「何歳ぐらいの人が来てるわけ？」「だいたい 20 代前半ぐらいかなあ」これは平均とその前後のちらばりを考えて言った発言？

・偏差(deviation)とは観測値と平均値の差。Ａさんは、平均より 5 歳年齢が高い→偏差＝５。

・標準偏差（SD=standard deviation）＝偏差の平均値をあらわすために偏差の 2 乗の平均値の平方根

＝平均して平均値からどのくらい離れているか



Ａさんの偏差　→　これの 2 乗の平均値の平方根が標準偏差（ＳＤ）

・ちなみに、偏差値とは、50＋10×偏差／SD

偏差値 60 とは、平均点 50 より 1SD 高い→平均よりも上の人のなかの平均

・平均０、分散（標準偏差）１にするには、偏差／標準偏差にすればよい→これを**「標準化」**という。どんなものでも比較が可能になる。平均年齢 40 歳で SD が 10 歳なら 45 歳の人は標準化すると 0.5。

・分布のかたちをあらわす歪度（山のピークが左か右に偏る）、尖度（山がとがっている程度）もある。

・正規分布かどうかの検定も可能。SPSS　[分析]－[記述統計]－[探索的]

・飛び離れておかしい値＝外れ値はないか確認

体重 230kg？　　ホント？入力ミス？分析するときこの人を含める？

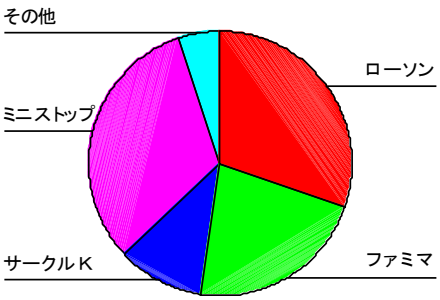
■質的データの測定結果

□視覚的に説明する

ｸﾞﾗﾌを書いてみる　SPSS　[ｸﾞﾗﾌ]－[円]　円じゃなくてもぜんぜん構わない

・まったくの少数派はどの程度いるのか→人数が少なすぎるものは除くか、多数派に含めるか判断

好きなコンビニの割合は？



□数値で説明する

・どういうカテゴリー（分類、グループ）が多かった、あるいは少なかったのか

→割合､%の説明　SPSS　[分析]－[記述統計]－[度数分布表]

「こないだ○○に行ったよ」「やっぱりカップルばっかり？」「9 割以上だと思う。家族連れも少しいたけど」

3）データとデータの関連を説明する

■組み合わせの種類は 3 種類→量と量、質と質、量と質→関連の見方は、基本的に 3 種類しかない！

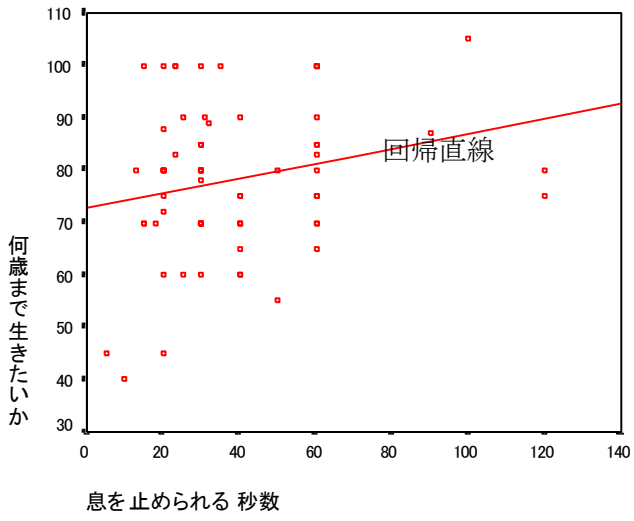
できごとの原因と結果(因果関係)を知り、将来を予測し、よりよい方向へコンロトールするために不可欠。

■量的データと量的データ

□視覚的に説明

・関連を視覚的に見るには、散布図（相関図）を見る。　SPSS　[ｸﾞﾗﾌ]－[散布図]

「息を止められる秒数」が長いほど「何歳まで生きたいか」という年齢が高い？



□関連を統計で説明

・関連の強さを相関係数（ピアソン pearson の r）で知ることができる。SPSS　[分析]－[相関]－[２変量]

線型回帰（直線で予測する）という方法。外れ値注意！

・相関係数は、－1 から 1 の間。0 の場合は相関なし。絶対値が 1 に近いほど相関は強い。

＋は正の相関＝いわゆる正比例、－は負の相関で反比例。

・相関係数がゼロかどうかの検定（無相関の検定）を行なえる。→有意確率に注目、0.05 未満なら有意という。

ただし、有意確率と関連の強さはまったく違うので注意！

ちなみに上の場合、相関係数 r＝0.23、有意確率 p=0.047

・r^2＝決定係数、相互に何%説明可能かを示す。分散の説明。

・変数Ｘと変数Ｙの**相関係数＝ＸとＹの共分散／（ＸのＳＤ×ＹのＳＤ）**

・**ＸとＹの共分散＝（Ｘの偏差×Ｙの偏差）の和／Ｎ**

→共分散が大きい＝一方の偏差が大きい人は、もう一方の偏差も大きい

＝ある人のＸが平均より離れているとき、その人のＹも平均から離れている

＝平均より長く生きたい人は、平均より息を止められる秒数も長い、その逆も

＝ＸもＹもともに変動しているということ=**共変動**ともいう

Ｘの偏差が 1、2、3 でＹの偏差も 1、2、3 としたら、掛け算の組み合わせで一番大きいのは？

1×3＋2×2＋3×1＜1×2＋2×3＋3×1＜1×1＋2×3＋3×2＝1×2＋2×1＋3×3＜1×1＋2×2＋3×3

ともにぴったり変動しているときに共分散が一番大きくなる。

・共分散をＳＤで割っているのは、**標準化**のため＝どんなものでも－１と１の間におさまるように。

回帰直線と相関係数と決定係数の関係とは？　変数が変数を説明するとは？

回帰直線は、Ｙの実際の値と回帰直線によるＸによる予測値（＝ａＸ＋ｂ）の差（＝予測の誤差）を最小にするように算出→**最小二乗法**＝誤差の 2 乗の和を最小にする方法

これにより回帰直線の式は下のように決まる

回帰係数ａ=ＸとＹの共分散／Ｘの分散

定数ｂ＝Ｙの平均値－回帰係数×Ｘの平均値

すなわち

Ｙの予測値＝（ＸとＹの共分散／Ｘの分散）Ｘ＋ＹのＸの平均値－（ＸとＹの共分散／Ｘの分散）×Ｘの平均値

Ｙの予測値をＹの平均値で予測するという方法（平均値は最も誤差の少ない予測値）

→Ｘを何倍かして（回帰係数倍）加えることで、Ｙの予測値をよくする方法への転換

決定係数は、このＹの予測値によってＹの分散がどの程度説明されたか＝Ｙの予測値の分散は、Ｙの分散の何％

説明率＝**Ｙの予測値の分散／Ｙの分散**

＝（Ｙの予測値－Ｙの平均値）^{2}の和÷人数／Ｙの分散

＝（（ＸとＹの共分散／Ｘの分散）×（Ｘ－Ｘの平均値））^{2}の和÷人数／Ｙの分散

＝（ＸとＹの共分散／Ｘの分散）^{2}×（Ｘ－Ｘの平均値）^{2}の和÷人数／Ｙの分散

＝（ＸとＹの共分散／Ｘの分散）^{2}×Ｘの分散／Ｙの分散

＝（ＸとＹの共分散）^{2}／Ｘの分散×Ｙの分散

＝（ＸとＹの共分散／（ＸのＳＤ×ＹのＳＤ））^{2}＝**（相関係数）^{2}＝決定係数**

■統計的検定とは？

帰無仮説＝関連がない、差がない

量×量　相関係数＝０

量×質　平均値の差＝０

質×質　比率の差＝０

起こった現象が、帰無仮説を前提として起こっていると考えると、大変起こりにくい現象になってしまわないか

その起こりにくさの確率を 0.05 未満としている

0.05 未満ならば、そのとき帰無仮説を棄却して

＝０ではないとする

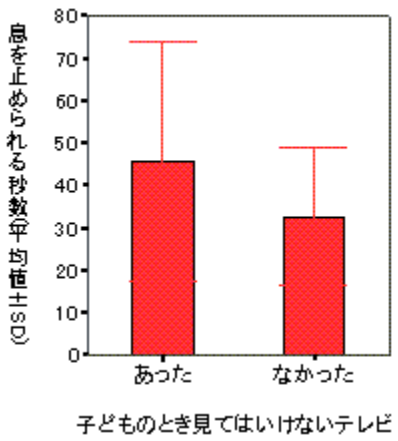
なぜ、有意確率　0.05 なのか　諸説あり・・・・

丁半ばくちで負け続けると・・・・

■量的データと質的データ

□視覚的に見ると

・質的変数におけるグループ(カテゴリー、分類)ごとに、平均値を比較する。SPSS　[ｸﾞﾗﾌ]－[棒]



子どものとき見てはいけないテレビ

エラーバー（標準偏差や標準誤差）をつけたい場合は SPSS　[ｸﾞﾗﾌ]－[ｲﾝﾀﾗｸﾃｨﾌﾞ]－[棒]

「親の方針でテレビの制限」がある→しつけ→がまん強い→「息止め長い」？

□統計的に見てみると数値で示せる（2 グループと 3 グループ以上は違う）

○２グループのとき

・平均値に有意な差があるかを検定するにはT検定を使う。

SPSS　[分析]－[平均の比較]－[独立したｻﾝﾌﾟﾙのＴ検定]

・「等分散性のための Levene の検定」で有意確率が 0.05 より大きいならば、「等分散を仮定する」の行（小さければ「等分散を仮定しない」の行）の「母平均値の差の検定」のところの有意確率を見る。グループの分散が同じかどうかで、ｔ検定の計算が違うから。0.05 よりも小さければ有意な差があるという。ちなみに、上の例では、p=0.021 で有意な差がある。

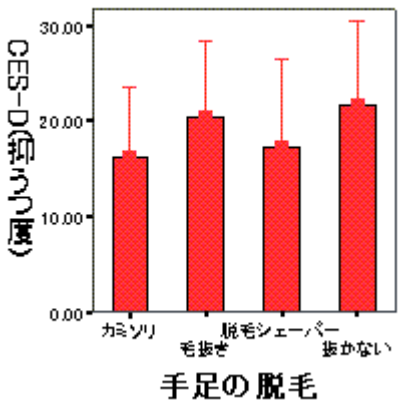
ｔ＝（2 グループの平均値の差）／大きな式（各グループの分散など）

分子は差　結局、差を標準的なものに

・教育の前後での成績の差、手術やケアの前後での QOL の差など、同じ対象者で、介入前後などの２時点のデータがあり、その値の変化を見たいときは、SPSS　[分析]－[平均の比較]－[対応のあるｻﾝﾌﾟﾙのＴ検定]

・サンプルが少なく、分布が正規分布とみなせない場合、SPSS　[分析]－［ノンパラメトリック検定］－［２個の独立サンプルの検定］または[２個の対応サンプルの検定]で Wilcoxon の符号和検定を使う。

○３グループ以上のとき



手足の脱毛

・ペアにしてＴ検定を繰り返すというのはやってはいけない。

1 回の測定に何回も検定をかける→有意になる確率が高まる

・一元配置分散分析を使う。

SPSS　[分析]－[平均の比較]－[一元配置分散分析]

各グループの平均値を比較するために、[オプション]ボタンを押して「記述統計量」をチェックして平均値を出力させる。

・F 検定である。有意確率が 0.05 より小さければ有意な差がある。検定の結果をそのまま言うと、有意ならグループの平均がみな同じだとはいえないということ。すなわち、グループ間で有意な差が 1 つ以上あるといえる。ちなみにこの例は、p＝0.12 で有意な差は認められない。

分散分析のしくみ

ある人の観測値＝全体の平均値＋あるグループに入っている効果＋個人の効果（誤差の効果）

グループの効果＝そのグループに属しているから平均値の高低が生じている

個人の効果（誤差の効果）＝グループ内でもその人だからという平均値の高低が生じている

この 2 つの効果を比較して、グループの効果が有意に大きいか比較している

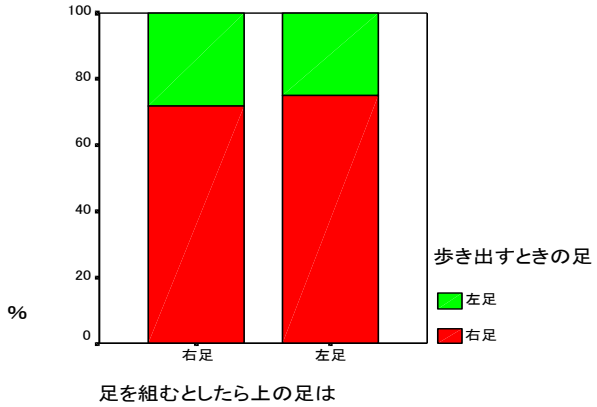
○３グループ以上の場合、分散分析ではどのグループとどのグループに差があるかまではいえない。
→多重比較を行なう　SPSS　[分析]－[平均の比較]－[一元配置分散分析]－[その後の検定]
SCHEFFE や TUKEY などをチェック。すべての組み合わせでペアにして検定を行なってくれる。

・一元配置分散分析のＦ検定でも、多重比較でも、どちらも有意になることが望ましい。
・多重比較は方法がたくさんあるが、自信を持って有意だと言いたいものを探すなら、多くやってみてすべて有意になる方がよい。

■質的データと質的データ

□関連を視覚的に見る

棒グラフを見て割合を比較する。　SPSS　[ｸﾞﾗﾌ]－[棒]



□数値で統計的に説明

・クロス表を作成し、％を確認する。
SPSS　[分析]－[記述統計]－[ｸﾛｽ集計表]
の[セル]ボタンで、「行」OR「列」にチェック。

・関連の強さについては、数値で示すことは多くないが、φ(ﾌｧｲ)係数やクラメール Cramer のＶが１に近いほど強い。

・関連がないかどうかの検定はχ²（カイ２乗）検定。
SPSS　[分析]－[記述統計]－[ｸﾛｽ集計表]の[統計]ボタンで「カイ２乗」をチェック。 Pearson のｶｲ 2 乗の有意確率を見る。P<0.05 ならば、有意な関連がある。

・２×２（四分表）なら、連続修正（イエーツ Yates の補正）の方を見る。この例は全く有意ではない。

・期待度数が５未満のセルが多いか（20%以上など）、最小期待度数が小さい（１未満）ときはカテゴリーの併合を検討する（χ²検定は有効ではない）。２×２なら Fisher の直接確率を用い、セルが多く、どちらかの変数を順序尺度として見ることができるならば、ノンパラメトリック検定を利用することも考える。

SPSS　[分析]－［ノンパラメトリック検定］

χ²検定のしくみ

クロス表において、2 変数にまったく関連がないという状況とは、どのような状況か。

	受講者	非受講者	計
喫煙者			５０
非喫煙者			５０
計	５０	５０	１００

全セルに期待度数が入るとまったく関連がない
セルの**期待度数＝（縦の計×横の計）／全体の計**
χ²＝（そのセルの度数－期待度数）²／期待度数
これを全セルについて合計したものがχ²値
χ²値が大きい＝**ズレが大きい**＝関連強い

まったく関連がない状態＝期待値が入る状態

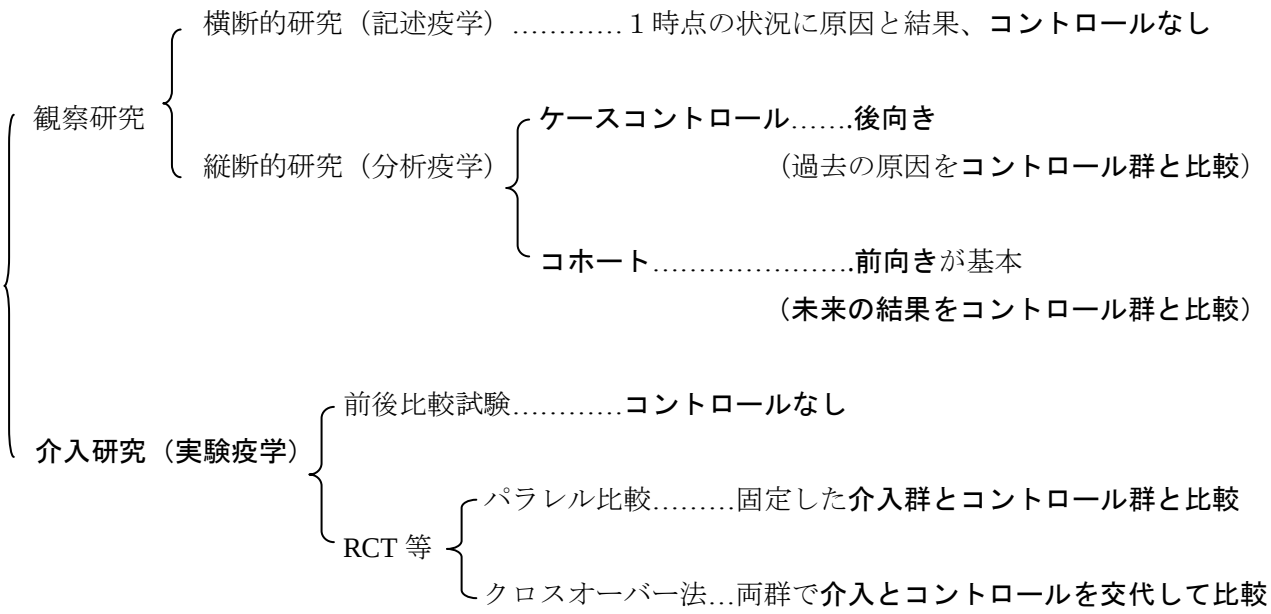
	受講者	非受講者	計
喫煙者	２５	２５	５０
非喫煙者	２５	２５	５０
計	５０	５０	１００

もっとも期待値との**ズレ**が大きい=関連が強い状態

	受講者	非受講者	計
喫煙者	０	５０	５０
非喫煙者	５０	０	５０
計	５０	５０	１００

■研究デザインによる統計手法の選び方（多変量解析以前）

主な（疫学）研究デザイン



1）観察研究　登場した相関係数、t 検定、分散分析、χ^2検定

2）前後比較試験

目的変数＝効果指標＝アウトカムが

（1）**質的データ**　2×2表　マクネマー（McNemar）検定＝χ^2検定　$\chi^2=(|a-d|-1)^2/(a+d)$

（2）**量的データ**　対応のある平均値の差の検定（t 検定）

3）RCT のパラレル比較、ランダム化していない比較試験など、コントロールのある介入研究

（1）**質的データ**　χ^2検定

（2）**量的データ**

・介入前後の２時点

〇差の差の検定＝介入群の前後の差の平均値とコントロール群の前後の差の平均値について t 検定

＝各ケースの前後の差を計算、その値について介入群とコントロール群で平均値を比較

〇共分散分析=介入前の値と後の値による散布図で回帰直線が平行に近い（傾きが大きく異ならない）場合、介入前の値を共変量、介入の有無を主効果とした共分散分析（SPSS 一般線型モデル)。とくに介入前で値が異なる場合は、こちらを行ってみる。介入前の値と介入の有無の交互作用が有意な場合、回帰直線の傾きが有意に異なり、介入前の値の大きさによって介入効果に差があることになる。

☆2 つの方法で結果が異なることもあるので、仮説、サンプリングと割付、交絡因子の確認、グラフで確認！

☆世間になぜかまだある、介入前の t 検定で有意差なし、各群前後の対応のある t 検定で介入群だけ有意差あり、よって介入効果がありました！　は問題あり！

・３時点以上の変化の差を見たいとき

反復測定（repeated measures）による分散分析

SPSS でのデータ変容（変換）

■質的データの変容

□数値を一括して変換する

○得点の逆転　いつも＝4、ときどき＝3、たまに＝2、ない＝1→いつも＝1、ときどき＝2、たまに＝3、ない＝4

○カテゴリーの併合　いつも＝4、ときどき＝3、たまに＝2、ない＝1→いつも、ときどき、たまに＝1、ない＝0

[変換]→[値の再割り当て]→[同一の変数へ]

・[**今までの値と新しい値**]ボタンを押す。左側の「今までの値」と右側の「新しい値」を設定

・1 つ設定するごとに[**追加**]ボタンを押し、「旧→新」のボックス内を確認
　ミスしたら[**変更**]ボタンや[**削除**]ボタンを押して訂正・削除

・終了したら下の[続行]ボタンを押し、[OK]ボタンを押す

・[**変換**]→[**値の再割り当て**]→[**他の変数へ**]にすると、古いのも残せる。新しい「変換先変数」の指定が必要。

□合計得点の計算

○　何とか尺度、何とかスケールで各項目の合計点を出す。

[変換]→[計算]

・新しい変数名を左上の「目標変数」のボックスに記入。右上に「数式」を書き込む。

・合計する関数 SUM を使うと便利　　= SUM(**変数名 1, 変数名 2, ...**)
　またはそのまま足し算　＝**変数名 1 + 変数名 2 + ・・**

□ある値の回答の数をカウント

○複数回答で○のついた数を数える＝○ありが 1 で入力なら 1 という値の個数を数える

[変換]→[出現数の計算]

・新しい変数名を左上の「目標変数」のボックスに記入。

・数える変数を選んだら、[**値の定義**]ボタンを押して、数える値を[**追加**]する。1 と 2 とか複数の数字もＯＫ。

■量的データの変容

○量的データを質的データに変更　年齢→10 歳ごとの年齢階級に分ける、65 歳以上と未満に分ける

[変換]→[値の再割り当て]→[同一の変数へ]もしくは[他の変数へ]

・[**今までの値と新しい値**]ボタンを押す。左側の「今までの値」と右側の「新しい値」を設定

・「今までの値」を「範囲」で指定する。どこからどこまで？○○～○○、最小値～○○、○○～最大値の 3 種。

■グループ別の分析

○グループ別にデータとデータの関連を見る　疾患別にケアとＱＯＬの関連を見る

[データ]→[ﾌｧｲﾙの分割]

・グループ化する変数を選択　「グループごとに分析」をチェック。

・以降の分析は全部グループ別の分析になる。戻すときは、同じところで「全てのケースを分析」をチェックすればよい。

データの入力と保存、Excel とのやりとり

■SPSS のデータエディタを使う

SPSS を起動－「無題　SPSS のデータエディタ」という空白のシートがあらわれる（もし，起動直後に「どのような作業を行いますか？」と聞いてきた場合は、「データに入力」を選ぶ。）。

□変数名の入力

表の左下のタブの「変数ビュー」をクリック。「名前」の欄に変数名を入力。

・「名前」の付け方の決まり＝半角８文字（全角であれば４文字）以内、半角数字で始めることはできない、ピリオド（.）で終わってはいけない、スペースや特殊文字（－　！　？　'　*）などは使えない（とくにハイフン（－）を使いたくなるが利用できないので注意）、大文字と小文字は区別しない

□データの入力

すべての名前の入力が終わったら、「データビュー」タブをクリックし、データ入力画面に。一番上の欄には変数の名前が並んでいるのが見える。数字は点キーと矢印キーで入力。１と入れても 1.00 となるが気にしない（同じこと）。無回答や非該当は矢印キーでとばす。ピリオド（.）が残る。

□データの保存

メニューの「ファイル」から、新規の保存の場合であれば「名前を付けて保存」を、同じ名前で保存するのであれば「データの上書き保存」を選ぶ。「名前を付けて保存」の場合はファイル名を付けるように要求されるので、名前を付ける。名前の最後には sav が自動的に付加される（拡張子と言う）。

□Excel への書き出し

「ファイル」－「名前を付けて保存」を選び、ファイルの種類を「Excel（*.xls）」にして、「保存」。

■Excel を使う

□変数とデータの入力

第１列目に変数名を，２列目以下にデータを入力。1 列目以外は SPSS のデータエディタを使う場合と同様。変数名は SPSS の「名前」の付け方の決まりを守ること。無回答や非該当は空欄にする。

□データの保存方法

通常の保存でＯＫ（拡張子が xls）。ただし、データが複数のシートにわたる場合は入力されている画面を表示しておくこと。シートごとに保存が必要。

□SPSS への取り込み

SPSS のメニューの「ファイル」－「開く」－「データ」を選択。「ファイルの場所」を指定し，「ファイルの種類」を「Excel(*.xls)」にする。取り込みたいファイルが出てきたらそれを選択し「開く」をクリック。つぎの画面で「データの最初の行から変数名を読み込む」にチェックしてＯＫを押す。取り込んだ後は「変数ビュー」で変数の数を確認し、新規で保存。