

8日目：項目のチェック（2）

昨日は、平均値などの基礎統計量を計算する試行錯誤へご招待しましたが（?）、今日は簡単にやってみます。そのためには、**psych** というパッケージが必要となりますが、パッケージのインストール&読み込みの詳しい方法は、後で説明します。

以下の説明は、**psych** パッケージのインストールが済み、いつもの練習用のファイルを読み込んでいるというところから始めます。

せっかくパッケージをインストールしてもらったのですが、先に **psych** パッケージを使わない、**summary** というコマンドを使ってみます。まずは以下のように入力し、実行してみてください。

summary(x)

ずらずらと、基礎統計量が出てきます。何が算出されているかをチェックすると、最小値 (Min.)、第1四分位 (1st Qu.)、中央値 (Median)、平均値 (Mean)、第3四分位 (3rd Qu.)、最大値 (Max.)、そして、あれば「NA」の数 (NA's) です。

Summaryは、(x)と指定しても警告は出てこないし、最小値、最大値も変数ごとにやってくれるし、「NA」も自動的に省いてくれるし、その数も出してくれる…と、いいことが多いのですが、問題は標準偏差を計算してくれないところ…。社会調査のようなデータには向くかもしれませんが、心理統計にはこれは痛い。

そこで、**psych**パッケージに登場してもらいます。

パッケージを使うには、まず欲しいパッケージを①インストールし、さらにそれを②読み込む必要があります。Rを起動しただけでは、パッケージは読み込んでくれません。インストールの方法は後で詳しく説明するとして、先に読み込みの方法から簡単に…

library(psych)

と、まずは入力します (**library()**は、パッケージを読み込むための関数です)。これを実行しても、Rコンソールには何の変化もありません。次に、

describe(x)

と入力して実行します。すると、欲しかった数値が！

解説するまでもないでしょうが、左から変数名、列番号(**var**)、ケース数(**n**)、平均値 (**mean**)、標準偏差(**sd**)、中央値(**median**)、トリムされた平均値(**trimmed**)、MAD: median absolute deviation (**mad**)、最小値(**min**)、最大値(**max**)、レンジ(**range**)、歪度(**skew**)、尖度(**kurtosis**)、標準誤差(**se**)です。ちなみに、「MAD、トリムとはなんぞや?」と思う人は、統計の本を読むなり、ググるなりしてください。

```
> library(psych)
> describe(x)
      var  n   mean  sd median trimmed  mad  min  max range  skew kurtosis  se
no      1 20 1010.50 5.92 1010.5 1010.50 7.41 1001 1020   19  0.00   -1.38 1.32
性別    2 20   1.70 0.47   2.0   1.75 0.00   1    2    1 -0.81   -1.41 0.11
年齢    3 20  19.35 0.59  19.0  19.25 0.00  19   21    2  1.30    0.58 0.13
b1      4 20   2.95 1.28   3.0   2.94 1.48   1    5    4  0.23   -1.26 0.29
b2      5 20   3.85 0.93   4.0   3.88 1.48   2    5    3 -0.09   -1.30 0.21
b3      6 20   4.45 0.69   5.0   4.56 0.00   3    5    2 -0.76   -0.72 0.15
b4      7 20   4.00 0.73   4.0   4.00 0.74   3    5    2  0.00   -1.19 0.16
b5      8 20   3.25 0.97   3.0   3.19 1.48   2    5    3  0.19   -1.12 0.22
```

さて、パッケージのインストールと読み込みについてです。方法はいくつもあります。

《インストール》

- メニューバーから

メニューバーの「パッケージ」をクリックし、「パッケージのインストール」を選択します。すると右図のようなパッケージの一覧が出てくるので、使いたいパッケージを選択します。

- コマンドを使って

`install.packages("パッケージ名")`と入力します。

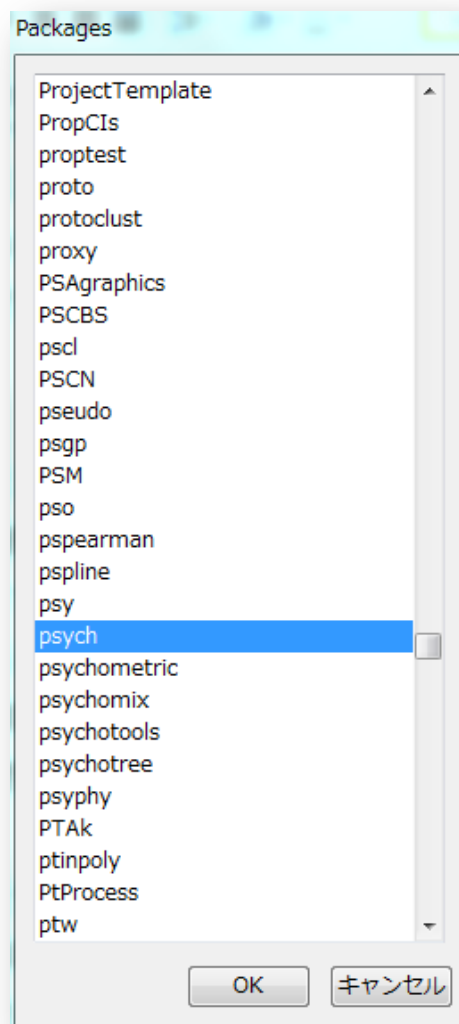
《読み込み》

- メニューバーから

メニューバーの「パッケージ」をクリックし、「パッケージの読み込み」を選択します。インストールのときのように、パッケージの一覧が出てくるので、読み込みたいパッケージを選択します。

- コマンドを使って

これは先にも説明しましたが、`library(パッケージ名)`と入力します。



次は、Rの命令の中身を見たり、ヘルプを見たりする方法です。Rコンソールの方で良いので、以下だけ（変数指定をしない）を入力して実行してください。

describe

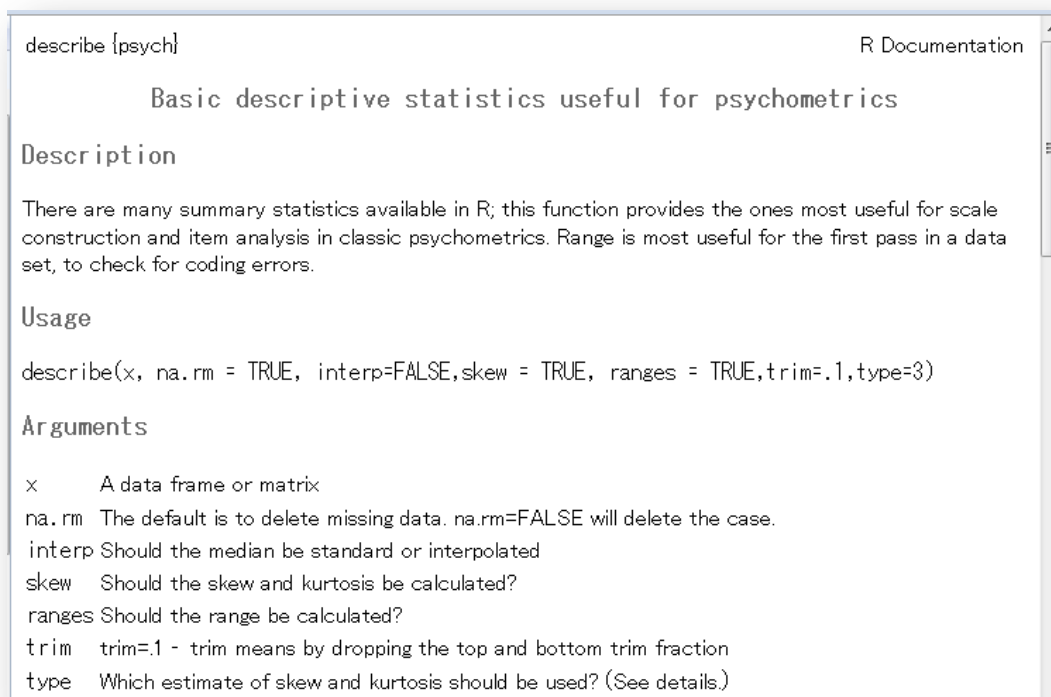
すると、一見でコンピュータのプログラムらしきものが表示されると思います。その通りで、これが **describe** の中身（プログラム）なのです。このようにすれば、中身を見ることができます（できないものも結構あります）。

次には、以下のように入力し、実行してください。

?describe

こちらは新しいウインドが開きます（私の場合、いつも利用しているインターネットブラウザで開きます）。これはRのヘルプ画面です。英語ですが、嫌がらずに眺めてみてください。まず **Description** で、概要の説明がされています。

Usageは、コマンドの詳しい説明です。そこには、**describe(x, na.rm = TRUE, interp = FALSE, skew = TRUE, ranges = TRUE, trim = .1, type = 3)**と記載されています。**na.rm = TRUE**以下はデフォルトの設定であり、何も指定しなければこの通りに実行されます。試しに、**describe(x)**の結果と、**describe(x, na.rm = TRUE, interp = FALSE, skew = TRUE, ranges = TRUE, trim = .1, type = 3)**の結果を比べてみてください。同じ出力結果になります。昨日は書かなければならなかった**na.rm = TRUE**も、**describe**では不要だったのも、それがデフォルトの設定だったからです。たとえば**mean**のヘルプを見ると、そちらでは**na.rm = FALSE**がデフォルトであることがわかります。



```
describe {psych}                                R Documentation
-----
Basic descriptive statistics useful for psychometrics

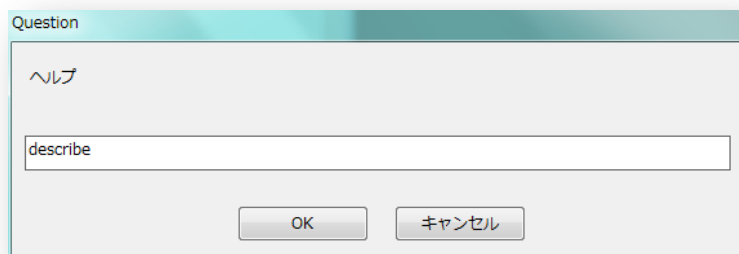
Description
-----
There are many summary statistics available in R; this function provides the ones most useful for scale construction and item analysis in classic psychometrics. Range is most useful for the first pass in a data set, to check for coding errors.

Usage
-----
describe(x, na.rm = TRUE, interp=FALSE, skew = TRUE, ranges = TRUE, trim=.1, type=3)

Arguments
-----
x      A data frame or matrix
na.rm  The default is to delete missing data. na.rm=FALSE will delete the case.
interp Should the median be standard or interpolated?
skew   Should the skew and kurtosis be calculated?
ranges Should the range be calculated?
trim   trim=.1 - trim means by dropping the top and bottom trim fraction
type   Which estimate of skew and kurtosis should be used? (See details.)
```

さらに下の方には、**Examples** もあります。このヘルプにはいろんな情報がありますので、積極的に見るようにしておくと、いろんな発見があると思います。

ちなみに、ヘルプを参照するには?**describe**以外にもあります。メニューバーの「ヘルプ」から「Rの関数（テキスト）」を選択する方法です。下のような画面が開いたら、ヘルプを見たい関数を入力します。



では、話を戻して、次に男女別に基礎統計量を求めることをやってみます。コマンドは **describe.by** です。まずはヘルプ探して見ることで、これ使い方を試行錯誤してみてください。

Examples もありますが、簡単な設定は以下のようなでしょう。これで性別に計算をしてくれます。

describe.by(x, x\$性別)

さて今日の最後に、このRでの計算結果をエクセルに移すことをやってみます。Rの出力のままでは論文の表としては使えません。何とかして右のような表に仕上げる必要があるでしょう。エクセルに結果を移すのはファイルを介してもできますが、簡単なのはコピペです。

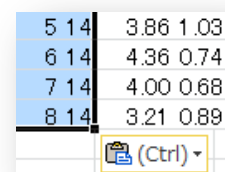
		人数	平均値	標準偏差	最小値	最大値
男性	年齢	6	19.17	0.41	19	20
	b1	6	2.83	1.47	1	4
	b2	6	3.83	0.75	3	5
	b3	6	4.67	0.52	4	5
	b4	6	4.00	0.89	3	5
	b5	6	3.33	1.21	2	5
女性	年齢	14	19.43	0.65	19	21
	b1	13	3.08	1.26	2	5
	b2	14	3.86	1.03	2	5
	b3	14	4.36	0.74	3	5
	b4	14	4.00	0.68	3	5
	b5	14	3.21	0.89	2	5

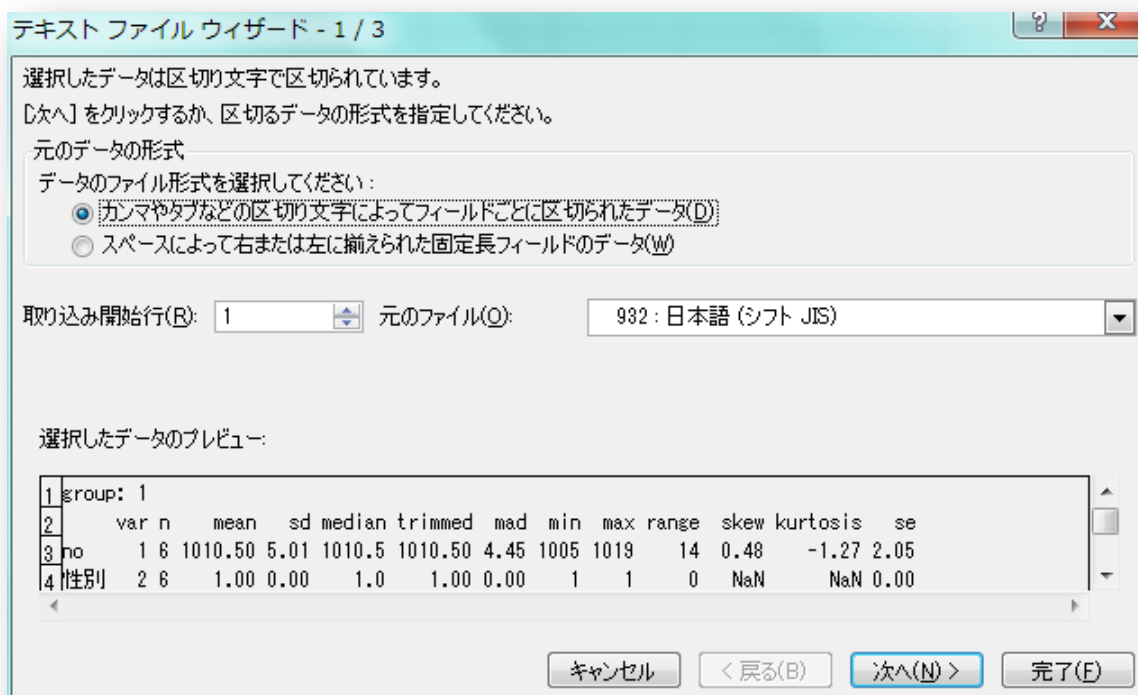
まず、R コンソールの **describe.by** の結果部分をコピーします。そしてエクセルのシートにペーストします。すると以下の図のようになります。

	A	B	C	D	E	F	G	H				
1	group: 1											
2	var n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
3	no 1 6	010.50	5.01	1010.5	1010.50	4.45	1005	1019	14	0.48	-1.27	2.05
4	性別 2 6	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
5	年齢 3 6	19.17	0.41	19.0	19.17	0.00	19	20	1	1.36	-0.08	0.17
6	b1 4 6	2.83	1.47	3.5	2.83	0.74	1	4	3	-0.39	-2.00	0.60
7	b2 5 6	3.83	0.75	4.0	3.83	0.74	3	5	2	0.17	-1.54	0.31
8	b3 6 6	4.67	0.52	5.0	4.67	0.00	4	5	1	-0.54	-1.96	0.21
9	b4 7 6	4.00	0.89	4.0	4.00	1.48	3	5	2	0.00	-1.96	0.37
10	b5 8 6	3.33	1.21	3.5	3.33	1.48	2	5	3	0.04	-1.88	0.49
11	-----											
12	group: 2											
13	var n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
14	no 1 14	1010.50	6.44	1011.0	1010.50	8.15	1001	1020	19	-0.10	-1.63	1.72
15	性別 2 14	2.00	0.00	2.0	2.00	0.00	2	2	0	NaN	NaN	0.00
16	年齢 3 14	19.43	0.65	19.0	19.33	0.00	19	21	2	1.04	-0.20	0.17
17	b1 4 14	3.00	1.24	2.5	2.92	0.74	2	5	3	0.67	-1.31	0.33
18	b2 5 14	3.86	1.03	4.0	3.92	1.48	2	5	3	-0.14	-1.52	0.27
19	b3 6 14	4.36	0.74	4.5	4.42	0.74	3	5	2	-0.58	-1.13	0.20
20	b4 7 14	4.00	0.68	4.0	4.00	0.00	3	5	2	0.00	-0.99	0.18
21	b5 8 14	3.21	0.89	3.0	3.17	1.48	2	5	3	0.22	-0.95	0.24
22												

次に、ペーストしたときに出てくるアイコン（右図）をクリックします。

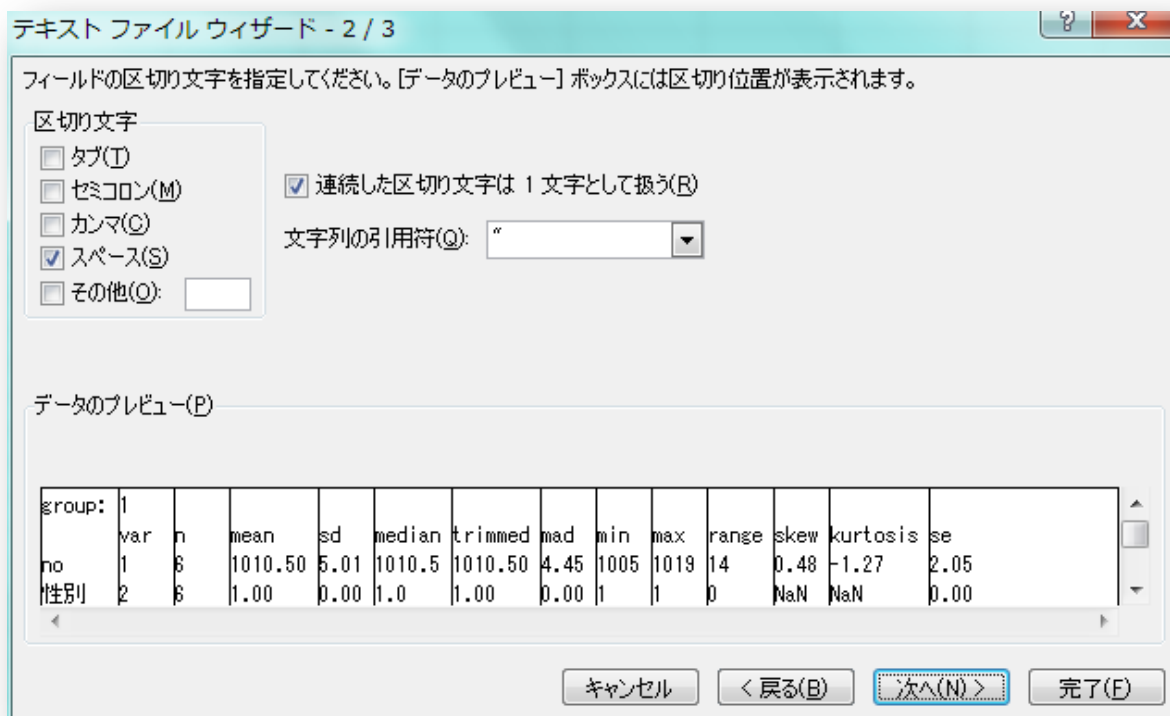
開いたメニューのうち下側の「テキスト ファイル ウィザードを使用する」を選択します。すると、次の図のようなウインドが開きます。





この画面ではさわるところはありません。R からコピーしてきたデータは、スペース（空白）によって整形されています。しかし、それは「固定長」ではないので、「データのファイル形式」は「区切り記号付き」のままでOKです。

「次へ」をクリックします。



このウィザードは、結構うまく区切りをつけてくれます。「区切り文字」で「スペース」を指定しなくても、たいていは「スペース」にチェックが入っていると思います。

また「データのプレビュー」には、区切りの部分に縦線が入っています。このまま続けると、この線の部分でデータを区切ってくれます。

これ以上特に触る部分もないので、「完了」をクリックします。

すると以下のように数値がセルに分けられていると思います。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	group:	1													
2	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se		
3	no	1	6	1010.5	5.01	1010.5	1010.5	4.45	1005	1019	14	0.48	-1.27	2.05	
4	性別	2	6	1	0	1	1	0	1	1	0	NaN	NaN	0	
5	年齢	3	6	19.17	0.41	19	19.17	0	19	20	1	1.36	-0.08	0.17	
6	b1	4	6	2.83	1.47	3.5	2.83	0.74	1	4	3	-0.39	-2	0.6	
7	b2	5	6	3.83	0.75	4	3.83	0.74	3	5	2	0.17	-1.54	0.31	
8	b3	6	6	4.67	0.52	5	4.67	0	4	5	1	-0.54	-1.96	0.21	
9	b4	7	6	4	0.89	4	4	1.48	3	5	2	0	-1.96	0.37	
10	b5	8	6	3.33	1.21	3.5	3.33	1.48	2	5	3	0.04	-1.88	0.49	
11	-----														
12	group:	2													
13	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se		
14	no	1	14	1010.5	6.44	1011	1010.5	8.15	1001	1020	19	-0.1	-1.63	1.72	
15	性別	2	14	2	0	2	2	0	2	2	0	NaN	NaN	0	
16	年齢	3	14	19.43	0.65	19	19.33	0	19	21	2	1.04	-0.2	0.17	
17	b1	4	14	3	1.24	2.5	2.92	0.74	2	5	3	0.67	-1.31	0.33	
18	b2	5	14	3.86	1.03	4	3.92	1.48	2	5	3	-0.14	-1.52	0.27	
19	b3	6	14	4.36	0.74	4.5	4.42	0.74	3	5	2	-0.58	-1.13	0.2	
20	b4	7	14	4	0.68	4	4	0	3	5	2	0	-0.99	0.18	
21	b5	8	14	3.21	0.89	3	3.17	1.48	2	5	3	0.22	-0.95	0.24	
22															

ここまできたら、後はエクセルで整形するだけですから、先のような表に仕上げるのはすぐでしょう。

これで8日目は終了です。明日は度数分布表を作ってみます。