Microsoft Excel を用いた演習の手引き

作成日： 2025 年 3 月 1 日
作成者：尾﨑米厚・金城文

● **演習用データの内容**

500 人（男性 160 人，女性 340 人）の健康診査の結果および生活習慣についての仮想データ（Microsoft Excel ファイル）です。

● **分析ツールの設定**

Microsoft Excel（以下，エクセル）の分析ツールを用いて統計学的な検討を行うには，まず「ファイル」→「オプション」→「アドイン」から，「分析ツール」をアクティブにしておく必要があります。

A データの種類と分布

① 度数分布表とヒストグラムの作成

量的データの場合，エクセルを用いて度数分布表をつくるには，以下のように行う（演習用データによる分析例）。一度にうまくいくとは限らないので，何度か挑戦して，そのデータにあった集計ができるようになるとよい。

a データの値の範囲を調べる（体重の集計の例）

①マウスのクリックとドラッグを用いて，集計したい値が入っている範囲を指定する（E2 から E501 の範囲；広い範囲を指定する時は範囲の最初のセルをクリックし，最後のセルにカーソルを移動して「Shift」キーを押しながら範囲の最後のセルをクリックすると，その間の全体を指定できる）。指定した範囲のセルは，異なる色になる。

②リボン（画面上部のさまざまなボタンが配置されている領域）の「数式」タブにある Σ（シグマと読む）の記号のプルダウンメニューに「最大値」「最小値」とあるので，それを順次クリックすると，データが入っている欄外に最大値と最小値が計算されて出てくる（最大値は，E501 の下に 99，最小値は，34 と出る）。度数分布表は，区切りのよい数値のところで，データを 10 前後から 20 区分くらいに分けると分布の様子がわかりやすいといわれる。

演習データの場合，体重の最大値は 99kg［関数の計算式では＝MAX（E2：E501）］，最小値は 34kg［＝MIN（E2：E501）］となる。したがって，データの範囲は，（最大値）－（最小値）＝99－34＝65 となる。今回は 5kg ごとの階級に分けて分析してみよう。なお，体重は明らかに男女差があると考えられるので，男女別に集計してみる。

b 階級ごとに集計する

①男女別に集計するには，まずデータを性別に並びかえる必要がある。先ほどと同様に範囲を指定し，「データ」→「並べ替え」で「最優先されるキー」を「性別」にする。
②表（ワークシート）のデータが入っていない場所に，区間を区切る値（30 以上 35 未満なら 35）を順番に入れていく。
③「データ」→「データ分析」を選び，そのなかの「ヒストグラム」を選択する。
④「入力範囲」のところで，集計したい範囲を指定し，「データ区間」のところで，先ほどつくった区間を区切る値が書かれている範囲を指定する。
⑤「グラフ作成」にチェックを入れて，「OK」をクリックすれば，度数分布表とヒストグラムが表示される（結果はデータとは別のシートに出る）。135 ページの**図 8-2** は男女をまとめたヒストグラムをつくるために，男女の集計結果を 1 つの表にまとめてから，「挿入」→「グラフ」で縦棒グラフにしたものである。

c 結果をながめる

男と女の異なる集団がそれぞれ，異なる平均値とばらつきをもって，しかし同じような山のような形をした分布をしていることが分かる。男女ともほぼ，体重の分布は正規分布に近いといえる。これにより，それぞれのだいたいの平均値やばらつきがわかる。さらに，この解釈を通して，このデータの代表値で適切なものは何かを決めることができる。

男女の対象者数が異なるために分布の違いがわかりづらいような場合は，相対度数（総数に対するそれぞれの階級の割合）を用いてヒストグラムを書いたほうがわかりやすい。

② 基本統計量の算出

これらの統計量を，エクセルでは，即座に計算してくれる。

①データの入力されているシートにおいて，「データ」タブの「データ分析」を選び，データ分析の選択肢のなかから，「基本統計量」を選び，「OK」をクリックする。
②「入力範囲」の右側のボタンをクリックして，その後，集計したいデータが入力されている範囲を指定する（男は E2：E161，女は E162：E501）。
③データが縦方向の「列」，横方向の「行」のいずれに並んでいるかで，どちらかにチェックを入れる（演習データであれば「列」）。
④出力オプションの「統計情報」にチェックを入れて「OK」とすれば，基本統計が計算されて出てくる（**表 1**）。

例題データの体重を男女別に集計すると，平均値，中央値，標準偏差などが計算される。男女では，平均体重は異なるが，標準偏差はあまり異ならない。したがって，ばらつきの度合いは似通っているといえる。また，変動係数は，男性 0.15，女性 0.16 となり，標準偏差は男性のほうが少しだけ大きいが，変動係数でみると，男女ほぼ等しい。

表 1　基本統計量の算出

男の体重		女の体重	
平均	63.28313	平均	53.23765
標準誤差	0.751483	標準誤差	0.472422
中央値（メジアン）	62.4	中央値（メジアン）	52.3
最頻値（モード）	60.5	最頻値（モード）	52.4
標準偏差	9.505597	標準偏差	8.711024
分散	90.35638	分散	75.88194
尖度	0.895696	尖度	1.629797
歪度	0.593127	歪度	0.884971
範囲	56.1	範囲	52.2
最小	42.9	最小	34
最大	99	最大	86.2
合計	10125.3	合計	18100.8
データの個数	160	データの個数	340

③ 分位数の決定

分位数の決定もエクセルを用いて検討できる。
①エクセルの「データ」タブから「データ分析」→「順位と百分位数」を選び，OK とする。
②次の画面で，入力範囲を指定して OK とすると，データが大きい順に並び，パーセンタイルも表示される。

たとえば，データの「収縮期血圧」を集計してみよう。前述の方法で，男女含めて 500 人のデータに順位をつける。たとえば，四分位数をみつけるには，第 1 四分位数は，25%目がある 113 と 114 の間（113.5 mmHg），第 2 四分位数（中央値）は，50%目がある 127 と 128 の間（127.5 mmHg），第 3 四分位数は，75%目がある 138 と 139 の間（138.5 mmHg）である。

④ 正規分布の値の取得

エクセルでは，＝NORMDIST（ある値，平均値，標準偏差，false）という関数を用いると，その平均値・標準偏差をもつ正規分布において，ある値が発生する確率が計算される（正確には確率密度が計算される）。

B 関連の指標

① 散布図の描画

エクセルで散布図を描くには，「挿入」タブで「散布図」を選び，2 つの変数が対応して入力されている範囲を指定するとよい。

たとえば，演習データの，腹囲と収縮期血圧で散布図を描くと，139 ページの**図 8-4** のようになる。散布図を描くとおおよその関係がわかる。また，外れ値がみつかれば，実際の値か，入力ミスなどの間違いによるものか調査票に戻り確認する必要があることもある。これは，1 つの外れ値のために大きな相関係数を示す場合があるからである。

2 つの変数の関連を検討したいときがある。2 つの数量データの関連を分析する方法は相関と回帰であり，2 つのカテゴリデータの関連をみるのがクロス集計による分析である。

② 相関の分析

相関係数も分析ツールを用いて計算できる。

①「データ」→「データ分析」→「相関」→「OK」と進み，データの範囲をドラッグする（データが対になっている場所を指定）。

②演習データでは，データの方向は「列」なので，「データの方向」は，列にして「OK」をクリックすると相関係数が計算される。

上の例だと，0.30 という相関係数（ピアソンの相関係数）が出てくる（弱いが相関がありそうである）。この分析ツールでは，3 つ以上の対応したデータセットでも，範囲を指定すれば，その中にある 2 つの変数のすべてのペアに対する相関係数を一度に計算してくれる。

③ 回帰分析と回帰直線の描画

エクセルの分析ツールにも単回帰分析は提供されている。「データ」→「データ分析」→「回帰分析」と進むと，データの範囲が出てくるので，対になっている 2 つのデータの範囲を指定すればよい（**表 2**）。

y を収縮期血圧，x を腹囲として計算すると，結果の 3 つ目の表の左に切片と傾きの結果が出てくる（下の「概要」参照）。切片 78.2，傾き（X 値 1 と記載されている）0.60 となる。腹囲が 1cm が増えれば，血圧が 0.60 mmHg 高くなる関係であることを示す。したがって，ある腹囲を与えれば，その人の収縮期血圧値を予測できることになる。傾きの t 値が 7.15 で p 値が 0 にきわめて近いので，この傾きは統計学的に有意に 0 ではないといえる。

1 番上の表の 3，4 行目の重決定 R2 は決定係数とよばれ（この場合 0.09），この単回帰のモデルで実際のデータの 9%を説明できていることを示す。けっして高い値には見えないが，低すぎるわけでもない。腹囲を測定するだけで，その人の収縮期血圧が，ある程度予測できることになる。

また，先に描いた散布図（119 ページ，図 8–4）のデータがプロットされた点の集まりあたりをクリックして指定し，その後右クリックすると表示されるメニューで，「近似曲線の追加」→「線形近似」を選ぶと，散布図に回帰直線が加わる。オプションで「グラフに数式を表示する」と「グラフに R2 乗値を表示する」をチェックしていると，さらに回帰式と R2 乗値も書き加えられる。

表 2　単回帰分析

回帰統計	
重相関 R	0.305
重決定 R2	0.093
補正 R2	0.091
標準誤差	17.372
観測数	500

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	1	15408.23	15408.23	51.06	0.00000000
残差	498	150289.58	301.79		
合計	499	165697.81			

	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	78.197	6.856	11.405	0.00000000	64.73	91.67	64.73	91.67
X 値 1	0.597	0.084	7.145	0.00000000	0.43	0.76	0.43	0.76

④ クロス集計表の作成（ピボットテーブル）

たとえば，喫煙状況と飲酒状況との関連をみるとしよう。

①まず，「挿入」タブの「ピボットテーブル」をクリックすると，「ピボットテーブルの作成」という画面が真ん中あたりにあらわれるので，「範囲」でデータの範囲を指定する（ワークシートすべてを指定すればよい）。

②ついで，「完了」をクリックすると，空白の表と，「ピボットテーブルのフィールドリスト」というボックスが右にあらわれる。

③喫煙と飲酒のボックスをチェックし，片方を「列ラベル」，もう片方を「行ラベル」にドラッグする（どちらが列でもいい）。

④下の Σ 値に「データの個数」となるように，「性」などすべてのデータにブランクがないような変数にチェックを入れてドラッグすると，左の広いところに**表 3** のような表がつくられる。

表 3　クロス集計表

	毎日飲酒	ときどき飲酒	飲酒なし	総計
喫煙なし	234	122	69	425
	55.1%	28.7%	16.2%	100.0%
喫煙あり	22	23	30	75
	29.3%	30.7%	40.0%	100.0%
総計	256	145	99	500

C 統計分析

① 信頼区間の推定

信頼区間の推定の際に，標準誤差（SE）にかける係数もエクセルで計算できる。たとえば，95％信頼区間だと，正規分布の上側確率が 0.025（5％の半分）であるため，＝NORMSINV（0.975）という関数を書くと 1.96 という値が表示される。平均値＋1.96×標準誤差より大きい値は 2.5％しかない，すなわち平均値±1.96×標準誤差の区間から外れる値は 5％しかないのである。この間に 95％の確率で母平均が存在するであろうと考えられる。

男女別の体重のデータを用いて信頼区間を計算すると**表 4** のようになる。平均値と標準誤差は，基本統計量の計算結果（付録 3 ページ，**表 1**）から抜き出して入れた。信頼の度合いが増す（信頼区間のパーセントが大きくなる）と上限と下限の幅が広くなっているのがわかる。

表 4　信頼区間

	平均値	標準誤差	90%信頼区間		95%信頼区間		99%信頼区間	
			下限	上限	下限	上限	下限	上限
男	63.28	0.75	62.04	64.52	61.81	64.76	61.34	65.22
女	53.24	0.47	52.46	54.02	52.31	54.16	52.02	54.46

② 独立性の検定（χ^2 統計量の評価）

先述したように，エクセルを用いてクロス集計表をつくるには，ピボットテーブルを使用する（付録 5 ページ）。

残念ながら，エクセルでは独立性の検定（カイ 2 乗検定）を実施する項目がないが，計算された χ^2 統計量の大きさが統計学的に有意差があるかどうか（帰無仮説を棄却できるかどうか）を判定することは，エクセルの関数で可能である。＝CHIDIST（χ^2,自由度）で p 値がでるので，5％の危険率で検定するのなら，$p<0.05$ で統計学的に有意に関係がある

（帰無仮説の棄却）となるし，1%の危険率で検定するなら，$p<0.01$ で有意に関係があるといえる。おもな χ^2 検定をするための集計表をデータファイルの別ワークシート「χ^2 検定」に作成してあるので，ご利用いただきたい。

③ 対応のある場合の *t* 検定

エクセルでは，＝TDIST（*t* 統計量，自由度，1 または 2）で *p* 値が出る（*t* 分布を使用）。最後の 1 または 2 には片側検定で 1，両側検定で 2 を入れる。通常 2 を使う。また，これらのプロセスを計算せずとも，エクセルの分析ツールを用いれば，一気に計算してくれる（**表 5**）。

①まず演習データの「メタボ判定」（C 列）を降順に並べかえると，メタボ判定が 3 の人（メタボリックシンドロームありの人）のみ，「介入後の腹囲」が記載してある（AB 列）。これは，メタボリックシンドロームの人に対して 12 週間の運動と栄養を中心とした介入を行ったあとの腹囲である。

②メニューの「データ」→「データ分析」→「*t* 検定：一対の標本による平均の検定」を選ぶ（「OK」をクリックする）。

③変数 1 の範囲を指定し（この場合は腹囲の値の 2 行目から 54 行目〔G2：G54〕），変数 2 の範囲も指定する（この場合は介入後の腹囲の値の 2 行目から 54 行目〔AB2：AB54〕）。

④仮説平均との差異のボックスには 0（ゼロ）を入れて，OK をクリックすると結果が計算される。仮説平均の差異を 0 にするのは，帰無仮説は，両群の平均値に差がない（差が 0）と仮定しているからである。

介入前（変数 1）の平均値 94.3cm，介入後（変数 2）の平均値 92.5cm であり，統計量（*t*）は 5.93 と算出される。5%の有意水準での *t* 値（両側検定）である 2.01 よりはるかに大きく，*p* 値も 0 に非常に近い値なので，統計学的に有意に腹囲の平均値が減少したといえる。このように，対応のある平均値の差の検定は，健康教室など介入の評価に用いることができる。

表 5　対応のある場合の *t* 検定

	変数 1	変数 2
平均	94.26037736	92.54716981
分散	35.29974601	38.06023222
観測数	53	53
ピアソン相関	0.94026285	
仮説平均との差異	0	
自由度	52	
t	5.925005914	
P(T<=t) 片側	0.00000013	
t 境界値 片側	1.674689154	
P(T<=t) 両側	0.00000025	
t 境界値 両側	2.006646805	

④ 分散が等しい場合の *t* 検定

エクセルでは，＝TDIST（*t* 統計量，自由度，2）と入れると *p* 値が出る。対応がある場合の検定と同様，これらのプロセスも，エクセルの分析ツールを用いれば，一気に計算してくれる。男性の喫煙の有無別にγ-GTP の値を比較するとしよう（**表 6**）。

①まず演習データの「性別」（A 列）および「喫煙」（X 列）を昇順に並べかえる（性別，喫煙状況別にデータを並べかえる）。

②「データ」→「分析」→「データ分析」→「*t* 検定：等分散を仮定した 2 標本による検定」を選ぶ（「OK」をクリックする）。

③変数 1 の範囲をドラッグして指定し（この場合は男性非喫煙者のγ―GTP 値である 2 行目から 113 行目〔P2：P113〕），変数 2 の範囲もドラッグして指定する（この場合は男性喫煙者のγ-GTP，114 行目から 161 行目〔P114：P161〕）。

④仮説平均との差異のボックスには 0（ゼロ）を入れて，OK をクリックすると結果が計算される。仮説平均の差異を 0 にするのは，帰無仮説は，両群の平均値に差がない（差が 0）と仮定しているからである。

非喫煙者のγ-GTP の平均値は，41.9，喫煙者のそれは，95.3 である。自由度は 158，*t* 統計量は－3.23，両側検定の *p* 値は，0.002 で帰無仮説が棄却される。この結果，喫煙者のほうがγ-GTP の値がわるいことが示された。

表 6　等分散を仮定した *t* 検定

	非喫煙者	喫煙者
平均	41.875	95.3125
分散	1427.461712	27581.11303
観測数	112	48
プールされた分散	9207.345332	
仮説平均との差異	0	
自由度	158	
t	-3.228111706	
P(T<=t) 片側	0.00075772	
t 境界値 片側	1.654554875	
P(T<=t) 両側	0.00151544	
t 境界値 両側	1.975092073	

⑤ 分散が等しくない場合の *t* 検定

まず分散がほぼ等しいかどうかの判断を，「データ分析」の「*F* 検定：2 標本を使った分散の検定」によって行う。この例だと *F* 検定を行うと *p*＝0.00 となり（**表 7**），等分散とはいえないので，分散が等しくない場合の検定が必要である。

次に，分析ツール内メニューの下の「分散が等しくないと仮定した 2 標本による検定」

を行うと，t 統計量−2.20，p 値 0.03 となり，帰無仮説は棄却される（**表 8**）。分散が等しいと仮定した場合と，結果の解釈は変化しなかった。

表 7　*F* 検定

	変数 1	変数 2
平均	41.875	31.875
分散	1427.461712	2708.196809
観測数	112	48
自由度	111	47
観測された分散比	0.527089356	
P(F<=f) 片側	0.003244367	
F 境界値 片側	0.677971457	

表 8　分散が等しくないと仮定した *t* 検定

	非喫煙者	喫煙者
平均	41.875	95.3125
分散	1427.461712	27581.11303
観測数	112	48
仮説平均との差異	0	
自由度	49	
t	-2.204941194	
P(T<=t) 片側	0.016091293	
t 境界値 片側	1.676550893	
P(T<=t) 両側	0.032182585	
t 境界値 両側	2.009575237	

⑥ 無相関の検定

エクセルの「データ分析」の「相関」では，相関係数しか計算されず有意かはわからない。エクセルでは「回帰分析」を用い，片方の変数の範囲を Y 範囲，もう一方の変数の範囲を X 範囲として分析を行えば，X の p 値として有意確率が算出される（付録 5 ページ，**表 2**）。

⑦ 同順位がある場合の *U* 検定

演習データから，男性におけるメタボリックシンドロームと飲酒量との関連を，ピボットテーブルを用いて作成した（**表 9**）。

飲酒量分類には，順序があるが，その度合いを数値にはできない。また，集計結果も正規分布しているようには見えない。メタボなし群とメタボ＋予備群の 2 群の間で，飲酒量に差があるかどうかについての検定をするには，同順位のデータがある（同じカテゴリは同じ

順位になる）場合の U 検定を行うことになる（通常，$\chi 2$ 乗検定が行われるが，U 検定も可能である）。帰無仮説は「2 群の飲酒量の平均順位に差がない」である。飲酒量が多い方を順位が高いとすれば，それぞれの平均順位は，以下のようになる。

・2 合以上：((20＋8) ＋1) /2＝14.5 位
・1～2 合未満：28＋ ((35＋21) ＋1) /2＝56.5 位
・1 合未満：28＋56＋ ((30＋19) ＋1) /2＝109 位
・飲酒なし：28＋56＋49＋ ((23＋4) ＋1) /2＝147 位

よって，メタボなし群の順位和＝14.5×20＋56.5×35＋109×30＋147×23＝8918.5 となり，メタボ＋予備群の順位和＝14.5×8＋56.5×21＋109×19＋147×4＝3961.5 となる。順位和が小さいほうの群を選び，順位和の期待値を計算すると，52× ((108＋52) ＋1) /2＝4186 となる（両群の標本の平均順位が等しいと仮定した場合）。分散は，(108×52) / (12×160 (160－1)) × [(1603－160) － {(283－28) ＋ (563－56) ＋ (493－49) ＋ (273－27)}] ＝69553.03 となるから，検定統計量は，z＝ (｜4186－3961.5｜) /＝$\sqrt{}$ 69553.03＝0.85 と，1.96 より小さくなるので (p＝0.40)，5%の有意水準で「有意差あり」とはいえない（帰無仮説を棄却できない）。すなわち，メタボ＋予備群の順位和は期待値より小さいが，より多く飲酒するものが多いとまではいえない。

ちなみに，同じデータを χ^2 検定すると (2×4 表)，χ^2＝5.56，p＝0.13 となり，有意差なしとの結果になる。p 値は，ノンパラメトリック検定より小さくなった。

表 9　飲酒量とメタボリックシンドロームの有無のクロス集計

	飲酒なし	1 合未満	1～2 合未満	2 合以上	計
メタボなし群	23	30	35	20	108
メタボ＋予備群	4	19	21	8	52
計	27	49	56	28	160

⑧ IF 関数文を用いたデータの分類

エクセルを用いてデータを分類しなおしたい場合，IF 関数を用いる。

たとえば，収縮期血圧，拡張期血圧から血圧分類（正常血圧＝1，境界域血圧＝2，高血圧＝3）を行いたいとする。演習データでは，H 列に収縮期血圧，I 列に拡張期血圧が入っている。データが入っている列の右端に新しい列を作り，2 行目のセル（列のタイトルのすぐ下）に，＝IF（AND（H2＜140，I2＜90），1，IF（AND（H2＜160，I2＜95），2，3））と入力するとその行のデータ（1 番目の人）が血圧分類のどれに当たるかという数値がでてくる。

この式の意味は，もしも最大血圧 140 未満，かつ最小血圧 90 未満であれば 1 と分類し，その残りのうちで最大血圧 160 未満かつ最小血圧 95 未満であれば 2 と分類し，それでも残れば（すなわち最大血圧 160 以上または最小血圧 95 以上）3 と分類する，ということである。このセル（2 行目）をコピーし，それ以下のすべてのセルへペースト（貼り付け）すれば，すべてのデータの血圧分類がたちどころにできる。

この方法を応用すると，住民のデータからある事業をよびかける対象者を抽出することもできる。たとえば，肥満（BMI 25 以上）で（必須），①血圧が正常血圧でない，②総コレ

ステロールが 230 mg/dL 以上，③喫煙のいずれか 2 項目が該当する人を抽出するなどの例である。血圧対象者，コレステロール対象者，喫煙対象者をそれぞれ 1 としてそれ以外を 0 とする列をつくり（血圧対象者：＝IF（J2＞1，1，0）　コレステロール対象者：＝IF（K2＜230，0，1）　喫煙対象者：＝IF（Q2＜4，1，0)），その次にその対象かどうかの結果の合計の列（仮に AD 列としておく）をつくり，最後に合計点が 2 以上で，BMI が 25 以上の対象者を抽出する列（＝IF（AND（G2＞＝25，AD2＞1)，1，0)）をつくると，対象者を抽出できる。