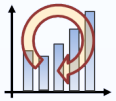
総務省　ICTスキル総合習得教材

【概要版】eラーニング用 

［コース3］データ分析 

3-3：基本統計量･クロス集計表の作成

	1	2	3	4	5
［コース1］データ収集					
［コース2］データ蓄積					
［コース3］データ分析			◆		
［コース4］データ利活用					

本講座の学習内容（3-3：基本統計量・クロス集計表の作成）

【講座概要】

- 数値データの尺度に基づく４つのデータの種類を説明します。
- 基本統計量を紹介し、Excel関数による導出方法を説明します。
- Excelのピボットテーブルを用いて、クロス集計表を作成する方法を紹介します。

【講座構成】

座学	[1] 数値データの尺度と種類
実習紹介	[2] Excel関数による基本統計量の導出
	[3] ピボットテーブルとクロス集計表

【学習のゴール】

- ✓ 数値データの４つの尺度を理解する。
- ✓ 基本統計量の意味を理解し、代表的な基本統計量が説明できる。
- ✓ Excel関数を用いて、基本統計量を導出できる。
- ✓ ピボットテーブルを用いて、クロス集計表や基本統計量に関する表を作成できるようになる。

数値データの尺度

◆数値データの尺度には「名義尺度 、順序尺度、間隔尺度 、比率尺度」の4種類があります。

- 基本統計量の説明の前段階として、数値データの尺度の4種の尺度を紹介します。

- 電話番号や郵便番号のように区別や分類のみのために用いられる番号を**名義尺度**といいます。
  - 郵便番号の数字をアルファベットに変更するように、名義尺度は数値を記号に変換してもその目的を果たせます。
  - 血液型の「A型を1型、B型を2型、AB型を3型、O型を4型」と呼べば、名義尺度といえます。
  - 名義尺度では、一致する（等しい）か、どうかのみに意味があり、大小関係に意味はありません。

- 地震の震度や5段階評価の満足度や成績のように、数値に大小関係（順序）はあるものの数値の間隔に意味はないものを**順序尺度**といいます。
  - 「震度3は震度2より揺れが強い」とは言えますが、「震度3は震度2と震度1が合わさった振動」や「震度3は震度2の1.5倍の揺れ」とは言えません。

- 温度や西暦のように目盛が等間隔で差に意味がありつつも、0や比に意味がない数値を**間隔尺度**といいます。
  - 温度の「1℃→2℃」と「2℃から3℃」は「同じ1℃の上昇」とはいえますが、「3℃は1℃の3倍の温度」とはいえません。
  - 0℃は水が凍る温度の融点としての意味はあっても、0℃でも温度がなくなるわけではありません。

- 重量や長さのように0に原点としての意味があり、間隔と比率の両方に意味がある尺度を**比率尺度**といいます。
  - 重量（kg）や長さ（cm）では、「50kgと100kg」「100cmと200cm」は、それぞれ2倍といえます。
  - 重さ0kg、長さ0cmは、何も存在しないことに対応しています。

□ この4種類の尺度によって、平均値などの基本統計量に意味があるかが異なってきます。

数値データの尺度と平均値

◆数値データの尺度によって、平均値に意味があるかが異なってきます。

- 4種類の尺度のうち、間隔尺度、比率尺度は平均値によってデータを代表する値を示すことができます。
  - 満足度などの5段階評価の順序尺度においても、便宜的に平均値による指標を表示することがあります。しかし、同じ順序の中で最上位の点数のラベルを5点から100点に変えた場合に平均値は変化するため、代表的な値の表示に適切とは言えません。

4種の数値データの尺度の特徴

	事例	大小比較	差分	比率	代表的な値の表示に適切な統計量
名義尺度	郵便番号 部屋番号	×	×	×	• 最頻値
順序尺度	震度 5段階評価	○	×	×	• 中央値 • 最頻値
間隔尺度	温度（℃）、 西暦	○	○	×	• **平均値** • 中央値 • 最頻値
比率尺度	重さ、長さ	○	○	○	• **平均値** • 中央値 • 最頻値

- 日本の住民の住所を1～47の都道府県番号（名義尺度）で表した場合、最も住民が多い都道府県（最頻値）は、「13番の東京都」とはいえますが、平均値を計算しても意味はありません。
  - 平均値を計算すると、20.7となりますが、都道府県番号「20番の長野県」「21番の岐阜県」とは無関係です。

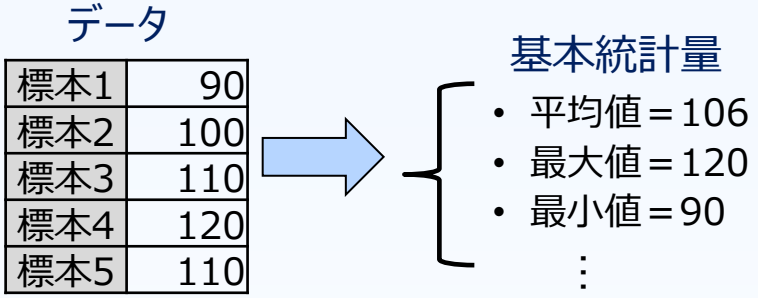
基本統計量

◆基本統計量とは、データセット全体の特徴をそれぞれ一つの値で要約する指標を指します。

- 気温、人間の身長のデータを分析対象とする場合、調査対象とした時間や人数だけデータレコード（標本数、サンプルサイズ）があります。
  - 統計学では、データの中にある一つ一つの観測値を「標本」や「サンプル」と言います。データという言葉はデータセット全体を指すのか、一つ一つの標本を指すのかが不明瞭になりやすいため、言葉を区別しています。
  - データセットに入っている標本の合計を標本数やサンプルサイズと言います。Excelの表形式のデータベース（リレーショナルデータベース）に格納した場合、標本数はデータレコード数（行数）に対応します。

- 標本数が多くなると、一つ一つの標本を確認してデータの特徴を把握することが困難になります。
  - 毎分収集している気温のデータは、1日のデータでも1440標本となり、一つ一つの標本を見て確認するだけでも手間がかかります。

- データセット全体の特徴（代表的な値やバラツキの程度）をそれぞれ一つの値で要約する指標を基本統計量といいます。
  - 基本統計量は、記述統計量や要約統計量と呼ばれることもあります。
  - 「代表的な値」の一つとして平均値が含まれ、平均値も基本統計量の一つです。
  - 最大値および最小値は、それぞれがデータセットの状態を一つの値で要約する指標であり、基本統計量に該当します。
  - 平均値が同じ100であっても、99～101の間に標本が散らばっている状況と、10～190の間に標本が散らばっている状況では、「バラツキの程度」が異なります。基本統計量の中には「バラツキの程度」を表す指標もあります。



データの代表的な値（平均値、中央値、最頻値）

◆ 平均値、中央値、最頻値は、それぞれデータセットの代表的な値を表す基本統計量です。

- ここではサンプルデータとして、2017年7月1日～9月30日の東京の天候データ（気温）を利用します。
  - 同時点の天気データが付いている各日の3時、6時、9時、12時、15時、18時、21時の1日7時点の644標本のデータを利用します。
  - 気象庁の「過去の気象データ・ダウンロード」（http://www.data.jma.go.jp/gmd/risk/obsdl/index.php）から、ダウンロードすることができます。

- 平均値（mean）は、標本の合計値を標本数で割ったものに対応し、標本から得られた平均値として標本平均とも呼ばれます。
  - セルに『= AVERAGE(データ範囲)』と入力することで、平均値が導出できます。

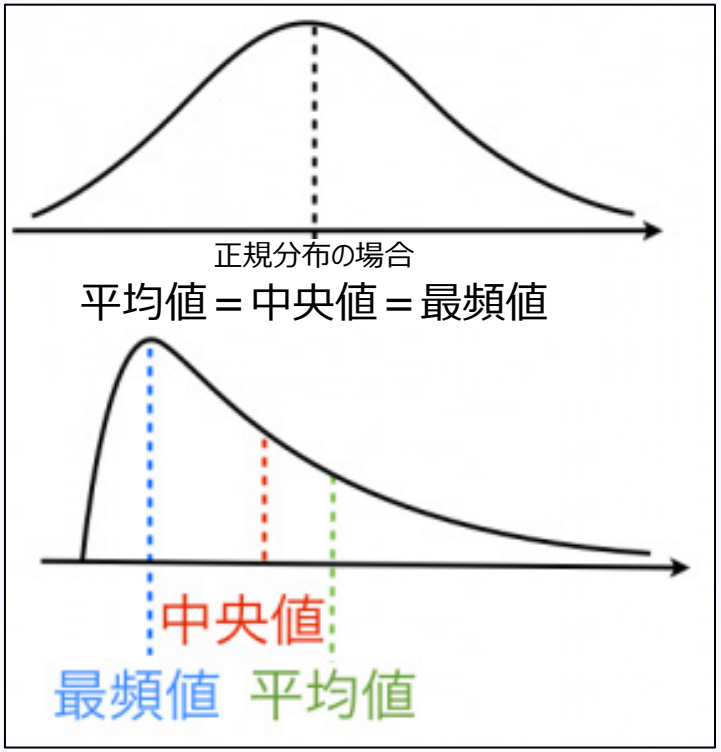
- 中央値(median)は、標本を大小関係で並べた際に中央の順位にある値を指します。
  - セルに、『= MEDIAN(データ範囲)』と入力することで、中央値が導出できます。
  - 順序尺度において、順序を変えない範囲で5点→100点などのラベル替えをしても、中央値は影響を受けないため、中央値は順序尺度においても意味を持つ指標です。

- 最頻値(mode)は、データの中で最も頻度が高い値を指します。
  - セルに『= MODE(データ範囲)』と入力することで、最頻値が導出できます。

サンプルデータにおける平均値、中央値、最頻値のExce出力

基本統計量の名称	値	Excel関数の入力
平均値	25.7	=AVERAGE(C2:C645)
中央値	25.7	=MEDIAN(C2:C645)
最頻値	27.4	=MODE(C2:C645)

平均値、中央値、最頻値のイメージ



標本のバラツキを示す指標（分散、標準偏差）

◆分散、標準偏差は、それぞれ標本のバラツキを表す基本統計量です。

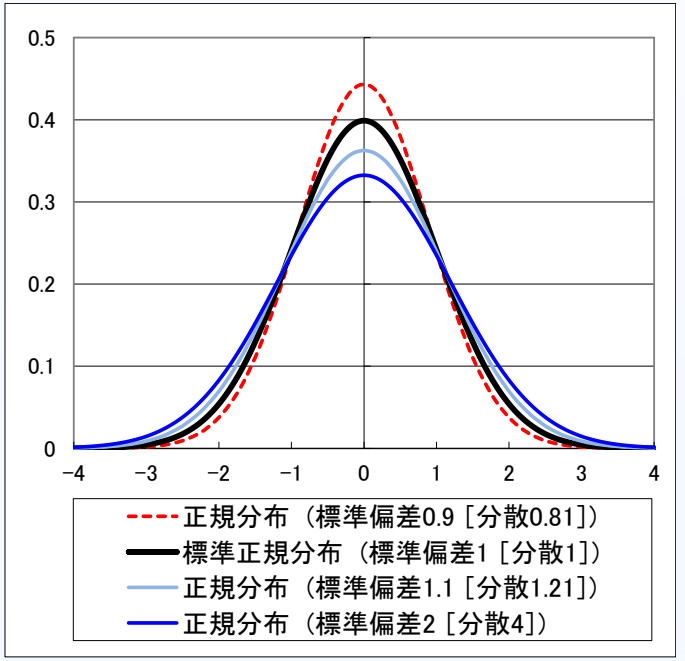
- 分散は標本のバラツキを示す指標であり、導出過程で偏差平方和を利用しています。
  - 各標本の標本平均からのズレとしての偏差（＝各標本の値-標本平均）には正と負の値が両方があり、全標本で偏差の総和をとると正と負が打ち消し合って0になります。このため、各標本の偏差を二乗することで負の偏差も正の値にしてから総和をとった偏差平方和を導出過程で利用して、バラツキの指標の分散を導出します。
- 偏差平方和を（標本数-1）で割ることで、標本分散と呼ばれる二乗しているバラツキの指標が導出できます。
  - セルに『＝ VAR.S(データ範囲)』と入力すると、標本に基づいて母集団の分散を偏りなく推定できる不偏分散が導出できます。
  - 標本数自体ではなく、（標本数-1）で割る理由や「偏りなく」の意味は、やや専門的になるため、関心がある方は統計学の入門書を参照して下さい。
- 標準偏差は、二乗していた尺度を元に戻したバラツキの指標です。
  - セルに『＝ STDEV.S((データ範囲)』と入力すると、標本に基づく標準偏差が導出できます。
  - 不偏分散は偏差を二乗してから総和をとった偏差平方和から導出しているため、二乗されたバラツキの指標となっています。標準偏差は、不偏分散の正の平方根をとることで、二乗されていた尺度を元に戻しています。

サンプルデータにおける平均値、中央値、最頻値のExce出力

基本統計量の名称	値	Excel関数の表記
（標本）分散	13.01	=VAR.S(C2:C645)
（標本）標準偏差	3.61	=STDEV.S(C2:C645)

- 統計学に立ち入らない方は、「分散」と「標準偏差」は、バラツキの指標と理解するだけで構いません。

標準偏差の正規分布への影響



標準偏差が大きくなると、分布の頂点は低くなり、分布の裾は広がります。

ピボットテーブルの利用

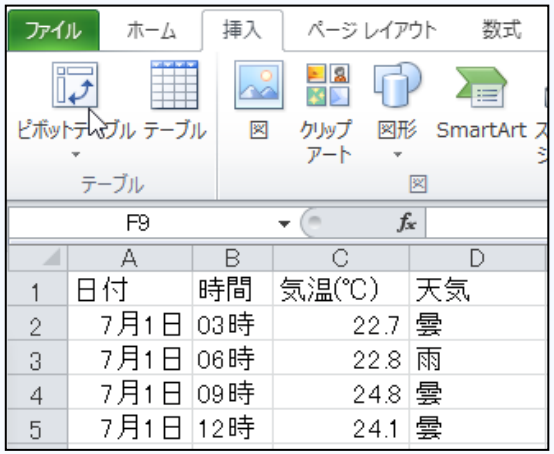
◆ Excelにはピボットテーブルと呼ばれるクロス集計表を簡単に作れる機能があります。

- オリジナルのデータから特定の2項目（例：性別と年齢層など）で行と列を作り、交わる部分に該当する件数を求めることをクロス集計といい、クロス集計を表に表したものをクロス集計表といいます。
- サンプルデータとして、基本統計量の導出において利用した2017年7月1日～9月30日（92日分）の東京の天候データ（気温と天気）を利用します。
- Excelのピボットテーブルを使うと簡単にクロス集計表を作ることができます。ピボットテーブルは「挿入」タブの中から選択することができます。
  - ピボット（pivot）は、「旋回する軸」「回転軸」を表す英単語です。

「ピボットテーブルの作成」のダイアログボックスが表示されたら、データの範囲を指定して「OK」をクリックして下さい。その後に現れる「フィールドリスト」から、クロス集計表に利用したい変数にチェックを入れます。

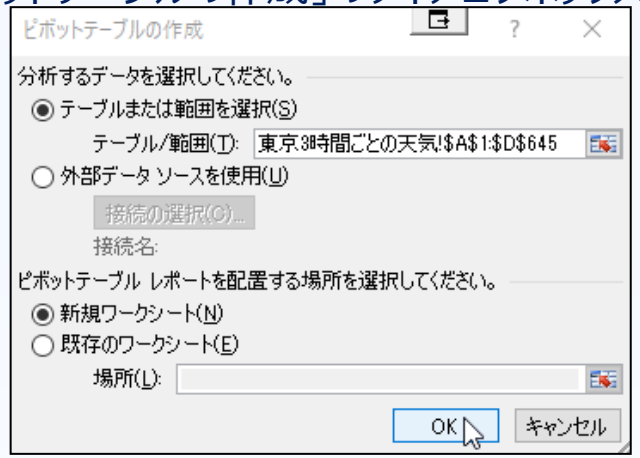
元のデータと同じシート内にピボットテーブルを作成する場合は「ピボットテーブルの作成」のダイアログボックスにおいて、「既存のワークシート」を選択し、ピボットテーブルを作成したい部分のセルを指定します。

「挿入」タブからピボットテーブルを選択

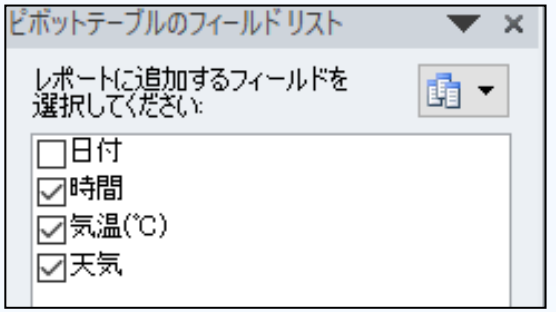


	A	B	C	D
1	日付	時間	気温(℃)	天気
2	7月1日	03時	22.7	曇
3	7月1日	06時	22.8	雨
4	7月1日	09時	24.8	曇
5	7月1日	12時	24.1	曇

「ピボットテーブルの作成」のダイアログボックス



「フィールドリスト」の選択



ピボットテーブルにおける「行ラベル」「列ラベル」「値」の設定

◆ピボットテーブルの作成には、「行ラベル」「列ラベル」と集計対象の値を設定します。

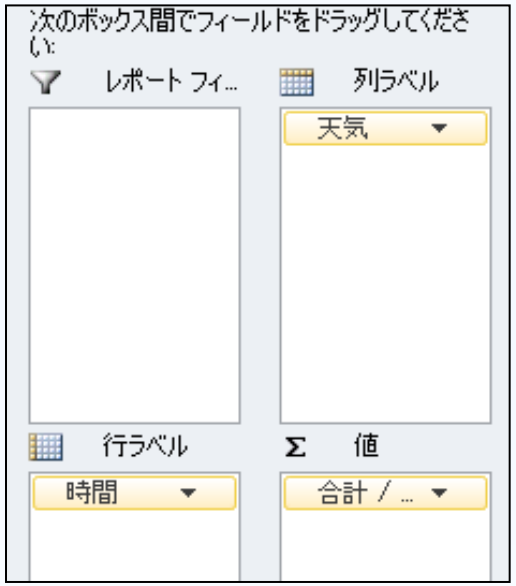
画面右側のフィールドリストから、クロス集計表において横側の行に入れたい項目を「行ラベル」、縦側の列に入れたい項目を「列ラベル」にドラッグ＆ドロップして移動させます。

- ここでは、時間を「行ラベル」、天気を「列ラベル」に移動させます。

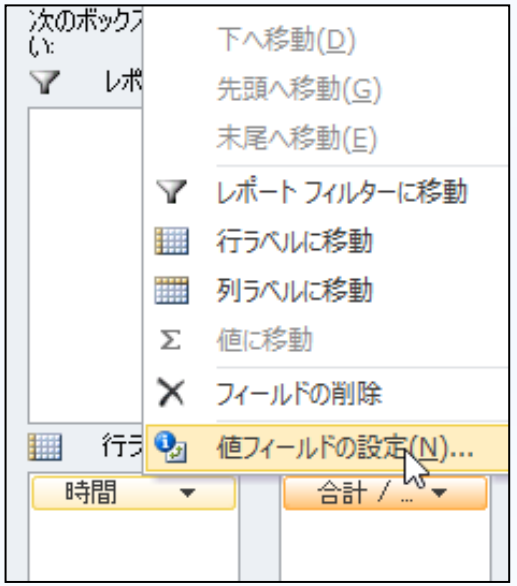
● ピボットテーブルの集計内容は初期設定で［合計値］になっているので、初期設定のままであれば、行と列が交差するセルには合計値が表示されます。

クロス集計表として行と列が交差するセルの頻度を見たい場合は、「合計/…」の部分をクリックし、「値フィールドの設定」から［データの個数］を選択します。
一方で、行と列が交差するセルに関する平均値が見たい場合は［平均］を選択します。

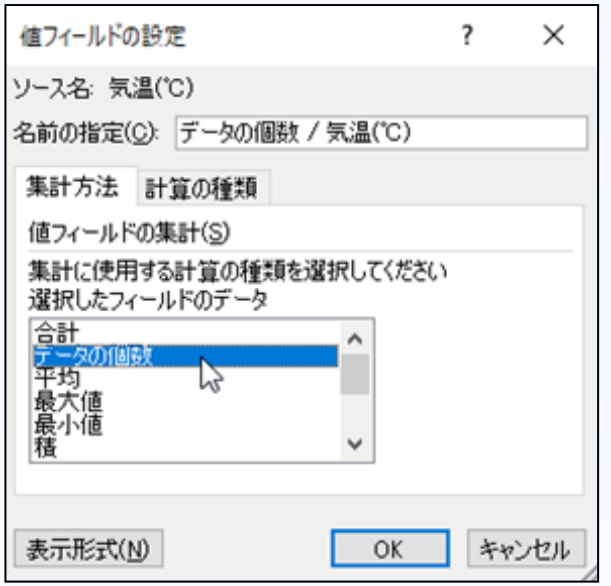
ドラッグ＆ドロップでラベルを設定



値フィールドを変更する



クロス集計表を見る場合は
「データの個数」を選択



ピボットテーブルによる集計表の表示

◆ピボットテーブルでは、データの個数に関するクロス集計表のみならず、平均値等の基本統計量に関する集計表を作成することができます。

- 「値フィールドの設定」において［データの個数］を選択していると、行と列が交わるセルにはデータの件数を示すクロス集計表を表示できます。
  - 2017年7月1日～9月30日（92日分）のデータであるため、クロス集計表の各に関する時点の合計件数は92となり、1日7時点の合計のデータ件数は644件です。
- 「値フィールドの設定」において［平均］を選択していると、行と列が交わるセルにはデータの平均値を示す集計表を表示できます。
  - ピボットテーブルでは［平均］の他にも［最大値］［最小値］［標本分散］［標本標準偏差］といった基本統計量を指定できます。

データの件数を示すクロス集計表

データの個数 / 列ラベル							
行ラベル	雨	快晴	晴れ	曇	薄曇	雷	総計
03時	15	5	22	43	7		92
06時	15	3	18	46	10		92
09時	10	2	20	55	5		92
12時	11	2	29	43	7		92
15時	11	3	31	33	13	1	92
18時	14	3	21	44	9	1	92
21時	12	7	14	45	13	1	92
総計	88	25	155	309	64	3	644

クロス集計表の右側の総計から、各時点に関するデータ件数は92件で、時点合計の標本数が644件となっていることが確認できます。

平均気温に当たる平均値を表示

平均 / 気温 列ラベル							
行ラベル	雨	快晴	晴れ	曇	薄曇	雷	総計
03時	21.6	22.9	24.4	23.7	25.3		23.6
06時	21.5	22.4	24.1	23.7	24.7		23.5
09時	21.7	29.5	27.4	25.9	27.6		26.0
12時	23.6	31.2	30.0	27.1	30.8		28.0
15時	23.8	28.6	29.5	27.3	30.2	25.1	28.0
18時	23.2	26.0	27.0	26.5	28.2	24.6	26.2
21時	22.2	24.9	24.8	25.4	25.8	22.7	24.9
総計	22.5	25.6	27.2	25.6	27.5	24.1	25.7

ピボットテーブルの平均値から、雨が降っている時間は、平均気温が下がっていること、夜間は日中よりも平均気温が低いことが確認できます。

□ ピボットテーブルを使うと、データの件数や基本統計量を簡単に表に示すことができます。