（付録A）補足

本文中に収めきれなかった事柄を、補足として掲載しました。

★（第1章「平均」補足）その他の平均

一般に「平均」といえば算術平均のことですが、それ以外の平均もいくつかあります。その中で、比較的よく使う平均として「幾何平均」と「調和平均」を紹介しておきます。

■幾何平均

算術平均では個々のデータを足し合わせて総度数で割りましたが、幾何平均は個々のデータを掛け合わせ、そのべき乗根を求めたものです。以下のように求めることができます。なお、幾何平均は相乗平均とも呼びます。

$$\text{幾何平均} = \sqrt[n]{X_1 \cdot X_2 \cdot \cdots \cdot X_n}$$

幾何平均は、経済成長率や投資収益率など、率の伸び縮みの平均を計算する際によく使います。例えば、3年間投資を行って、1年目は＋9%、2年目は－15%、3年目は＋12%の投資収益率になったとしましょう。この場合、伸び率の平均を幾何平均で計算すると、以下のようになります。

$$\sqrt[3]{(1+0.09)\times\{1+(-0.15)\}\times(1+0.12)} = 1.0124$$

つまり、この例では投資した金額が1年あたり平均で1.0124倍になることになります。いい換えると、3年間の投資収益率の平均は年1.24%だったということです。

■調和平均

調和平均は、個々のデータの逆数[※]を取り、その平均を求めて、さらにその逆数を取ったものです。調和平均は以下のように求めることができます。

$$\text{調和平均} = \frac{1}{\frac{1}{n}\left(\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_n}\right)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}}$$

調和平均は、速度の平均を計算するときなどに使います。例えば、ある区間を、行きは時速30Km、帰りは時速40Kmで通った場合、平均の速度は時速何Kmでしょうか？

行きが時速30Kmで、帰りが時速40Kmだから、平均は両者の間の時速35Kmでは？と思う方がいるかも知れません。しかし、それは正しくありません。

仮に、区間の距離をA（km）としましょう。すると、行き／帰りにかかった時間は、距離を時速で割れば求められるので、それぞれA/30（時間）、A/40（時間）です。この区間を往復すると、距離はAの2倍です。また、往復にかかる時間は、行きと帰りのそれぞれの時間を合計すればよいので、A/30＋A/40となります。これらから、往復の間の平均速度を求めると、以下のようになります。

$$\text{平均速度} = \frac{\text{往復の距離}}{\text{往復にかかった時間}} = \frac{2A}{\frac{A}{30}+\frac{A}{40}} = \frac{2A}{A\left(\frac{1}{30}+\frac{1}{40}\right)}$$

この式の分母／分子を2Aで割り算すると、最終的には以下のような式になり、平均速度は調和平均で求められることが分かります。

$$\text{平均速度} = \frac{1}{\frac{1}{2}\left(\frac{1}{30}+\frac{1}{40}\right)}$$

(※) 逆数

ある数値の分母と分子を入れ替えた値のことです。

この式を実際に計算して平均速度を求めると、時速34.29Kmとなります。これは、算術平均の時速35Kmとは違った値になっています。

なお、算術平均／幾何平均／調和平均の間には、以下の関係式が成り立ちます。

算術平均≧幾何平均≧調和平均

また、上式でイコール（＝）が成立するのは、すべてのデータが同じ値の場合のみです。

■Excelの関数で幾何平均／調和平均を求める

Excelには幾何平均／調和平均を計算する関数もあり、それぞれ「GEOMEAN」「HARMEAN」という名前です。書き方は以下のようになります。

幾何平均：=GEOMEAN(<データのセル範囲>)
調和平均：=HARMEAN(<データのセル範囲>)

例えば、A1～A5セルにデータを入力しているとして、その幾何平均を計算するには、計算先のセルに以下の式を入力します。

=GEOMEAN(A1:A5)

★（第1章「モードとメディアン」補足）データの傾向を表す各種の値

　モードとメディアン以外に、データの傾向を表す値がいくつかあります。それらを紹介します。

■パーセンタイル（分位点）

　パーセンタイル（日本語では「分位点」）は、メディアンの考え方を拡張したものです。データを小さいものから順に並べたときに、x%目の順位に来る値のことを、「xパーセンタイル」と呼びます。例えば、下から10%目の順位の値は「10パーセンタイル」と呼びます。
　特に、25パーセンタイルと75パーセンタイルがよく使われ、前者を「第一四分位点」、後者を「第三四分位点」と呼びます。ちなみに、メディアンは50パーセンタイルの値になりますが、これを「第二四分位点」とも呼びます。

●パーセンタイルと四分位点をExcelで求める

　パーセンタイルは「PERCENTILE」関数で求められます。書き方は以下の通りです。

　=PERCENTILE(<データのセル範囲>,<パーセント>)

　例えば、A1～A100セルにデータが入っているときに、10パーセンタイルの値を求めるには、セルに以下の式を入力します。

　=PERCENTILE(A1:A100,10%)

また、四分位点は「QUARTILE」関数で求められます。書き方は以下の通りです。

=QUARTILE(<データのセル範囲>,<位置の指定>)

「位置の指定」の部分には、表7.1の値を使って、どの四分位点の値を求めるかを指定します。

例えば、A1～A100セルにデータが入っているときに、第一四分位点の値を求めるには、セルに以下の式を入力します。

=QUARTILE(A1:A100,1)

表7.1　位置の指定に使う値

値	位置
0	最小値（＝0パーセンタイル）
1	第一四分位点（＝25パーセンタイル）
2	第二四分位点（＝50パーセンタイル＝メディアン）
3	第三四分位点（＝75パーセンタイル）
4	最大値（＝100パーセンタイル）

■データの散らばり具合を表す値

データの散らばり具合を表す値としては、分散と標準偏差が重要ですが、それ以外にもいくつかの値があります。

●レンジ

散らばり具合の表し方として、最も簡単なものはレンジ（range）です。レンジは、データ群の最大値から最小値を引いた値のことです。

レンジは単純で分かりやすい値です。ただ、データ群の中に飛び抜けて大きい（または小さい）値があると、その値によってレンジが大きくなってしまう、という問題があります。

●四分位偏差

レンジでは、上に述べたような問題が出ることがあります。そこで、飛び抜けた値を排除したレンジとして、四分位偏差を使うこともあります。四分位偏差は、第三四分位点の値から第一四分位点の値を引いて求められます。

なお、四分位偏差は、第三四分位点と第一四分位点の差の半分とする場合もあります。

●平均偏差

偏差の絶対値を平均して、散らばり具合を求めることも考えられます。これを平均偏差と呼びます。式で書くと以下のようになります。

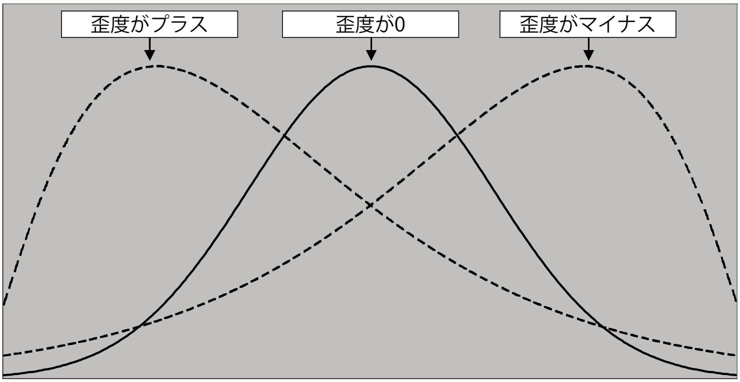
$$\text{平均偏差} = \frac{\left|X_1 - \overline{X}\right| + \left|X_2 - \overline{X}\right| + \cdots + \left|X_n - \overline{X}\right|}{n}$$

レンジや四分位偏差は、データ群の中のごく一部のデータだけを使った値なので、あまり良い値だとはいえません。一方、平均偏差はすべてのデータを元にして計算していますので、レンジや四分位偏差よりは良い値です。ただ、絶対値は数学的に扱いにくいので、平均偏差はあまり使われず、分散や標準偏差を使う方が多いです。

●歪度

歪度（「わいど」と読みます）は、データの分布が左右対称かどうかを示す値です。歪度が0に近いほど、分布が左右対象であることを示します。また、歪度がプラスの値なら分布の右裾が長く（右に歪んだ分布）、マイナスなら左裾が長い（左に歪んだ分布）ことを示します（図7.1）。

図7.1　データの分布と歪度の関係



●尖度

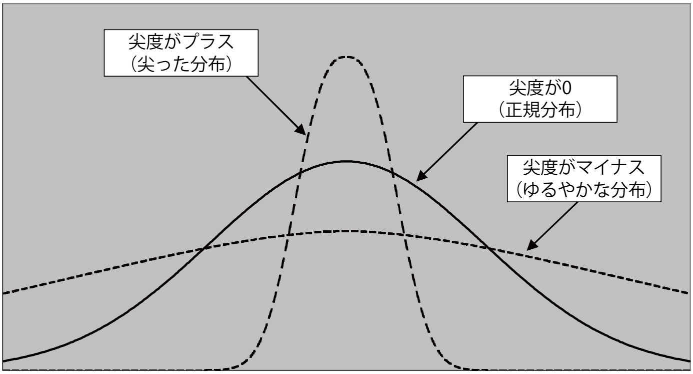
　ヒストグラムを作ると、その形が釣鐘状になることがよくあります。その形の理論上の極限を、「正規分布」と呼びます（正規分布の詳細については、本文第3章を参照)。

　データの分布が正規分布に近いかどうかを表す値として、尖度（「せんど」と読みます）があります。Excelには「KURT」という関数があり、尖度を求めることができます。

　尖度が0に近いほど、データの分布が正規分布に近いことを表します。また、尖度がプラスの値の場合、正規分布よりも尖った分布（平均付近にデータが集中した分布）になっていることを表します。一方、マイナスの値の場合は、ゆるやかな分布（平均から遠いところにもデータが広がっている分布）になっていることを表します（図7.2)。

　なお、上で説明した尖度に3を足して、「正規分布の場合は尖度が3」とすることもあります。

図7.2　データの分布と尖度の関係



■Excelの関数で散らばり具合を求める

レンジ／四分位偏差／平均偏差／歪度／尖度をExcelで求めることもできます。

●レンジを求める

レンジは、データ群の最大値から最小値を引けば求められます。Excelで最大値／最小値を求めるには、それぞれ「MAX」「MIN」関数を使います。書き方は以下の通りです。

最大値：=MAX(<データのセル範囲>)
最小値：=MIN(<データのセル範囲>

これらの関数を組み合わせて、以下のような式を作れば、レンジを求めることができます。

=MAX(<データのセル範囲>)－MIN(<データのセル範囲>)

例えば、A1セル～A10セルにデータが入力されている場合、レンジは以下の式で求められます。

=MAX(A1:A10)－MIN(A1:A10)

●四分位偏差を求める

四分位偏差は第三四分位点の値から第一四分位点の値を引けば良いので、以下の式で求めることができます。

=QUARTILE(<データのセル範囲>,3)－QUARTILE(<データのセル範囲>,1)

例えば、A1セル～A10セルにデータが入力されている場合、四分位偏差は以下の式で求められます。

=QUARTILE(A1:A10,3)－QUARTILE(A1:A10,1)

●平均偏差を求める

平均偏差を求めるには、「AVEDEV」という関数を使います。この関数の書き方は以下の通りです。

=AVEDEV(<データのセル範囲>)

例えば、A1セル～A10セルにデータが入力されている場合、平均偏差は以下の式で求められます。

=AVEDEV(A1:A10)

●歪度／尖度を求める

歪度／尖度を求めるには、それぞれ「SKEW」「KURT」という関数を使います。これらの関数の書き方は以下の通りです。

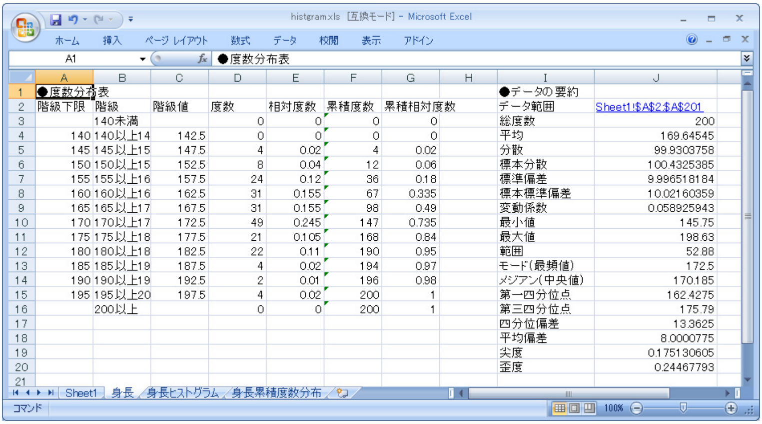
=SKEW(<データのセル範囲>)
=KURT(<データのセル範囲>)

■度数分布表ワークシートの「データの要約」について

本文XXページで、統計アドインを使って度数分布表を作る方法を解説しました。このワークシートには「データの要約」という部分があり、度数分布表の対象となるデータで、平均などを計算したものが表示されます（画面7.1）。それぞれの値の意味については、ここまでに解説した通りです。

なお、要約の途中にある「標本分散」「標本標準偏差」は、本文のXXページを参照してください。

画面7.1　度数分布表ワークシートの「データの要約」



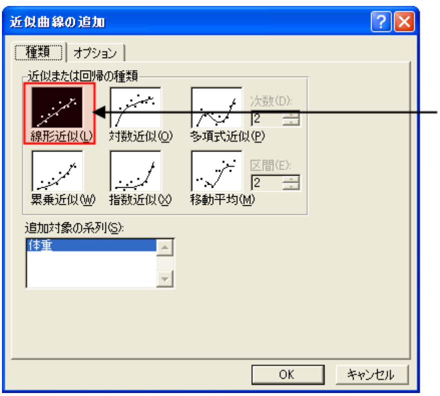
	A	B	C	D	E	F	G	H	I	J
1	●度数分布表								●データの要約	
2	階級下限	階級	階級値	度数	相対度数	累積度数	累積相対度数		データ範囲	Sheet1!$A$2:$A$201
3		140未満		0	0	0	0		総度数	200
4	140	140以上14	142.5	0	0	0	0		平均	169.64545
5	145	145以上15	147.5	4	0.02	4	0.02		分散	99.9303758
6	150	150以上15	152.5	8	0.04	12	0.06		標本分散	100.4325385
7	155	155以上16	157.5	24	0.12	36	0.18		標準偏差	9.99651818 4
8	160	160以上16	162.5	31	0.155	67	0.335		標本標準偏差	10.02160359
9	165	165以上17	167.5	31	0.155	98	0.49		変動係数	0.058925943
10	170	170以上17	172.5	49	0.245	147	0.735		最小値	145.75
11	175	175以上18	177.5	21	0.105	168	0.84		最大値	198.63
12	180	180以上18	182.5	22	0.11	190	0.95		範囲	52.88
13	185	185以上19	187.5	4	0.02	194	0.97		モード(最頻値)	172.5
14	190	190以上19	192.5	2	0.01	196	0.98		メジアン(中央値)	170.185
15	195	195以上20	197.5	4	0.02	200	1		第一四分位点	162.4275
16		200以上		0	0	200	1		第三四分位点	175.79
17									四分位偏差	13.3625
18									平均偏差	8.0000775
19									尖度	0.175130605
20									歪度	0.24467793
21										

★（第2章「回帰」補足）Excel 2003の「近似曲線の追加」機能の使い方

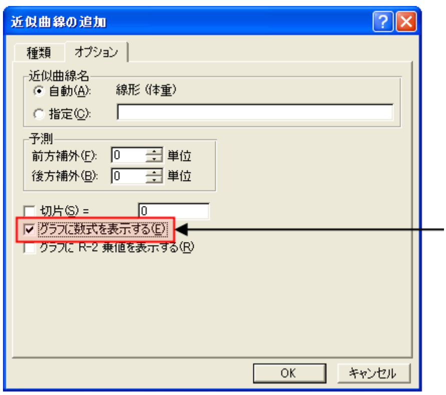
Excel 2003で、「近似曲線の追加」機能を使って回帰曲線を求めるには、以下の手順を取ります。

①[グラフ(C)]→[近似曲線の追加(R)...]メニューを選び、「近似曲線の追加」ダイアログボックスを開きます。
②「種類」タブの「近似または回帰の種類」で「線形近似(L)」をクリックして選択します（画面7.2）。
③「オプション」タブに切り替え、「グラフに数式を表示する(E)」のチェックをオンにします（画面7.3）。また、回帰直線に名前をつけたい場合は、「近似曲線名」部分の「指定(C)」をオンにし、その右の欄に線の名前を入力します。
④「OK」ボタンをクリックします。

画面7.2　「種類」タブで「線形近似(L)」を選択する



画面7.3　「オプション」タブで「グラフに数式を表示する(E)」のチェックをオンにする



★（第3章「確率変数の期待値と分散」補足）離散型確率分布の期待値と分散の例

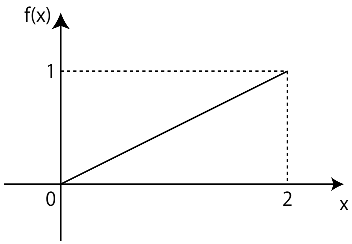
■確率変数の期待値の例

例

確率密度関数が以下の式で表されるような、連続型の確率分布があるとします。この確率密度関数をグラフで表すと、図7.3のようになります。この確率変数の期待値を求めてみましょう。

$$f(x)=\begin{cases}\dfrac{1}{2}x & (0 \leqq x \leqq 2) \\ 0 & (\text{その他の}x)\end{cases}$$

図7.3　確率密度関数のグラフ



本文XXページで述べたように、期待値は変数xに確率密度関数を掛け算して、それを−∞〜∞まで積分することで求められます。実際に計算すると、以下のようになります。

$$E(X)=\int_{-\infty}^{\infty}xf(x)dx$$
$$=\int_{-\infty}^{0}x\cdot 0dx+\int_{0}^{2}x\cdot\frac{1}{2}xdx+\int_{2}^{\infty}x\cdot 0dx$$
$$=\left[\frac{1}{6}x^3\right]_0^2=\frac{4}{3}$$

4/3＝1.333・・・で、範囲の中央である1よりは、やや2に寄った値になっています。確率密度関数（図7.3）から分かるように、2に近い値ほど取る確率が高いので、期待値も2に寄った値になっています。

■確率分布の分散の例

例

確率変数Xの確率分布が、図7.3のようになっている場合に、その分散を求めてみましょう。

Xの期待値E(X)は、4/3でした（期待値を求める例を参照)。ここから、分散は以下のように求めることができます。

$$V(X)=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$
$$=\int_{-\infty}^{0}\left(x-\frac{4}{3}\right)^2\cdot 0dx+\int_{0}^{2}\left(x-\frac{4}{3}\right)^2\cdot\frac{1}{2}xdx+\int_{2}^{\infty}\left(x-\frac{4}{3}\right)\cdot 0dx$$
$$=\int_{0}^{2}\left(\frac{1}{2}x^3+\frac{4}{3}x^2+\frac{8}{9}x\right)dx$$
$$=\left[\frac{1}{8}x^4+\frac{4}{9}x^3+\frac{4}{9}x^2\right]_0^2=\frac{22}{3}$$

★（第3章「正規分布」補足）正規分布に従う乱数をExcelで発生させる

Excelには、正規分布に従う乱数を発生する機能があります。正規分布のシミュレーションなどに使うことができます。

■分析ツールをオンにする

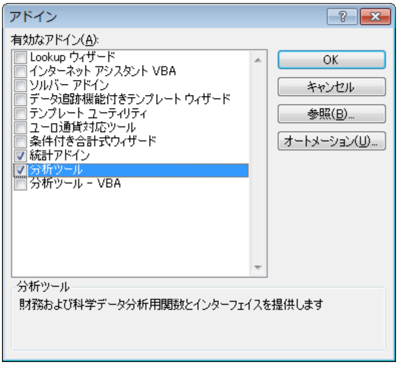
この後の手順で、「分析ツール」というアドインを使います。このアドインをオンにしていない場合は、以下の手順でオンにします。

●Excel 2003での手順

以下の通りです。

①[ツール(T)]→[アドイン(I)...]メニューを選び、「アドイン」のダイアログボックスを開きます。
②アドインの一覧で、「分析ツール」のチェックをオンにして、「OK」ボタンをクリックします（画面7.4）。

画面7.4　分析ツールアドインをオンにする

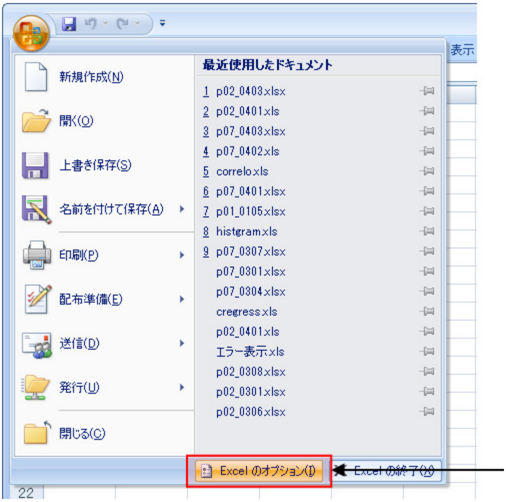


●Excel 2007／2010での手順

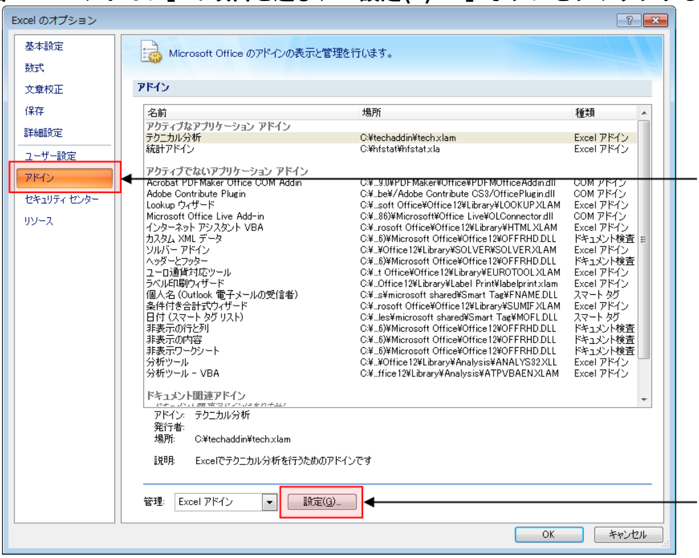
以下の通りです。

①Officeボタン（2010ではリボンの「ファイル」）をクリックし、メニューの下端の方に表示される「Excelのオプション」（Excel 2007では「オプション」）をクリックして、「Excelのオプション」ダイアログボックスを開きます（画面7.5）。
②ダイアログボックス左端で、「アドイン」の項目を選びます（画面7.6）。
③アドイン一覧の画面が表示されますので、「設定(G)...」のボタンをクリックします（画面7.6）。
④「アドイン」のダイアログボックスが開きますので、Excel 2003と同様の手順で分析ツールアドインをインストールします。

画面7.5　「Excelのオプション」ボタンをクリックする



画面7.6　「アドイン」の項目を選び、「設定(G)...」ボタンをクリックする



●アドイン一覧に「分析ツール」が出ない場合

Excelのインストール方法によっては、アドインの一覧に「分析ツール」が出な

いことがあります。その場合は、Excelのインストールプログラムを再度起動して、分析ツールを追加でインストールします。

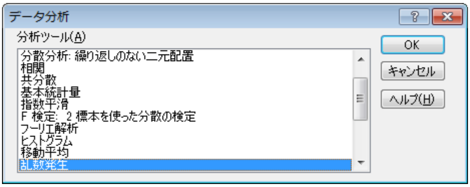
■正規分布に従う乱数を分析ツールで発生させる

Excelの「分析ツール」という機能を使うと、正規分布に従う乱数を発生させることができます。

まず、「データ分析」のダイアログボックスを開きます。Excel 2003では、[ツール(T)]→[分析ツール(D)...]メニューを選びます。また、Excel 2007／2010では、「データ」のリボンで「データ分析」のボタンをクリックします。

ダイアログボックスが開いたら、一覧の中で「乱数発生」を選んで（画面7.7）、「OK」ボタンをクリックします。

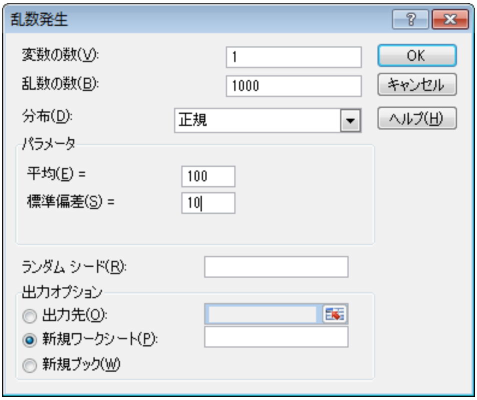
画面7.7　「データ分析」ダイアログボックスで「乱数発生」を選ぶ



すると、「乱数発生」ダイアログボックスが開きますので、以下のように設定します（画面7.8）。

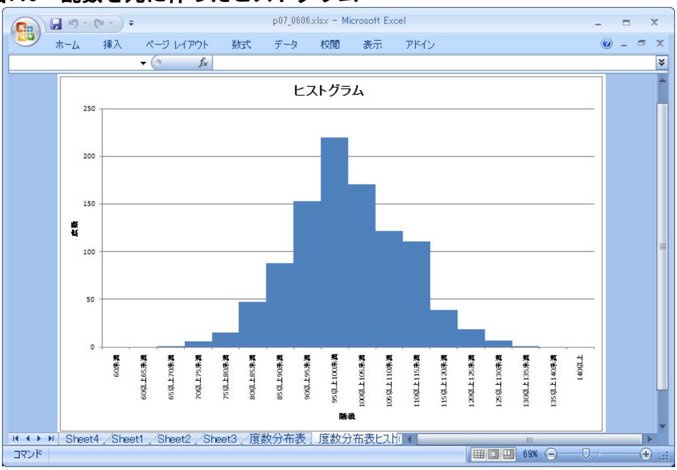
①「変数の数(V)」
「1」を入力します。
②「乱数の数(B)」
発生させる乱数の個数を指定します。
③「分布(D)」
「正規」を選びます。
④「パラメータ」
「平均(E)」と「標準偏差(S)」の欄に、正規分布の平均と標準偏差を入力します。
⑤「出力オプション」
乱数の出力先を指定します。

画面7.8　正規分布に従う乱数を発生させる



　これで「OK」ボタンをクリックすると、出力先のワークシートに指定した個数の乱数が出力されます。その乱数を元に、30ページの手順でヒストグラムを作ると、正規分布のシミュレーションができます（画面7.9）。

画面7.9　乱数を元に作ったヒストグラム



■正規分布に従う乱数を関数で発生させる

平均μ、標準偏差σの正規分布に従う乱数は、Excelの関数を使って発生させることもできます。関数の式は以下のようになります。

=NORMINV(RAND(),μ,σ)

例えば、平均10、標準偏差3の正規分布に従う乱数は、以下の式で発生させることができます。

NORMINV(RAND(),10,3)

★（第4章　各種区間推定の補足）

■標本数が十分に多く母分散が既知の場合の母平均の区間推定の考え方

母平均がμで母分散がσ^2の母集団があるとします。このとき、中心極限定理により、標本の数が十分に多ければ、母集団の分布の形に関係なく、標本平均$\overline{X}$の分布は$N\left(\mu,\sigma^2/n\right)$の正規分布に従います。

次に、以下のように$\overline{X}$を標準化（XXページ参照）し、変数Zを作ります。すると、この変数Zは標準正規分布N(0,1)に従うことになります。

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\cdots \text{①}$$

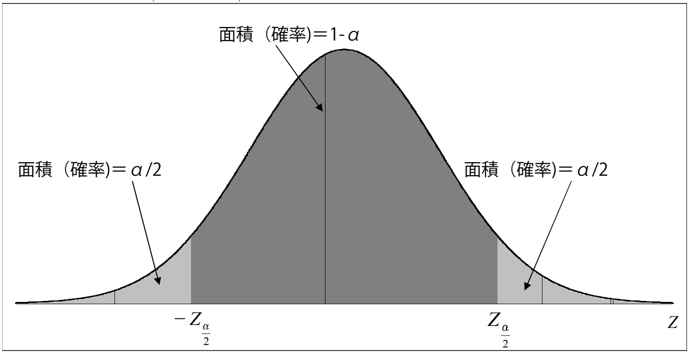
ここで、図7.4のように両側100α%点を考えてみましょう。下側のパーセント点を$-Z_{\alpha/2}$、上側のパーセント点を$Z_{\alpha/2}$とします。すると、パーセント点の考え方から、左端／右端の薄い灰色の部分の面積（確率）はそれぞれα/2となり、残りの濃い灰色の部分の面積（確率）は1－αになります（図7.4)。

これは、変数Zが$-Z_{\alpha/2}$以上$Z_{\alpha/2}$以下の値を取る確率が、1－αになることを意味しています。式で書くと以下の通りです。

$$P\left(-Z_{\frac{\alpha}{2}}\leqq Z\leqq Z_{\frac{\alpha}{2}}\right)=1-\alpha\cdots \text{②}$$

例えば、α＝0.05（5%）とすると、変数Zが$-Z_{0.025}$以上$Z_{0.025}$以下の値を取る確率は、1－0.05＝0.95（95%）ということになります。

図7.4　変数Zが$-Z_{\alpha/2}$ 以上$Z_{\alpha/2}$ 以下の値を取る確率が1－αになる



ここで、②式のP()の括弧の中に、①式のZを代入します。すると、以下の式ができます。

$$-Z_{\frac{\alpha}{2}} \leqq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqq Z_{\frac{\alpha}{2}} \quad \cdots \text{③}$$

次に、この式のそれぞれの辺に$-\sigma/\sqrt{n}$を掛けて、以下のように変形します。

$$-\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}} \leqq \mu-\overline{X} \leqq \frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}}$$

さらに、この式のそれぞれの辺に$\overline{X}$を足して、以下のように変形します。

$$\overline{X}-\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}} \leqq \mu \leqq \overline{X}+\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}}$$

この式は元の③式と同じことを意味しています。したがって、②式の括弧の中を上の式に置き換えて、以下のように書くことができます。

$$P\left(\overline{X}-\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}} \leqq \mu \leqq \overline{X}+\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}}\right)=1-\alpha$$

この式は、以下のようなことを意味しています。

「母平均μが$\overline{X}-\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}} \leqq \mu \leqq \overline{X}+\frac{\sigma}{\sqrt{n}}\cdot Z_{\frac{\alpha}{2}}$の範囲の値を取る確率は、1－αである」

これが、分散σが既知で標本数nが大きい場合の、母平均の区間推定になります。

■母分散が未知の正規母集団の場合の母平均の区間推定の考え方

正規母集団からn個の標本を抽出した場合に、その標本平均$\overline{X}$と標本分散s^2から、母平均μを推定することができます。

本文XXページで述べたように、$\overline{X}$から以下の変換を行って変数tを作ると、tは自由度n－1のt分布に従います。

$$t=\frac{\overline{X}-\mu}{s/\sqrt{n}} \quad \cdots \text{①}$$

このtが両側100α%点の間の値を取る確率は、1－αです（図7.5)。式で表すと以下のようになります。

$$P\left(-t_{\frac{\alpha}{2}} \leqq t \leqq t_{\frac{\alpha}{2}}\right)=1-\alpha \quad \cdots \text{②}$$

この式の括弧の中の式を取り出し、tに①式を代入すると、以下のようになります。

$$-t_{\frac{\alpha}{2}} \leqq \frac{\overline{X}-\mu}{s/\sqrt{n}} \leqq t_{\frac{\alpha}{2}}$$

次に、この式の各辺に$-s/\sqrt{n}$を掛けて、以下のように変形します。

$$-t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leqq \mu - \overline{X} \leqq t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

そして、この式の各辺に$\overline{X}$を足して、以下のように変形します。

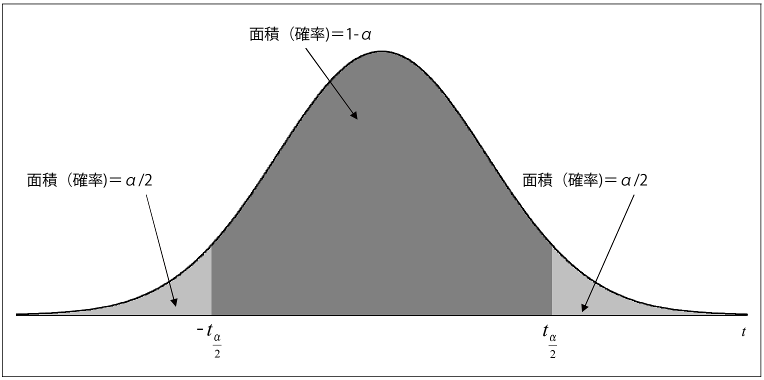
$$\overline{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leqq \mu \leqq \overline{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

この式は、②式の括弧の中を変形したものですので、以下が成り立ちます。

$$P\left(\overline{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leqq \mu \leqq \overline{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

つまり、母平均μは1－αの確率で、$\overline{X} - t_{\alpha/2} \cdot s/\sqrt{n}$から$\overline{X} + t_{\alpha/2} \cdot s/\sqrt{n}$の間の値を取ることになります。ここまでで、母平均μの100(1－α)%信頼区間を求めることができました。

図7.5　tが両側100α%点の間の値を取る確率は1- α



■母分散の区間推定

本文XXページで述べたように、標本の数がnで、標本分散がs^2、母分散がσ^2の場合、以下の式でχ^2を求めると、このχ^2は自由度n－1のカイ二乗分布に従います。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \cdots ①$$

したがって、χ^2が両側100α%点の間の値を取る確率は、1－αになります。式で表すと以下の通りです。また、図で表すと図7.6になります。

$$P\left(\chi^2_{1-\frac{\alpha}{2}} \leqq \chi^2 \leqq \chi^2_{\frac{\alpha}{2}}\right) = 1-\alpha \quad \cdots ②$$

ここで、この式の括弧の部分を取り出し、そこにあるχ^2に①式を代入します。すると以下のようになります。

$$\chi^2_{1-\frac{\alpha}{2}} \leqq \frac{(n-1)s^2}{\sigma^2} \leqq \chi^2_{\frac{\alpha}{2}}$$

次に、この式の分子と分母を入れ替えます。すると、不等号の向きも逆になって、以下のようになります。

$$\frac{1}{\chi^2_{\frac{\alpha}{2}}} \leqq \frac{\sigma^2}{(n-1)s^2} \leqq \frac{1}{\chi^2_{1-\frac{\alpha}{2}}}$$

次に、この式の各辺に$(n-1)s^2$を掛け算して、以下のように変形します。

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \cdots ③$$

この③式は、上の②式の括弧の中を変形したものですので、②式は以下のように変形できることになります。

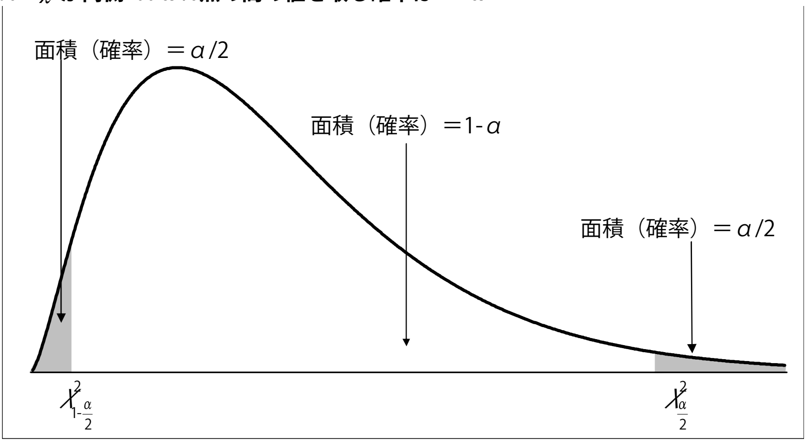
$$P\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

つまり、母分散は、1－αの確率で、以下の信頼区間の間の値を取ることになります。

下側信頼限界 $= \dfrac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}$　上側信頼限界 $= \dfrac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$

これで、母分散の推定ができました。

図7.6　χ^2が両側100α％点の間の値を取る確率は1－α



■母比率の区間推定

ここまでの推定のパターンは、標本平均や標本分散がどのような確率分布に従

うかを考え、その両側100α％点から、母平均や母分散の信頼区間を計算する、という手順でした。標本比率を推定する場合も、同じ考え方をします。

●標本比率の分布は正規分布で近似できる

仮に、母集団がAとBの2つだけから構成されているものとします。また、母集団全体のうち、Aが占める割合を母比率とし、それをpで表すことにします。その母集団からn個の標本を抽出した場合に、Aの個数をXとすると、Xは0からnまでの値を取る確率変数になり、二項分布に従います。また、期待値E(X)＝npで、分散V(X)＝np(1－p)です。

例えば、母集団のうち、Aが占める割合（母比率）が0.7だとします。この母集団から10個の標本を抽出して、Aの個数をXとすれば、Xはn＝10、p＝0.7の二項分布に従うことになります。

ただ、標本の数が多ければ、期待値E(X)＝npで分散V(X)＝np(1－p)の二項分布は、同じ期待値／分散の正規分布で近似することができました（本文XXページ参照）。

ここで、確率変数Xを、標本の数のnで割ってみます。Xは標本中のAの個数です。また、nは標本全体の個数です。したがって、X/nは標本全体の中でAが占める割合になり、標本比率になります。これをPで表すことにします（$P=X/n$）。

上で述べたように、Xの分布は、期待値E(X)＝np、分散V(X)＝np(1－p)の正規分布で近似することができます。PはXをnで割ったものですので、この分布も正規分布で近似することができます。また、その期待値E(P)と分散V(P)は、以下のように求められます。

$$E(P)=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}\cdot np=p$$

$$V(P)=V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{1}{n^2}\cdot np(1-p)=\frac{p(1-p)}{n}$$

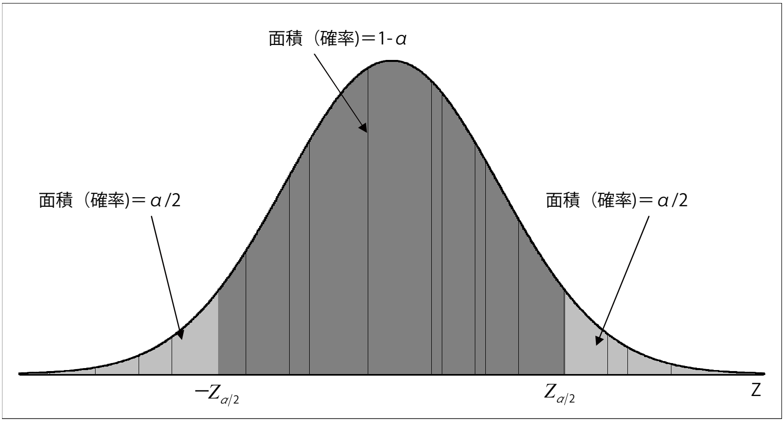
次に、Pを以下のように標準化してZを求めると、このZは標準正規分布に従います。

$$Z=\frac{P-E(P)}{\sqrt{V(P)}}=\frac{P-p}{\sqrt{p(1-p)/n}}\quad\cdots\ ①$$

したがって、標準正規分布の両側100α%点を$-Z_{\alpha/2}$ ／$Z_{\alpha/2}$ とすれば、上のZの値が両側100α%点の間の値を取る確率は、1－αになります（図7.7）。これを式で表せば、以下のようになります。

$$P\left(-Z_{\frac{\alpha}{2}}\leqq Z\leqq Z_{\frac{\alpha}{2}}\right)=1-\alpha\quad\cdots\ ②$$

図7.7　Zが$-Z_{\alpha/2}$ から$Z_{\alpha/2}$ の間の値を取る確率は1－α



●信頼区間を求める

次に、先ほどの②式の括弧の部分を取り出し、そこにあるZに①式を代入します。すると以下のようになります。

$$-Z_{\frac{\alpha}{2}}\leqq\frac{P-p}{\sqrt{p(1-p)/n}}\leqq Z_{\frac{\alpha}{2}}$$

ここで推定したいのは、母比率のpです。そこで、上の式を○≦p≦△の形に変形します。ただ、この式を見ると分母と分子の両方にpがあり、しかもルートの中

にもpがあって、簡単には変形できません。

そこで、分母にある母比率pを、標本比率Pで置き換えます。標本の数が十分に多ければ、標本比率Pは母比率pとそう大きく違わないはずなので、置き換えても大丈夫だと考えるわけです。すると、以下のようになります。

$$-Z_{\frac{\alpha}{2}} \leqq \frac{P-p}{\sqrt{P(1-P)/n}} \leqq Z_{\frac{\alpha}{2}}$$

次に、この式の各辺に$-\sqrt{P(1-P)/n}$を掛け算して、以下のように変形します。

$$-Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n} \leqq p-P \leqq Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n}$$

最後に、この式の各辺にPを足して、以下のように変形します。

$$P-Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n} \leqq p \leqq P+Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n} \quad \cdots ③$$

この式は、前ページの②式で括弧の中を変形したものです。したがって、この部分を②式の括弧に戻して、以下のようになります。

$$P\left(P-Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n} \leqq p \leqq P+Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n}\right)=1-\alpha$$

ここまでで、母比率pの信頼区間を求めることができました。信頼区間の下側信頼限界と上側信頼限界は、それぞれ以下の式で表されます。

上側信頼限界$= P-Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n}$

下側信頼限界$= P+Z_{\frac{\alpha}{2}} \cdot \sqrt{P(1-P)/n}$

■自由度

自由度とは、「自由に動ける変数の数」のことを表します。
例えば、標本と標本平均には以下のような関係があります。

$$\overline{X} = \frac{X_1 + X_2 + \cdot\cdot\cdot + X_{n-1} + X_n}{n}$$

この式は以下のように変形することができます。

$$\left(X_1 - \overline{X}\right) + \left(X_2 - \overline{X}\right) + \cdot\cdot\cdot + \left(X_{n-1} - \overline{X}\right) + \left(X_n - \overline{X}\right) = 0$$

ここで、$Y_i = X_i - \overline{X}\left(i = 1,2,\cdot\cdot\cdot,n\right)$のように置くと、上の式は以下のように変形できます。

$$Y_1 + Y_2 + \cdot\cdot\cdot + Y_{n-1} + Y_n = 0$$

さらに、上の関係式は以下のように変形できます。

$$Y_n = -\left(Y_1 + Y_2 + \cdot\cdot\cdot + Y_{n-1}\right)$$

つまり、Y_nはY_1～Y_{n-1}によって決まることになり、自由な値を取ることができません。したがって、自由度は標本の数のnから1減って、(n-1)になるわけです。

★（第6章「重回帰分析」補足）LINEST関数

Excelには統計関数が多数用意されていますが、重回帰分析を行う関数として「LINEST」があります。

■LINEST関数の書き方

LINEST関数の書き方は、以下のようになります。

=LINEST(目的変数のセル範囲,説明変数のセル範囲,定数,補正)

「定数」の部分では、回帰式の定数項を強制的に0にするかどうかを指定します。この値を「FALSE」にすると、定数項が0になります。一方、「TRUE」にすると、通常通り定数項も計算されます。

「補正」の部分では、「標準誤差」などの値も計算するかどうかを指定します（標準誤差については、より詳しい統計の書籍等を参照してください）。「TRUE」を指定すると標準誤差等も計算され、一方「FALSE」を指定すると標準誤差等は計算しません。なお、ここでTRUEを指定すると、決定係数を計算することもできます。

また、LINEST関数は、一度に多数の値を返す関数になっています。そのため、1つのセルに式を入力するのではなく、複数のセルにまとめて式を入力します（配列数式にします）。

「補正」の部分をFALSEにする場合は、縦は1行／横は（説明変数の数＋1）列のセル範囲を選択した状態でLINEST関数の式を入力し、Ctrlキーを押しながらEnterキーを押します。

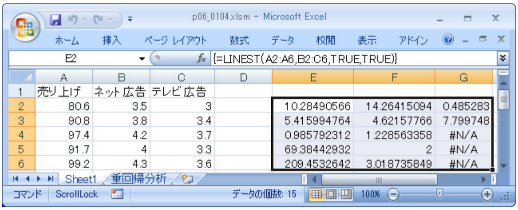
また、「補正」の部分をTRUEにする場合は、縦は5行／横は（説明変数の数＋1）列のセル範囲を選択した状態でLINEST関数の式を入力し、Ctrlキーを押しながらEnterキーを押します。

■LINEST関数の結果の見方

結果の1行目には、偏回帰係数と定数項が出力されます。セル範囲の左端のセルの値がa_n、その右のセルがa_{n-1}・・・というように順に出力され、右端のセルが定数項a_0になります。

また、「補正」に「TRUE」を指定した場合、2行目以降に標準誤差等が出力され、3行目の1列目に決定係数が出力されます（画面7.10)。

画面7.10　LINEST関数の結果の例



■偏回帰係数／定数項／決定係数だけを取り出す

LINEST関数の結果から、特定の偏回帰係数や、定数項／決定係数だけを取り出すこともできます。

説明変数の数がn個の場合、k番目の偏回帰係数a_kは、以下の式で求めることができます。

=INDEX(LINEST(目的変数のセル範囲,説明変数のセル範囲),n-k+1)

定数項は以下の式で求めることができます。

=INDEX(LINEST(目的変数のセル範囲,説明変数のセル範囲),n+1)

また、決定係数は以下の式で求めることができます。

=INDEX(LINEST(目的変数のセル範囲,説明変数のセル範囲,TRUE,TRUE),3,1)