第5章

データ分析の実践演習 I

この章から，具体的なデータセットを題材に，色々な切り口でデータ分析を行っていきます。まずは，第 4 章までに学んだ範疇でできる分析演習から始めます。まずは記載の通りに追体験を行って，ビッグデータの分析に慣れていきましょう。定着を図ることを希望する場合は，具体的な手順が書かれたところを極力見ないで，再現できるかどうか試すようにしましょう。

5.1 購買データの分析 1（グラフの使い分け）

表のような架空のコンビニの 200 件の購買データ（年齢・購入額）について，利用客の年齢層と購入金額の関係について分析して，いろいろなグラフの使い分けを学んでいきます。

	A	B	C
1	No	年齢	購入額
2	1	56	440
3	2	65	2000
4	3	13	530
5	4	36	1470
6	5	56	520
7	6	26	860
8	7	49	370
198	197	21	730
199	198	27	350
200	199	62	920
201	200	51	900

まずはデータをみて，言えそうなことを考えて（できれば仮説を立てて）みてから分析していきましょう。

STEP1. 利用者の年代別人数を調べる。

まず利用者の年代別人数を調べます。いくつか方法があるので紹介します。

方法 1（年齢でソートして，調べる）

データを別の場所(例えば M ～ O 列)に移して，年齢を基準に並べ替えを行います。［データ］→［範囲を並べ替え］→［並べ替えオプション］で，年齢のある列を基準に昇順に並べ替えをします。

あとは，ソートされた表を見て人数を数えます。

M	N	O
No	年齢	購入額
3	13	530
28	13	140
50	13	200
62	13	120
187	13	100
58	14	210

方法 2（COUNTIF 関数を利用）

元のデータの B 列から，数値条件を指定して数えていきます。累積度数を先に調べると楽です。

19 歳以下の人数を調べるために，G3 セルに

```
=COUNTIF(B$2:B$201,"<=19")
```

と入力し，一度 G7 までオートフィルで複写します。そしてあとから G4 セルは「29」，G5 セルは「39」と変えていきます。度数欄に戻って，F3 セルは「=G3」，F4 セルは「=G4-G3」と入力して，F4 から F7 へオートフィルで複写します。最後に相対度数は，H3 セルに「=F3/200」と入力して，H7 へオートフィルで複写します。

G3 =COUNTIF(B$2:B$201,"<=19")

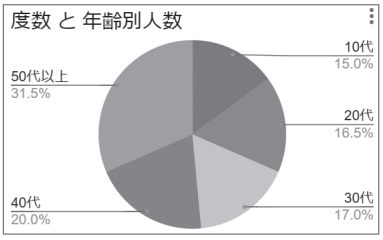
	D	E	F	G	H
1					
2		年齢別人数	度数	累積度数	相対度数
3		10代	30	30	0.15
4		20代	33	63	0.165
5		30代	34	97	0.17
6		40代	40	137	0.2
7		50代以上	63	200	0.315

方法 3（COUNTIFS 関数を利用）

度数から埋める場合は，条件が複数必要になるので，COUNTIFS 関数を使います。

F4 セルには，「=COUNTIFS(B$2:B$201,"<=29",B$2:B$201,">=20")」を入力して，F7 セルまでオートフィルで複写し，条件の数値を変えていきます。

最後に E2:F7 セルを範囲選択して,「円グラフ」で可視化します。相対度数を表示しなくとも，割合が表示できます。「比較的どの年代も均等に利用している店舗」であることがわかります。



STEP2. 購入額の度数分布とヒストグラムを出力

FREQUENCY 関数を用いるときは，「階級の上限値」を用います。度数分布のラベルは手入力でも容易ですが，例えば，E11 セルに「=(F10+1)&"~"&F11」と入力して E14 セルまでオートフィルで複写することでも記述できます。

E11 | =(F10+1)&"～"&F11

	D	E	F
8			
9		度数分布	階級上限値
10		～200	200
11		201～500	500
12		501～1000	1000
13		1001～1500	1500
14		1501～2000	2000
15		2001～	3000

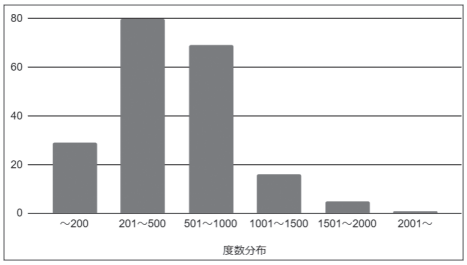
G10 セルに「=FREQUENCY(C2:C201,F10:F14)」と入力すると，全階級の度数分布が出力できます（F15 セルはその他扱いのため外しています）。

G10 | =FREQUENCY(C2:C201,F10:F14)

	D	E	F	G
8				
9		度数分布	階級上限値	度数
10		～200	200	29
11		201～500	500	80
12		501～1000	1000	69
13		1001～1500	1500	16
14		1501～2000	2000	5
15		2001～	3000	1

E9 ～ G15 セルを範囲選択して，「縦棒グラフ」で描きます（グラフエディタで系列から「階級上限値」を削除します）。

200 円～ 1000 円の購入額の人が多いことがわかります。



STEP3. 年代別の購入額の四分位数と箱ひげ図を出力

次に，購入額のばらつき具合を調べるために，四分位数を年代別に調べることにします。

STEP1 の方法 1 にあるように，購入額のデータを M ～ O 列にコピーして，年齢順にソートしておきます。四分位数は QUARTILE 関数を用います。2 つ目の引数は「第 n 四分位数の n」に等しく，最大値は第 4 四分位，中央値は第 2 四分位，最小値は第 0 四分位とみなします。

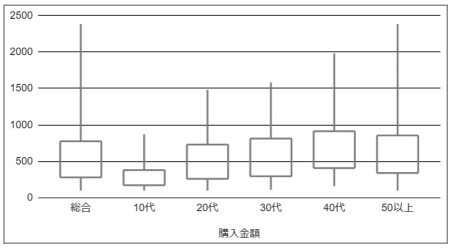
（オートフィルが十分活かせるわけではないため，各年代（列）の最大値の欄を埋めた後，縦にオートフィルで複写し，引数を少しずつ変えていくようにします。）

例えば，G18 セルには「=QUARTILE($O$2:$O$31,4)」と入力し，縦と横それぞれオートフィルを行い，セル番号と引数を適宜変えていくようにします。

G18 ▾ | fx =QUARTILE($O$2:$O$31,4)

	D	E	F	G	H	I	J	K
16								
17		購入金額	総合	10代	20代	30代	40代	50以上
18		最大値	2380	870	1480	1580	1980	2380
19		第3四分位数	775	380	730	812.5	912.5	855
20		第1四分位数	280	172.5	260	295	407.5	340
21		最小値	100	100	100	110	160	100
22		中央値	475	255	420	520	590	520

E17:K22 を範囲選択して，グラフは「ローソク足チャート図」（グラフの選択画面で下方にスクロールします。4.3 節参照）を選択し，図のようにチェックボタンを選択すると，箱ひげ図として出力されます。



これを見ると，「20 代以上の箱の位置は変わらない（中央値前後 50% の人の購入金額はほぼ同じ）」ものの，「第 3 四分位数以上の購入額の人は，年齢が上がるにつれて増えていく」ことがわかります。

STEP4. 年代別の購入額度数分布とヒストグラム・帯グラフを出力

STEP3 同様 M ～ O 列の年齢順にソートしたデータを利用して，例えば，G25 セルに

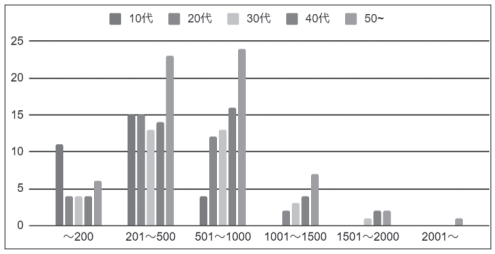
```
=FREQUENCY($O$2:$O$31,$F$25:$F$29)
```

と入力して K25 セルまでオートフィルで複写し，O 列のセル番号を修正します。

G25 =FREQUENCY($O$2:$O$31,$F$25:$F$29)

	D	E	F	G	H	I	J	K
23								
24		度数分布	階級上限値	10代	20代	30代	40代	50~
25		~200	200	11	4	4	4	6
26		201~500	500	15	15	13	14	23
27		501~1000	1000	4	12	13	16	24
28		1001~1500	1500	0	2	3	4	7
29		1501~2000	2000	0	0	1	2	2
30		2001~	3000	0	0	0	0	1

E24:K30 を範囲選択して，「縦棒グラフ」として出力すると下のようになります（系列から階級上限値は削除します）。



さらに，各年代の購入額の人数比をみるには，**帯グラフ**（「**100% 積み上げ横棒グラフ**」）を用います。データ範囲は 2 つに分けてグラフエディタで変更をします（「E24:E30,G24:K30」と記述し，左右に範囲を結合します）。その他次図のようにチェックボタンを選択します。