「3」も「飲酒習慣」に替えておきましょう（図12）。その上で、「飲酒分類」をフィールドリストにドラッグして消します。同じように「高血圧判定」もグループ化を図り、1 を「正常血圧」2-4 を「血圧異常」とします。「高血圧判定 2」を作り、もとの「高血圧判定」を消します。これで飲酒と血圧のクロス集計が完成しました。ページの選択で、男女の結果（図13）、男女別の結果（図14、15）が得られます。

性別	(すべて)		
データの個数 / 年齢	高血圧症判定2		
飲酒分類2	正常血圧	血圧異常	総計
非飲酒習慣	678	616	1294
飲酒習慣	147	148	295
(空白)	2	3	5
総計	827	767	1594

図13　男性＋女性の結果

この結果を見ますと、飲酒習慣のない人たちでは、正常血圧の人が異常のある人より多いようです（678>616）。そして、飲酒習慣のある人たちでは、正常血圧者と血圧異常者が拮抗しています（147 vs 148）。しかし、これは意味のある、つまり飲酒習慣のない人では血圧が正常の人が多い、と断言していいのでしょうか。そう言えるためには統計学的な検定が必要になります。これは次節で分析します。

性別	1		
データの個数 / 年齢	高血圧症判定2		
飲酒分類2	正常血圧	血圧異常	総計
非飲酒習慣	163	124	287
飲酒習慣	122	134	256
(空白)	1	1	2
総計	286	259	545

図14　男性の結果

性別	2		
データの個数 / 年齢	高血圧症判定2		
飲酒分類2	正常血圧	血圧異常	総計
非飲酒習慣	515	492	1007
飲酒習慣	25	14	39
(空白)	1	2	3
総計	541	508	1049

図15　女性の結果

男性の結果を見ると、飲酒習慣のない人たちでは、正常血圧の人の方が多く、逆に飲酒習慣のある人たちでは、血圧異常の人の方が多い、と出ています。これも統計的に意味があるかどうかが問題です。

女性の結果では、習慣の有無にかかわらず正常血圧の人が多いようですが、飲酒習慣者がかなり少ないため、比較に意味があるのか疑わしそうです。

いずれにしても、このようにピボットテーブルを用いて分析することで、いろいろ考えさせられる結果を得られることが分かります。また、地域の健康指標について自分たちで思いついた考えを数字の上で確かめてみることができるでしょう。

＊　＊　＊　＊　＊　＊　＊　＊

ピボットテーブルの使い方はこのような例題の他、はるかに多様ですが、一つだけ補足しておきます。

例題では2要因についての頻度をみるため、図8でデータアイテムに持ってくる項目は空白セルがなければ「何でもよい」と言いましたが、ここに測定値（体重や検査値など測定できる値）を持ってきて、平均値などを見ることができます。

〔例題3〕 飲酒習慣のある人とない人では、食生活も異なる可能性があります。例えば中性脂肪値が異なるのでしょうか。

〔解説〕

中性脂肪値という測定値をデータアイテムとすると、例えば飲酒習慣という要因別に平均値を見ることもできます（図16－17）。

図13－15の状態から、列フィールド（「高血圧症判定 2」）とデータアイテム（ID 番号）をフィールドリストにドラッグして消し、代わりに中性脂肪をデータアイテムとします。データアイテム

領域は「合計表示」になっているので、「平均値」に替えます。

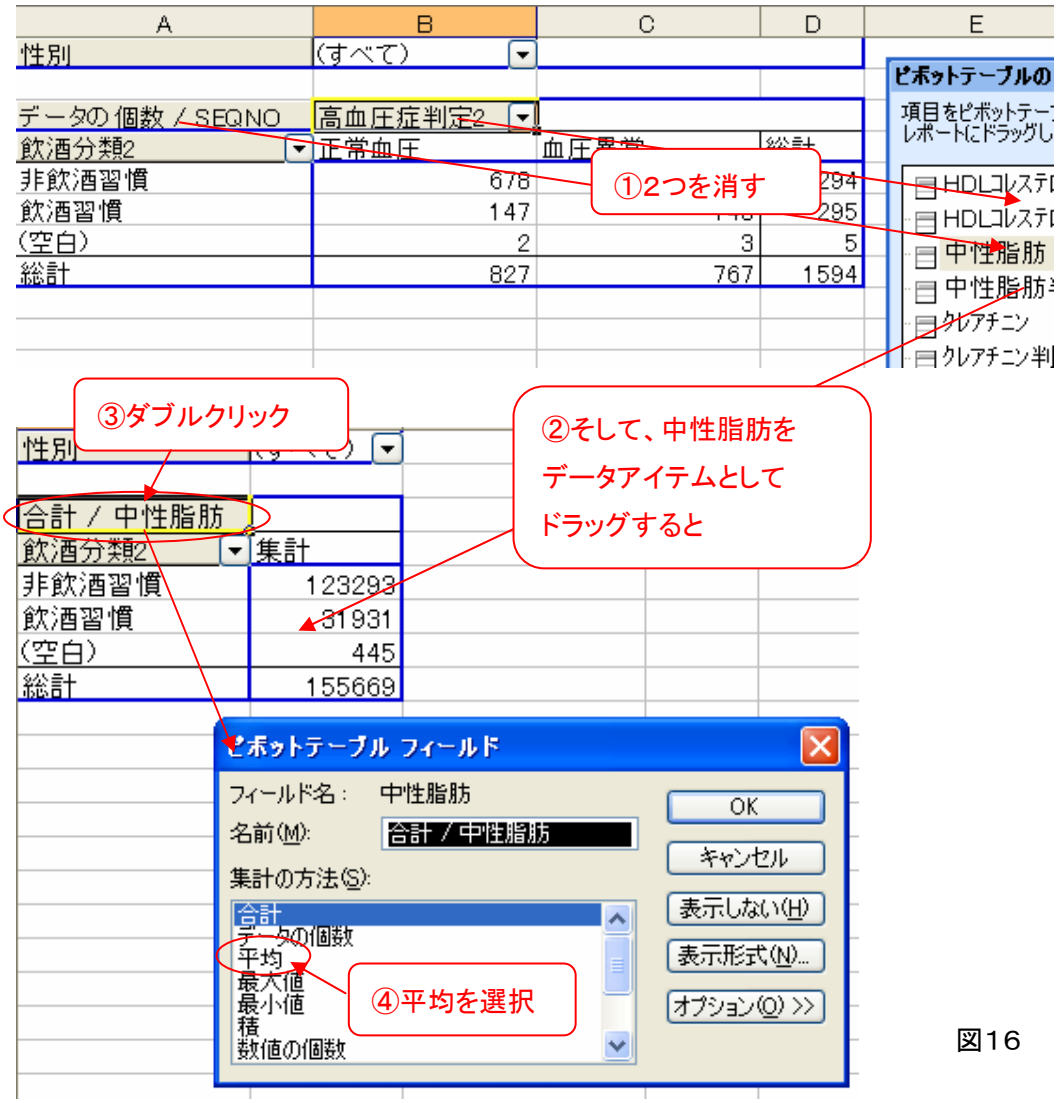

図16

性別	(すべて)
平均 / 中性脂肪	
飲酒分類2	集計
非飲酒習慣	95.3
飲酒習慣	108.2
(空白)	89.0
総計	97.7

図17

図 16 がその結果です。これを見ると飲酒習慣者の方がそうでない人に比べ、中性脂肪値

が高いように見えます。ただし、これも検定が必要です。

今までのピボットテーブルの使い方をまとめますと:

・列や行、またページのフィールドでは男女、喫煙習慣、飲酒習慣、疾病の判定など、分類できる項目・要因を持ってきます。BMI や検査値などの測定値を列・行・ページに持ってくると列などが多くなりすぎて適切ではありません。ただし、グループ化すれば行や列に適すものになりえます。

・データアイテムとしては、列や行の要因別の度数だけ見るのであれば何を持ってきても構いません。ただし、空白のない項目を使います。

・データアイテムとして、測定値を持ってくることで、要因(列や行)毎の平均値や最大値などの統計量を比べることができます。

・いずれにしても、これらで得られた要因差が確かにありそうだと言えるかは統計的な検定をしなければいけません。ピボットテーブルだけではそこまで言えません。

(3)統計学的検定

ピボットテーブルを用いた分析等で、ある要因に関し集団間の差異が得られても、それが統計的に有意かどうかは、検定をしてみなければ分かりません。検定法そのものは専門書に譲りますが、例えば図14では以下のような 2x2 の表が得られました。このような結果が得られることは非常に多く、度数の独立性検定(χ^2 検定)をできるようにしておくことは大変有意義です。

性別	1		
データの個数 / 年齢	高血圧症判定2		
飲酒分類2	正常血圧	血圧異常	総計
非飲酒習慣	163	124	287
飲酒習慣	122	134	256
(空白)	1	1	2
総計	286	259	545

図 14(再掲)

幸いエクセルは基本的な統計学的検定をサポートしていますので、χ^2 検定をエクセルの関数を用いて行ってみましょう。

図18で説明します。まず、図14の 2x2 の数字を合計も含めコピーし、適当な場所に(新しいシートなど)貼り付けておきます(「実測値」と表題を付けておくと良いでしょう)。その下に 2x2 の期待値を算出しておきます(図の 19 行と 20 行)。15、16 行は算出のための式を載せていますが、実際にはこの式を打ち、[Enter]を押しますと、19、20 行の数字が出るのです。期待値は実測値の計から算出します。そして、22 行の関数式によりχ^2 検定により p 値を算出し有意水準(0.05 とするのが一般的です)と比較し、それより小さいと、差が統計的に有意

ということになります。ここで用いた関数式は 22 行の赤字の部分ですが、この式を実際に打って確定([Enter]を押す)しますと 23 行の 0.033 になります。22 行は参考で、実際にはこのような行はでません。

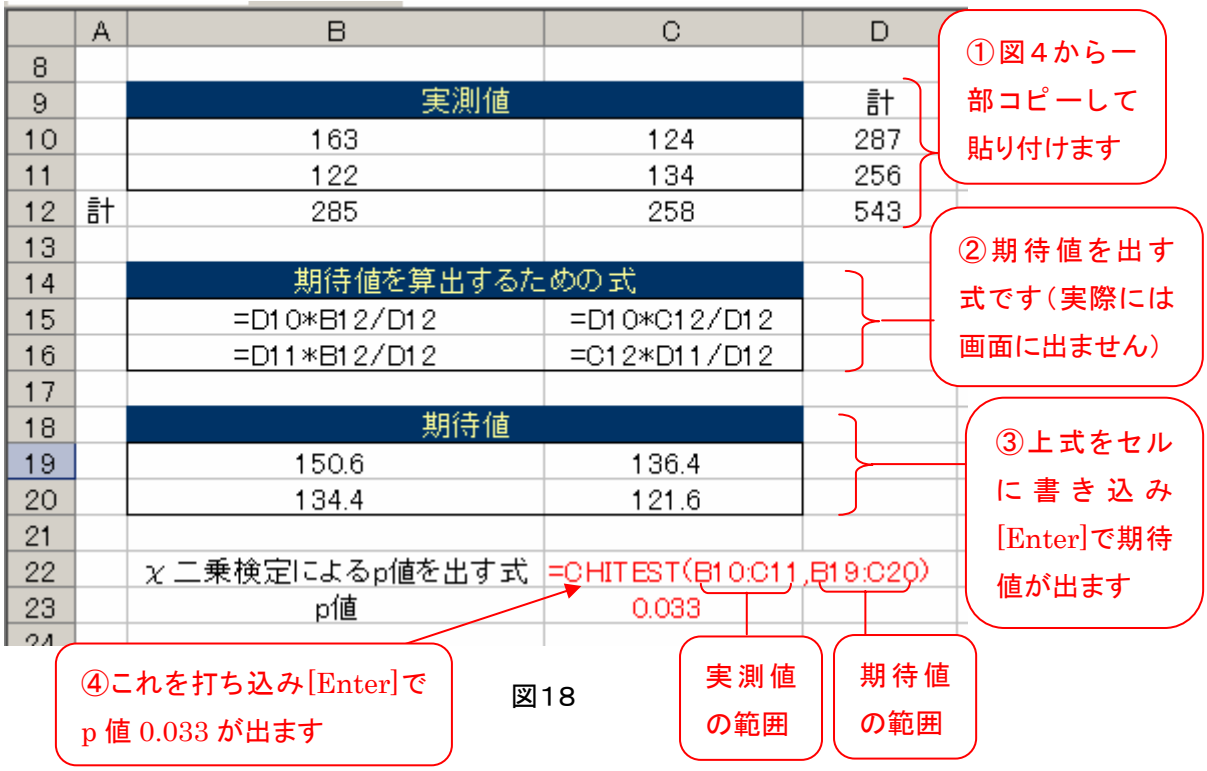

図18

5%水準で差は有意と認められたということは、つまり、男性の場合、飲酒習慣のある集団はそうでない集団に比べ、血圧異常を生じる頻度が多いということです。

女性の場合はどうでしょう(図15)、男女コミ(図13)ではどうでしょう、ぜひ χ^2検定を行ってみて下さい。

エクセルは χ^2検定以外にも、t 検定(統計関数「TTEST」)、相関係数(同「CORREL」)などをサポートしており、これだけでも非常に有益です。適宜使っていただきたいものです。例えば、図 16 で示しましたように飲酒習慣者の方がそうでない人に比べ、中性脂肪値が高いように見えます。これは t 検定を行うことで、有意差かどうか分かります。ここではお示ししませんが、やってみようと思う人は TTEST を使って検定してみて下さい。この場合は生の測定値を用い、オートフィルタを活用します。

エクセルで統計学的検定を使うのはやや煩雑で、本格的には SPSS などの統計のソフトを用いた方が良いと思います。統計専用のソフトとは別に、エクセル上で用いる(アドイン・ソフトと言います)検定ソフトもありますので、利用するのも良いでしょう。

4. 最後に

今回提示した健診データの分析はほんの一例に過ぎません。埋めていけば何かしら結果の出るツールはこれまでもありますし、今後も医療費適正化計画の関連で大変多く世に出されるものと思います。しかし、自分の問題意識に従って、実際にデータを扱いながら考えることの意義は大きなものがありますし、それは今後も変わらないでしょう。ぜひデータ分析にトライしてほしいものです。

（参考）

活用した健診機関のオリジナルの基本健康診査結果は、列数が 275 という仕様になっています。エクセル（Excel 2003 までのバージョン）では列に 256 という制限があり、そのままでは不完全にしか読み込めません。従って不要と思われる列を削除して列数を 256 以下にしたいわけですが、CSV ファイルをテキストとして開いて削除するのは殆ど不可能です。

そこで最初に CSV を表形式で表示してくれるソフト（例えばフリーソフトの Cassava Editor http://www.vector.co.jp/soft/winnt/business/se162309.html で入手可能）を用いて列が 256 以下になるように修正します。それが可能なのは、実は予備用など不要な列が結構あるからです。

〔Cassava 等 CSV ファイルを表形式で表示するソフトを使用したデータの修正法〕

最初に項目名ファイルを読み込み、不要と思われる項目を選んでおく（不要とした項目番号を記録しておく）。次いで、新たにデータファイル（これには一般的に項目名はありません）を読み込み、不要として選んだ項目の列を1列ずつ削除する。15 列（以上）削除し、任意の名前を付けて保存して終了する。これで、エクセルで使える準備が整ったことになります。