

R Commander を用いた統計解析の基礎(2)

岡田 昌史

1. 変数の操作

それでは、前回の復習からいってみましょう.まずは身長(height)と体重(weight)から、BMIを 求めて、それをデータセットに新しい列として追加してみます.

BMI は、体重 / (身長)² ですから、height と weight を使って書くと、weight / (height ^2) あるいは、weight / (height * height) などと書けますね。R では、かけ算を " * "、割り算を " / "、ベき乗を " ^ "であらわします。これは、Excel などと一緒ですね。

現在,データセット figures6 には, height と weight の 2 つの列(このように,名前のついた一 連のデータを「変数」,正確には「ベクタ変数」と呼びます)しかありませんが,ここに BMI を新 しい変数として追加してみましょう.「データ」メニューから,「アクティブデータセット内の変数 の管理」を選び,「新しい変数を計算...」をクリックします.



「新しい変数の計算」ウィンドウが出ますから、「新しい変数名」のところに "BMI", 「計算式」のところに "weight / (height ^ 2) " をいれて, 「Ok」を押しましょう.

2007/8/8

R Commander を用いた統計解析の基礎



☆新しい変数の計算			
現在の変数(ダブルクリッ	クして計算式のテキ	キストボックスに入力)	-
height weight		<u> </u>	
↓ ^{weight} 新儿\\实数么	計算式	-	
BMI	weight /	(height ^ 2)	
OK + *	・ンセル	ヘルプ	

おっと、ここでもし、

i	7/		
	2 mil) 2	変数 BMIがる 変数 に上	すでに存在します。 :書きしますか?
		Yes	No

というのが出てしまった方は、おそらく前回の演習がうまく進んでおり、すでに "BMI" 変数が作成されているということです. 上書きしてもかまいませんので、"Yes" を選んでください.

それでは、うまく変数が追加されているか、「データセットを表示」で確認してみましょう.

74	figures6		
	height	weight	BMI
1	1.75	60	19.59184
2	1.	72	22.22222
3	1.63	57	20.93664
4	1.90	90	24.93075
5	1.74	95	31.37799
6	1.91	72	19.73630

このように、3列目に「BMI」が追加されていれば成功です.

2. データの可視化

それでは、次にRコマンダーを用いて、データの特性を示すさまざまな表やグラフを描いてみま しょう.あらゆるデータ解析において、まず最初に必要になるのは、データを表やグラフにまとめ て、その全体像を把握することです.Rは「対話的」な統計解析環境なので、可視化の機能は比較 的充実しています.



ここでは、さまざまなデータを扱いますが、その1つ1つを毎回 Excel などから読み込んでいて は大変なので、書籍「R による医療統計学」(原書 "Introductory Statistics with R")に掲載されてい るサンプルデータがまとめて収録されているパッケージ、「ISwR」を読み込んで利用します.

まずは,	ISwR パック	アージを読み込ん	でみましょう.	こんな感じですね?
------	----------	----------	---------	-----------



ISwR パッケージを読み込むと, R コマンダーの「データ」メニュー,「パッケージ内のデータ」 から,「アタッチされたパッケージからデータセットを読み込む...」でいろいろなデータを読み込 めるようになります.

アクテ

2007/8/8

R Commander を用いた統計解析の基礎

∽んパッケージからデータを読み込む	
パッケージ(ダブルクリックして選択) car datasets <mark>ISwR</mark> survival	データセット(ダブルクリックして選択) hellung intake juul juul2
または データセット名を入力: OK キャンセル	ヘルブ

「パッケージからデータを読み込む」ウィンドウが出たら,「パッケージ」のところで "ISwR", 「データセット」では "juul" を選択し, Ok を押してください. これで, ISwR パッケージに内蔵さ れている juul という名前のデータセットが読み込まれ, アクティブデータセットとなります.

6	٢	3	0
	ΒY	NC	SA

<mark>7%</mark> juul							×
	age	menarche	sex	igA	tanner	testvol	
1	NA	NA	NA	90	NA	NA	
2	NA NA	NA	NA	88	NA	NA	
3	NA	NA	NA	164	NA	NA	
4	NA	NA	NA	166	NA	NA	
5	NA	NA	NA	131	NA	NA	
6	0.17	NA	1	101	1	NA	
7	0.17	NA	1	97	1	NA	
8	0.17	NA	1	106	1	NA	
9	0.17	NA	1	111	1	NA	
10	0.17	NA	1	79	1	NA	
11	0.17	NA	1	43	1	NA	
12	0.17	NA	1	64	1	NA	
13	0.25	NA	1	90	1	NA	
14	0.25	NA	1	141	1	NA	
15	0.42	NA	1	42	1	NA	
16	0.50	NA	1	43	1	NA	
17	0.67	NA	1	132	1	NA	
18	0.75	NA	1	43	1	NA	
19	0.75	NA	1	36	1	NA	
20	1.00	NA	1	86	1	NA	
21	1.16	NA	1	44	1	NA	
22	1.50	NA	1	68	1	NA	
23	1.50	NA	1	89	1	NA	
24	1.58	NA	1	101	1	NA	
25	1.67	NA	1	115	1	NA	
26	1.67	NA	1	53	1	NA	
27	1.75	NA	1	94	1	NA	
28	1.83	NA	1	95	1	NA	
29	1.92	NA	1	76	1	NA	
30	2.00	NA	1	79	1	NA	-

「データセットを表示」をさせて みると、ウィンドウの右側にス クロールバーがでていますね. このデータは 1339 行もある大 きなもので, age, menarche, sex,igf1,tanner,testvol という 6 つの変数が格納されています. これは、学校での健康診断の際 に, 思春期の子供のインスリン 様成長因子(IGF-I)の分布を調べ たデータです. "NA" という値が 目立ちますね. これは, "Not Available"の略,つまり,いわ ゆる欠損値です. R では欠損値 のことを一貫して "NA" と呼びま す.

それではまず,このデータの特性をあらわす「記述統計量」を表にまとめて出力してみましょう.「統計量」メニューから「要約」を選び,さらに「アクティブデータセット」を選択してください.

7% R コマンダー			_
ファイル 編集 データ	統計量 グラフ モデル	分布 Demos ツール ヘルプ	
Rada データセット: ju スクリプトウィンドウ data(juul, pac: showData(juul,	要約 分割表 平均 日本 分散 ノンパラメトリック検定 次元解析 モデルへの適合	アクティブデータセット 数値による要約 頻度分布 欠測値を数える 層別の統計量 相関行列 相関の検定 シャピロ・ウィルクの正規性の検定	フティブモデルなし> ont'), maxwidth=80, max:

2007/8/8

R Commander を用いた統計解析の基礎



出力ウィンドウのほうに、青文字で要約結果が出力されたことと思います.

出力ウィンドウ			
> showData(juul,	placement='-20+200	', font=getRcmdr	('logFont'), maxwic
> summary(juul)			
age	menarche	sex	igf1
Min. : 0.170	Min. : 1.000	Min. :1.000	Min. : 25.0
1st Qu.: 9.053	1st Qu.: 1.000	1st Qu.:1.000	1st Qu.:202.2
Median :12.560	Median : 1.000	Median :2.000	Median :313.5
Mean :15.095	Mean : 1.476	Mean :1.534	Mean :340.2
3rd Qu.:16.855	3rd Qu.: 2.000	3rd Qu.:2.000	3rd Qu.:462.8
Max. :83.000	Max. : 2.000	Max. :2.000	Max. :915.0
NA's : 5.000	NA's :635.000	NA's :5.000	NA's :321.0
tanner	testvol		
Min. : 1.000	Min. : 1.000		
1st Qu.: 1.000	1st Qu.: 1.000		
Median : 2.000	Median : 3.000		
Mean : 2.640	Mean : 7.896		
3rd Qu.: 5.000	3rd Qu.: 15.000		
Max. : 5.000	Max. : 30.000		
NA's :240.000	NA's :859.000		
		5	
4			

これは、アクティブデータセットを構成するそれぞれの変数について、その値の最小値,4分位 値、平均値、最大値、欠損値の数をもとめたものをまとめて表示しています.これらの値をみるこ とで、そのデータがどのように「分布」しているのか、すなわち、どちらかというと小さい値が多 いのか、大きい値が多いのか、また平均値と中央値にどのくらいずれがあるのか、欠損値はどのく らいあるのか、といったことがすぐにわかります.

ただ, sex, menarche, tannerの各変数については, 4分位値や最大, 最小値がちょっと「きり のいい」値になりすぎている感じです. また, testvol に関しては, 欠損値が多すぎることがわか るでしょう. これは, sex, menarche, tanner はそれぞれ性別, 初経の有無, ターナーの成熟度の 5段階分類を示しているためです. これらはカテゴリカルデータなのですね.

そして、testvol に関しては、精巣体積です. つまり、男子にしか測定していないため、女子では欠損値になっているのです.

それでは次に、データの分布を知るために用いるもっとも基本的な図であるヒストグラムを描い てみましょう.「グラフ」メニューから「ヒストグラム…」を選んでください.



76 R コマンダー		
ファイル 編集 データ 統計量	グラフ モデル 分布 ロ	Demos ツール ヘルプ
Randa F-Qtyr: juul F-	色パレット インデックスプロット	ットを表示 モデル: <mark><アクティブモデルなし></mark>
スクリプトウィンドウ	ヒストグラム	
data(juul, package="] showData(juul, placen summary(juul)	幹葉表示… 人 箱ひげ図… QQブロット…	nt=getRcmdr('logFont'), maxwidt
.FA <- factanal(~age4	散布圆 数左回结网	ctors=1, rotation="varimax", so
.FA	取伸込1729… 垢れ線ガラフ	
remove(.FA)	2014 60家グラフ 冬代付き数左回	
sapply(juul, functior	来住りて取勿凶… 死物のつかに)) # NA counts
	〒2300フロット 株式ニュ	
	1年クラブ…	
	□ □ りフノ □ つ: 与二 ピニュ →	
•	3/八元クフノー ・	
出力ウィンドウ ・	クラノをノア1 ルビ1未1子 『	
0.677 0.940	0.005	
0.077 0.049	0.005	
Loadings:		
Fector1		
raccorr		

「ヒストグラム」ウィンドウがでたら、変数として「age」を選択、あとはそのままで「Ok」を 押してみましょう.

7 % ヒストグラム	_ 🗆 🗙
変数(1つを選択)	
age igf1 menarche sex	
区間の数: <自動>	
軸の尺度	
頻度の計算 💽	
バーセント 〇	
密度 〇	
OK キャンセル	ヘルフ

ヒストグラムが無事出力されたでしょうか?年齢分布が圧倒的に 20 歳以下にかたよっていることがわかるでしょう.



_ 🗆 🗙

R Graphics: Device 2 (ACTIVE) ファイル 履歴 サイズ変更



デフォルトでは、ヒストグラムの棒1本あたりに含まれる年齢の範囲は、自動的に決まってしま います.このデータでは少数の高い年齢層の影響を受けて、年齢の範囲が広めになっているようで す.もう少し細かいヒストグラムにしてみましょう.「ヒストグラム」ウィンドウで、「区間の数」 を 40 ぐらいにします.

74 ヒストグラム	
変数(1つを選択)	
igf1 menarche sex	
区間の数: 40	
軸の尺度	
頻度の計算 💽	
パーセント 〇	
密度 〇	
OK キャンセル	ヘルブ



区間分けが細かくなったグラフが出力されました。14-16 歳あたりや、2-4 歳あたりに落ち込ん でいる部分があることが検出できましたね。

次に箱ひげ図を見ていきたいと思いますが,これは層別にしたほうが見やすくなりますから,ま ずはこのデータに含まれる,数値で入力されているカテゴリカルデータ(sex, menarche, tanner)が カテゴリカルデータであることをRに教えてあげましょう.

Rでは、カテゴリカルデータのことを「因子(Factor)」と呼びます。因子分析などの因子とは若 干ニュアンスが違い、データの層別に使われる要因、といった意味合いのようです。

「データ」メニューから「アクティブデータセット内の変数の管理…」を選び, さらに「数値変数を因子に変換…」を選んでください.



74 R コマンダー	
ファイル 編集 データ 統計量 グラフ モデル 分布	Demos ツール ヘルプ
Read データセッ スクリプトウィ データセットの日ード オータセッ データセットの日ード ガータワンボート データのインボート Hist(juul アクティブデータセット boxplot(j) アクティブデータセット内の変数の管理 data(IgM, アクティブデータセット内の変数の管理 data(red.cell.folate, package="ISwF showData(red.cell.folate, placement boxplot(folate~ventilation, ylab="f summary(red.cell.folate) tapply(red.cell.folate) tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate tapply(red.cell.folate	セットを表示 モデル: マクティブモラ eaks=40, col="darkgray") 変数の再コード化 新しい変数を計算 データセットに観測値番号を追加 変数の標準化 数値変数を区間で区分 因子水準を再順序化 因子に対する対比を定義 変数名をつけ直す データセットから変数を削除
> showData(red.cell.folate, placement	nt='-20+200', font=getRcmdr

「数値変数を因子に変換」ウィンドウが出たら、「変数」で「menarche」「sex」「tanner」を選ん で、「新しい変数名または複数の変数に対する接頭文字列」に「c.」と入力しましょう. こうする ことで、これらの変数のコピーに、「これはカテゴリカルデータですよ」という印をつけた新しい 変数、つまり因子化した変数は、もとの変数名の頭に "c." をつけた、"c.menarche" とか "c.sex" と か "c.tanner" という名前で作成されます.

7%数値変数を因子に変換		
変数 (1 つ以上選択)	因子水準	
igf1	水準名を指定	•
sex	数値で	0
tanner 🗾		
新しい変数名または複数の変数に対する接頭文字列:	c.	
OK キャンセル へ.	ルプ	
uacarruur, packaue- IJWK INT		

「Ok」を押すと、まずは menarche について聞かれます.



74水準名	c.menarche	
数值	水準名	
1	なし	
2	あり	2
0	ĸ	キャンセル

juul データにおける menarche の値は, 1: なし, 2: あり ですから, そのように水準名を入力し ましょう. 続けて c.sex です. こちらは 1: 男性, 2:女性です.

74水準	洺 c.sex			
数值	水	準名		
1	男	性		
2	女	性		
	OK	+	ャンセル	

最後に c.tanner です. これはローマ数字の I-V にしましょうか.

▓水準名。	c.tanner	<u> </u>
数值	水準名	
1	1	
2	II	
3	III	
4	IV	
5	M	
OK	く キャン・	t IV
4 5 OK	V M K = キャン・	t IL

Ok を押せば, 無事3つの因子変数が新しく作成されたはずです. 「データセットを表示」でみて みましょう.

2007/8/8 R Commander を用いた統計解析の基礎



1	% juul										×
		age	menarche	sex	igf1	tanner	testvol	c.menarche	c.sex	c.tanner	
·	614	59.00	NA	1	134	NA	NA	<na></na>	男性	<na></na>	
	615	60.00	NA	1	105	NA	NA	<na></na>	男性	$\langle NA \rangle$	
	616	63.00	NA	1	160	NA	NA	<NA $>$	男性	$\langle NA \rangle$	
I	617	65.00	NA	1	92	NA	NA	<na></na>	男性	<NA $>$	
1	618	67.00	NA	1	222	NA	NA	$\langle NA \rangle$	男性	<na></na>	
,	619	70.05	NA	1	NA	NA	NA	<na></na>	男性	<na></na>	
1	620	77.00	NA	1	126	NA	NA	$\langle NA \rangle$	男性	<NA $>$	
1	621	78.00	NA	1	90	NA	NA	$\langle NA \rangle$	男性	<na></na>	
J	622	79.00	NA	1	119	NA	NA	<NA $>$	男性	$\langle NA \rangle$	
1	623	80.00	NA	1	122	NA	NA	$\langle NA \rangle$	男性	<na></na>	
J	624	81.00	NA	1	112	NA	NA	$\langle NA \rangle$	男性	$\langle NA \rangle$	
1	625	81.00	NA	1	87	NA	NA	$\langle NA \rangle$	男性	$\langle NA \rangle$	
Ц	626	83.00	NA	1	149	NA	NA	<NA>	男性	<NA $>$	
1	627	83.00	NA	1	104	NA	NA	<NA>	男性	$\langle NA \rangle$	
Л	628	0.25	NA	2	51	1	NA	<NA>	女性	I	_
1	629	0.91	NA	2	25	NA	NA	<NA $>$	女性	$\langle NA \rangle$	
J	630	2.64	1	2	NA	1	NA	なし	女性	I	
1	631	3.25	NA	2	250	NA	NA	<na></na>	女性	$\langle NA \rangle$	
	632	5.12	1	2	NA	1	NA	なし	女性	I	
1	633	5.68	1	2	NA	1	NA	なし	女性	I	

それでは、因子によって層別した箱ひげ図を作成しましょう.「グラフ」メニューから「箱ひげ 図…」を選んでください.

74 R コマンダー		
ファイル 編集 データ 統計量	グラフ モデル 分布	Demos ツール ヘルプ
Rada F-Styr: juul F-	色パレット インデックスプロット	ットを表示 モデル: <mark><アクティブモデルなし></mark>
スクリプトウィンドウ	ヒストグラム	
tapply(red.cell.folat	幹葉表示 箱ひげ図	ntilation=red.cell.folate\$ventila
juul\$c.menarche <- fa juul\$c.sex <- factor juul\$c.tanner <- fact juul\$c.menarche <- fa juul\$c.sex <- factor juul\$c.sex <- factor juul\$c.tanner <- fact	QQプロット 散布図 散布図行列 折れ線グラフ 条件付き散布図 平均のプロット 棒グラフ 円グラフ	e, labels=c('なし','あり')) c('男性','女性')) abels=c('男性','女性')) e, labels=c('なし','あり')) c('男性','女性')) abels=c('I','II','III','IV','V')) nt=getRcmdr('logFont'), maxwidth=;
●	3次元グラフ グラフをファイルで保存	

「箱ひげ図」のウィンドウが出たら、変数には「age」を選び、「層別のプロット」ボタンを押しましょう.



<mark>7%</mark> 箱ひげ図	<u>_ ×</u>
変数(1つを選択)	
age	
igf1	
sex	
マウスで外れ値情報を表示 🗌	
層別のプロット	
OK キャンセル	ヘルプ

「質的変数」ウィンドウが開いたら,因子化した変数を1つ選ぶことができます.c.tannerを選んでみましょう.

<mark>7%</mark> 質的変数	
層別変数 (1つ選択)	
c.menarche	
<u>c.tanner</u>	
OK キャンセ	IV

「層別: c.tanner」となっているのがわかりますね. ここで "Ok" を押せばグラフがかかれます.

7%箱ひげ図	
変数(1つを選択)	
igf1	
マウスで外れ値情報を表示	
Varyat Citariner	
OK キャンセル	ヘルプ





c.tanner は成熟度の指標ですから、年齢が上がるにつれて進んでいることがわかります. でも、 Vのところはかなり年齢の高い人が含まれていますから、その影響で I-IV がわかりにくくなってい ますね. では、年齢が 20 歳以下の人のデータだけをみてみましょう.

まず,データセットからその部分集合を抽出します.「データ」メニューから「アクティブデー タセット」を選び,さらに「アクティブデータセットの部分集合を抽出」としましょう.



7% R コマンダー						_
ファイル 編集	データ 統計量	グラフ モデル	分布	Demos ツール ヘルプ		
$\mathbf{R}_{\mathbf{d}_{\mathbf{f}}}$ $\neq - 2 \mathbf{d}_{\mathbf{f}}$	新しいデータセッ データセットのロ・ データのインポー	ኑ - ሾ ኑ	•	・セットを表示 モデル: <mark><アクティブモデルなし></mark>		
data(juul	パッケージ内のデ		+	アクティブデーカセットの設け	1	
juul\$c.me juul\$c.se	アクティブデータt	2ット内の変数の管	理♪	アクライブデータセットを新しくする		
juul\$c.ta juul\$c.me	nner <- fact narche <- fa	tor(juul\$ta actor(juul\$	nner mens	アクティブデータセット内の変数		
juul\$c.se	x <- factor nner <- fact	(juul\$sex, cor(juul\$ta	labe nner	ケー人名の設定 アクティブデータセットの部分集合を抽出		
showData(juul, placer	ment='-20+2	oo', ″ 、	アクティブなデータセット内の変数を積み重ねて結合する… 欠測値のあるケースを削除…	þ,	maxł
boxpioc (a	gene.canner,	, yrab- age	<u> </u>	アクティブデータセットの保存		
•				799479 - SE9FØIDX# - F]	
出力ウィンドウ	1					実

「データセットから部分的に抽出」ウィンドウが開いたら、「すべての変数を含む」がチェックさ れていることを確認し、「部分集合の表現」として、20歳以下ということで "age <= 20" を入力. そして、抽出結果のあたらしいデータセットの名前として「myjuul」などとして "Ok" を押します.

7%データセットから部分的に抽出	
すべての変数を含む 🔽	-
または	5
変数(1つ以上選択)	
age	5
c.menarche	-
C.sex	ŧ
c.tanner	F
部分集合の表現	=
age <= 20	
T F	
 新しいデータセットの名前	Ļ
myjuul	
OK キャンセル へい	ע ד

それでは, myjuul で, さきほどと同様の箱ひげ図をかいてみましょう. 以下のようになりましたか?

2007/8/8 R Commander を用いた統計解析の基礎





では、次に連続(量的)変数同士のプロットをやってみましょう。最初にためしたときと同様、 「グラフ」メニューから、「散布図...」を選択してください。







<mark>7/</mark> 散布図	
x 変数(1 つ選択)	E数(1 つ選択) e ▲ unarche
点を確認する	
×の値にゆらぎを与えて表示	
⊻の値にゆらぎを与えて表示	
周辺箱ひげ図	
最小2乗直線	
平滑線	
スムージングの幅	50
部分集合の表現	
<全ての有効なケース>	
層別のフロット	
×軸ラベル	⊻ 軸 ラベル (合計)
プロットするパラメータ	
プロットするパラメータ	〈自動〉 」
点の大きさ	1.0
軸テキストの大きさ	1.0
軸ラベルのテキストの大きさ	1.0
0K キャンセ	ル ヘルフ

「散布図」ウィンドウが出たら、x 変数として "age", y 変数として "igf1" を選択して "Ok" とします。 R Commander を用いた統計解析の基礎



年齢とIGF-I の値の関係を示すプロットが出力されました。軸の周辺に、それぞれの変数の箱ひ げ図も表示されているので、x 軸, y 軸変数の分布も同時に把握することができます(「周辺箱ひげ 図」チェック)。それと、ノンパラメトリック回帰による平滑化曲線(「平滑線」チェック)、回帰直 線(「最小2乗直線」チェック)も同時にプロットされています。





プロットに対応する、age と igfl の相関係数も計算してみましょう。「統計量」メニューの「要約」、「相関の検定…」です。

74 R 🖃	マンダー							
ファイル	編集 デー	-タ 統計量	グラフ モデル	/ 分布	Demos	ツール	ヘルプ	
R . <i>∓</i>	ータセット:	■ 要約		ト アクテ・	(ブデータセ	Zwh		V: マクティブ
Cindr	-	└── 分割表		▶ 数値() ▶ 岐鹿/	こよる要約.	•••		
200	フトワイン	<u>トリ</u> 一 - 比率		「 - 死限(2) ▶ - 欠割(i)	511中… 直を對える。			
scat	terplot(terplot(ig: 此中 ig: 分散		▶ 層別 (D統計量…			TRUE labe
mvju	ul.femal	ig. .e . ノンパラ	メトリック検定	▶ 相関(INOL, TABE.
scat	terplot(ig: 次元解	新	相関(D検定…			TRUE, labe:
myju	ul\$logig	r£1モデル^	の適合	 シャピ 	ロ・ウィルク	の止規性	的検定…	
scat	terplot(logigf1~a	age, reg.l	line=lm	, smoo	th=TR	UE, lai	bels=FALSE, k
cor.	test(myj	uulage,	myjuulaig	jii, ai	ternat	ive="	two.si	ded", method≕
							1	
_								
出力ウ	イントワー							
76 1	目期の検索	2	N					
			NG					
変	数(2つ道							
aq	e							
iqf	1							
log	jigf1							
me	narche		•					
†B†	関のタイプ	7		対立仮	R			
e	アソンの新	aia室相関 🤇	2	両側				
7	מנקקש	D.随伤 (坦阳 ~	0.0			
- <u>۲</u>	ンドールの	りタウ ()	相関 >	0 0			
	ОK	±	センセル		~	њ 7	1	
	UN		1.5.670			~ ~		

変数として age と igfl を選んで、ピアソンの積率相関係数を選ぶと、出力ウィンドウに計算結 果が出ます。 2007/8/8 R Commander を用いた統計解析の基礎



標本相関係数 0.642, 母相関係数 =0 の帰無仮説は非常に小さい p 値で棄却されていますね。

しかし、この図をみると、年齢が7歳ぐらいまでは直線へのあてはまりがよさそうですが、それ 以降は IGF-I のばらつきがかなり大きくなっていますね。成長ホルモンに関連するものですから、 年齢とは関係がありそうな気がしますが…他にも関係する要因があるのかもしれません。そこで、 性別で層別してみましょう。

「散布図」ウィンドウにも、「層別のプロット」というボタンがありましたね。早速、c.sex で層 別をしてみましょう。

7% 質的変動	_ 🗆 🗡
屠別変数 (1つ選択)	
c.menarche	
c.sex	
c.tanner 🚽	
層別して線を描く 🔽	
OK キャンセル	



そうすると、プロットの点が男性と女性で色分けされて、下図のようなグラフが出ます。



...うーん、あんまりきれいにわかれた感じではありませんね。

どうも、igfl が年齢がすすむにつれて、直線的ではなく指数的に増えている感じを受けます。で は、igfl を「対数変換」して、もう一度プロットしてみましょう。weight と height から BMI をつ くったときを思い出して、igfl を対数変換した変数、logigfl を作成してください。「データ」メ ニューの「アクティブデータセット内の変数の管理」、その先の「新しい変数を計算...」です。



74新しい変数の計算		
現在の変数(ダブルクリ:	ックして計算式の	テキストボックスに入力)
age c.menarche [因子] c.sex [因子] c.tanner [因子]		
新しい変数名	計算式	
logigf1	log(igf1)	
	4	Þ
OK 🗕 🕈 🕈	ンセル	~11~

そして、age と logigfl でプロットしてみると...



前よりは直線によくのってきましたね。相関係数も計算してみましょう。



標本相関係数 0.736, その 95%信頼区間 0.704-0.765 ということで、igfl そのままよりはよく相関しています。

ただ、15歳以降でちょっと減少傾向に転じているような...

では、今度は初経の有無で層別してみましょう。c.menarche は女性にしかないので、まずは データセットからの部分抽出で、女性だけのデータセット、myjuul.female をつくります。やり方 はわかりますね。

7%データセットから部分的に抽出	<u>_ </u>
すべての変数を含む 🔽	
または	
変数(1つ以上選択)	
age	
c.menarche	
c.sex	
c.tanner 🗨	
部分集合の表現	
c.sex ==	
T E	
新しいデータセットの名前	
myjuul.female	
OK キャンセル	ヘルブ

そして、myjuul.female における age と igfl の散布図を、c.menarche で層別して描いてみま しょう。

2007/8/8 R Commander を用いた統計解析の基礎





初経の前後で増加傾向から、減少傾向に転じていることがわかりますね。

3. カテゴリカルデータの集計

myjuul データセットを用いて、カテゴリカルデータの処理についてもう少しみていきましょう。 「統計量」メニューから「分割表」とたどると、<u>変数の因子化をする前は</u>一部のメニューがかすれ た文字になっていて無効化されていたと思います。





7∦ R コマンダー	
ファイル 編集 データ	統計量 グラフ モデル 分布 Demos ツール ヘルプ
R _{dr} F-&tyt: 2007F072F0 data(juul, pac) juul\$c.menarche juul\$c.sex <- : juul\$c.tanner - mujuul <- subset	要約 学校 平均 比率 クンパラメトリック検定 大元解析 モデルへの適合 ・ たき、 たま、 たま、 モデル: マク 2元表、 多元表、 シース表の入力と分析、 ・ たま、 1abels=c('なし','あり 1abels=c('I','II','III','III') マーク マーク シース表、 シースま、 シースたる、 シース表、 シース表、 シース表、 シース表、 シースま、 シースま、 シースま、 シースま、 シースま、 シースた、 シースま、 シースま、 シースま、 シースま、 シースま、 シースたる シースたる シースたる、 シースたる シースたる シースたる シースたる シースたる シースたる シー
boxplot(age~c.t data(juul2, pac	anner, ylab="age", xlab="c.tanner", data=myju: :kage="ISwR")

それが、因子化をしたあとの現在のデータセットでは、「2元表…」などが選択できるように なっています。これは、因子化をすることによって、このデータセットのなかのいくつかの変数は カテゴリカルデータである、ということがRに伝わったためです。Rはデータが数量データなの か、カテゴリカルデータなのかをいつも見ています。それに応じて処理を変更してくれるのです。 それでは、「分割表」から「2元表…」を選んでみましょう。

7∕4 R コマンダー	
ファイル 編集 データ	統計量 グラフ モデル 分布 Demos ツール ヘルプ
Red データセット: n スクリプトウィンドウ scatterplot(we: data(juul, pac) juul\$c.menarche	要約 ア# C. 日本…上本本示 モデル: マア (分割表 2元表… V 平均 多元表… V 比率 2元表の入力と分析… ALSE, labels: 分散 ノンパラメトリック検定 menarche, labels=c(!なし, 」本
juul\$c.sex <- : juul\$c.tanner - myjuul <- subse boxplot(age~c.t	次元解析 モデルへの適合 abels=c('男性','女性')) nner, labels=c('I','II','III' et(juul, subset=age <= 20) tanner, ylab="age", xlab="c.tanner", data=myju"
▲	
出力ウィンドウ	

この「2元表」とは、いわゆる2×2の分割表です。行、列それぞれに指定したカテゴリカル変数の値に基づいて、それぞれの条件を満たすデータが「いくつあるのか」を集計して表にしてくれます。では、まず c.sex と c.menarche でみてみましょうか。

R Commander を用いた統計解析の基礎



762元表 _□×
行の変数(1つ選択) 列の変数(1つ選択) c.menarche c.sex c.tanner のの変数(1つ選択)
パーセントの計算 行のパーセント 〇 列のパーセント 〇 総計のパーセント 〇
パーセント表示無し 💿 🕟 仮説検定 独立性のカイ2乗検定 🔽
カイ2乗統計並の要素 期待度数の表示 「 フィッシャーの正確検定 「 部分集合の表現
<全ての有効なケース> OK キャンセル

上のように指定して「Ok」を押すと...

```
出力ウィンドウ
> .Table
    c.menarche
c.sex なし あり
男性 0 0
女性 368 263
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
    Pearson's Chi-squared test
data: .Table
X-squared = NaN, df = 1, p-value = NA
> remove(.Test)
```



出力ウィンドウの上のほうに、c.sex と c.menarche でクロス集計をした結果が表示されます。 c.sex は性別, c.menarche は初経の有無ですから、当然男性ではすべてゼロになりますね。何らか のデータ入力ミスがあれば、ここで気がつくことができます。では、今度は c.menarche と、c. tanner で集計してみましょう。

```
出力ウィンドウ
```

```
.Table <- xtabs(~c.menarche+c.tanner, data=myjuul)</pre>
 .Table
          c.tanner
c.menarche I
               II III
                        IV.
                             v
      なし 221 43 32
                        14
                             2
      あり
           1
              1
                   - 5
                       26 190
 .Test <- chisq.test(.Table, correct=FALSE)</pre>
  .Test
        Pearson's Chi-squared test
data:
       .Table
X-squared = 463.5167, df = 4, p-value < 2.2e-16
```

このように集計されます。初経の有無と、ターナーの成熟度分類の値には何らかの関係があるように見えますね。出力ウィンドウの下のほうには、この分割表(今回は成熟度分類が5段階なので、2x2ではないですが)に対してカイ2乗検定を行った結果が出ています。p値が非常に小さく、この2つのカテゴリ変数は独立しているという仮説が棄却できることがわかります。

なお、ここではやりませんが、分割表が 2x2 になったときには、「フィッシャーの正確検定」に チェックをいれるべきだったことも忘れずに。

4. 1標本または2標本の検定

ここまでで、カイ2乗検定と相関係数の検定はやりましたが、おそらくもっともよく知られてい る検定、t 検定についてはまだでしたね。figures6 に戻って、まずは1標本の検定からやってみま しょう。データセットを figures6 に指定しなおしたら、「統計量」メニューから「平均」,「1標本

2007/8/8

R Commander を用いた統計解析の基礎



t 検定」にいきます。

7% R コマンダー		
ファイル 編集 データ	統計量 グラフ モデル	、分布 Demos ツール ヘルプ
Rada データセット: fi	要約 分割表	▶ 50 編集 データセットを表示 モデル: <ア
<pre>X0U7F0xvF0 myjuul.female - scatterplot(ig: myjuul\$logigf1 scatterplot(log cor.test(myjuu) cor.test(myjuu) cor.test(myjuu)</pre>	平均 比率 分散 ノンパラメトリック検定 次元解析 モデルへの適合 L\$age, myjuul\$ig L\$age, myjuul\$ig	1標本 t 検定… 独立サンブルの t 検定… 対応のある t 検定… 1元配置分散分析… 多元配置分散分析… 1, alternative="two.sided", me ogigf1, alternative="two.sided", me
出力ウィンドウ		
cor		

0 641842

weight は平均 74.33, 最大値 95 でしたから、この母集団は体重がちょっと重そうです。そこで、 体重の母平均値が 90kg である、ということを帰無仮説、90kg とはいえないということを対立仮説 として、両側検定をしてみましょう。

<mark>74</mark> 1標本の t 検定	_ 🗆 🗙
変数(1つを選択)	
Seight weight	
対立仮説	
母平均 ⊭ ≠ ⊭ 0 € 仮説: ⊭ = 9	0
母平均 🛛 < 🖉 〇	95
母平均 〃 > 〃 0 〇	
OK キャンセル	ヘルブ

母平均値 μ に 90 を入力、<u>対立仮説</u>に母平均 $\mu \neq \mu_0$ を指定して、Ok を押してみます。



> t.test(figures6\$weight, alternative='two.sided', mu=90, conf.level=.95)
 One Sample t-test
data: figures6\$weight | I
t = -2.4882, df = 5, p-value = 0.05528
alternative hypothesis: true mean is not equal to 90
95 percent confidence interval:
 S8.14796 90.51870
sample estimates:
mean of x
74.33333

ということで、標本平均は 74.33 ですが、母平均値が 90kg である、という仮説は p=0.055 で ぎりぎりですが棄却することができませんでした。そのため、母平均値が 90kg ということも否定 できませんが、p 値がぎりぎりであることに対応して、母平均値の 95%信頼区間も 58.1-90.5 です から、たとえば母平均値が 91kg である、という仮説ですと棄却されることがわかります。そこま では重くない集団のようです。

それでは、2標本の場合、すなわち、平均値の<u>差の</u>検定に移りましょう。使うデータは ISwR パッケージ内の "energy" データセットです。「データ」→「パッケージ内のデータ」→「アタッチ されたパッケージからデータセットを読み込む...」で、アクティブデータセットに指定しましょ う。

74	energy	- D ×	eı
	expend	stature	4
	0.01		エ
1	9.21	opese	15
2	7.53	lean	
3	7.48	lean	や
4	8.08	lean	差
5	8.09	lean	
6	10.15	lean	
7	盘.40	lean	
: 8	10.88	lean	
9	6.13	lean	
10	7.90	lean	
11	11.51	obese	
12	12.79	obese	
13	7.05	lean	
14	11.85	obese	
15	9.97	obese	
16	7.48	lean	
17	8.79	obese	
18	9.69	obese	
19	9.68	obese	
20	7.58	lean	
21	9.19	obese	
22	8.11	lean	

energy データセットは、やせている女性と太っている女性で、2 4時間でのエネルギー消費量を測定したものです。変数 expend が エネルギー消費量(単位はメガジュール)、stature は体型で、すで に因子化されています。

やせている場合と太っている場合で、エネルギー消費量の平均値に 差があるかどうかが興味深いところです。

2007/8/8

R Commander を用いた統計解析の基礎



2標本のt検定は、「統計量」→「平均」→「独立サンプルのt検定…」です。

74 R コマンダー			
ファイル 編集 データ	統計量 グラフ モデル	一分布 Demos ツール ヘ	いげ
Rade データセット: e	要約 ▶ 分割表 ▶	編集 データセットを表示	: モデル:[
t.test (figures) t.test (figures) t.test (figures) t.test (figures) t.test (figures)	1 1 比率) 分散) ノンパラメトリック検定) 次元解析) モデルへの適合)	1編本 t 復定 独立サンプルの t 検定 対応のある t 検定 1元配置分散分析 多元配置分散分析 tive='two.sided',	560, conf Mu=70, co mu=90, co mu=90, co mu=91, co
data(energy, pa showData(energy	ackage="ISwR") 7, placement='-20)+200', font=getRcm	ndr ('logFo

「独立サンプルのt検定」ウィンドウがでますから、差をとりたい「グループ」をあらわす因子 として,体型データの stature,実際のエネルギー消費量をあらわす連続変数として、expend を指 定します。グループ間で、測定値の母分散が等しいと考えられる場合は「等分散と考えますか?」 に「Yes」をつけますが、ここでは必ずしもそうとは考えられないかなと思うので、「No」にして おきましょう。「Yes」の場合は Student のt検定、「No」の場合は Welch のt検定となります。

<mark>74 独立サンブルの t 検定</mark>	
グループ(1つ選択)	目的変数(1つ選択)
<u> </u>	A
stature	expend
<u></u>	<u> </u>
差: <層別因子が選択されて!	いません>
対立仮説 信頼水準	等分散と考えますか?
両側 💿 .95	Yes O
差 < 0 🔘	No 🤉
差 > 0 〇	4
OK キャンセ	:ル ヘルフ

検定結果は出力ウィンドウです。

```
Welch Two Sample t-test

data: expend by stature

t = -3.8555, df = 15.919, p-value = 0.001411

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.459167 -1.004081

sample estimates:

mean in group lean mean in group obese

8.066154 10.297778
```



帰無仮説、「2群でエネルギー消費量の平均値には差はない」は、p値 0.001 ということで棄却 されました。すなわち、やせている場合と太っている場合でエネルギー消費量には差が認められた ことになります。この結果には、各群のエネルギー消費量の標本平均値もでていますね。それぞれ 8.07 と 10.3 です。

エネルギー消費量はおそらく正規分布すると思われますから、t 検定で問題ありませんが、母集 団の年齢構成にそもそも偏りがあることなどがわかっており、エネルギー消費量が正規分布しない ことが明らかな場合などには、t 検定を使うことは適切ではありません。このような場合、ノンパ ラメトリックな手法を用います。2標本の平均値の差の検定におけるノンパラメトリック法は、 Wilcoxon の順位和検定です。「統計量」→「ノンパラメトリック検定」→「2標本ウィルコクソン 検定…」にあります。

7∦R コマンダー		
ファイル 編集 データ	統計量 グラフ モデル	分布 Demos ツール ヘルプ
Radio データセット: Ce	要約●分割表●平均●)編集 データセットを表示 モデル: <mark><ア</mark> く
t.test(figures) t.test(figures) t.test(figures) t.test(figures) data(energy, page)	比率) 分散) ノンパラメトリック検定) 次元解析) モデルへの適合)	tive='two.sided', mu=70, conf. tive='two.sided' mu=90, conf. 2標本ウィルコクソン検定 対応のあるウィルコクソン検定い 1, conf.
showData(energy t.test(expend~:	7, placement='-2 stature, alterna	0+200', font=getRcmdr('logFont' tive='two.sided', conf.level=.9

変数の指定方法は、t検定の場合とほとんど同じですね。

7%2標本ウィルコクソン検定		_ 🗆 🗙
グループ(1つ選択)	目的変数(1つ選択)
stature	expend	
≝∶lean - obese		_
対立仮説	検定のタイプ	
両側 💿	デフォルト	•
± <0 ○ 🔓	正確	0
差 > 0 〇	正規近似	0
	連続修正を用いた 正規近似	C
OK キャンセ	(J)	ルフ

2007/8/8 R Commander を用いた統計解析の基礎



結果はp値が0.002と、t検定のときに比べて若干大きくなっていますが、2群が同じ母集団から抽出されたという帰無仮説は棄却され、やはりエネルギー消費量には2群で差があるという結論となります。ただ、t検定のほうが正規分布を仮定しているため、データから利用できる情報量が多くなり、差の検出力があがります。したがって、p値はt検定を用いた場合のほうが小さくなるのです。

```
> wilcox.test(expend ~ stature, alternative="two.sided", data=energy)
```

Wilcoxon rank sum test with continuity correction

```
data: expend by stature
W = 12, p-value = 0.002122
alternative hypothesis: true location shift is not equal to 0
```

ではでは、次に「対応がある」データの場合に進みましょう。データは同じく ISwR パッケージ にある "intake" です。

🎀 inta 💶 🗖 🗙		
	pre	post
1	5260	3910
2	5470	4220
3	5640	3885
4	6180	516b
5	6390	5645
6	6515	4680
7	6805	5265
8	7515	5975
9	7515	6790
10	8230	6900
11	8770	7335

intake データは小さいものですが、11名の女性の月経前10日間と 月経後10日間で、エネルギー換算した食事摂取量を比較したもので す。月経前10日間の食事摂取量が変数 pre,月経後10日間が post に格納されています。1人の女性について2回の測定が行われてい るので、pre 群と post 群は「対応がある」2群ということになりま す。対応のある2群の平均値の差の検定には、やはりt 検定という名 前がついています。「統計量」→「平均」→「対応のあるt 検定...」 で呼び出してみましょう。





<mark>7%</mark> 対応のある t 検定	
第1の変数(1つ選択)	第2の変数(1つ選択)
post	post A
	pre
対立仮説	信頼水準
両側 🖲	.95
差 < 0 〇	
差 > 0 〇	
OK キャン1	セルーニーへルフ
	N

第1の変数、第2の変数として、それぞれの群のデータを示す変数を指定します。Ok ボタンを 押せば、結果が出力ウィンドウに表示されます。

Paired t-test

```
data: intake$pre and intake$post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1074.072 1566.838
sample estimates:
mean of the differences
1320.455
```

結果は通常のt検定と似ていますが、"Paired" t-test となっていることに注意しましょう。

p値は非常に小さく、月経前と月経後で摂取エネルギー量に差があったことが示されました。 一方、この検定もノンパラメトリック手法で行うこともできます。「統計量」→「ノンパラメト リック検定」→「対応のあるウィルコクソン検定…」です。

> wilcox.test tintake\$pre, intake\$post, alternative='two.sided', paired=TRUE)
 Wilcoxon signed rank test with continuity correction
data: intake\$pre and intake\$post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0

結果はやはり大きくは変化しませんが、p値は 0.00384 と、t検定に比べて大きくなっています。

5. 回帰分析

続いて、データを数理モデルにあてはめることで、変数間の直線的な関係の有無を明らかにする 解析手法、回帰分析に入っていきましょう。散布図のところで若干、これに近いことをしました が、回帰分析とは、



$$y_i = \alpha + \beta x_i + \epsilon_i$$

というように、x 変数で y を説明することが適切かどうかを検討することです。 それでは、R コマンダーで実際に回帰分析を行ってみます。今度のデータはやはり ISwR パッ ケージ、thuesen データセットです。

	7%	thuesen	
		blood.glucose	short.velocity
	1	15.3	1.76
	2	10.8	1.34
	3	8.1	1.27
	4	19.5	1.47
	5	7.2	1.27
	6	5.3	1.49
	7	9.3	1.31
	8	11.1	1.09
	9	7.5	1.18
	10	12.2	1.22
	11	6.7	1.25
	12	5.2	1.19
	13	19.0	1.95
	14	15.1	1.28
1	15	6.7	1.52
	16	8.6	NA
1	17	4.2	1.12
1	18	10.3	1.37
	19	12.5	1.19
	20	16.1	1.05
	21	13.3	1.32
	22	4.9	1.03
	23	8.8	1.12
	24	9.5	1.70

データセットに含まれるのは blood.glucose と short. velocity という変数。これは、I 型糖尿病の患者 24 人 における、空腹時血糖値と、心エコーの結果から求め られた心室内周短縮速度の平均値を示しています。心 室内周短縮速度は、心臓の収縮機能を示す指標の1つ で、この変動が血糖値の変化で説明できるかが興味あ るところです。このように「ある変数でほかの変数を 説明できるかを検討する」際には、変数間の関係を数 式で示した「数理モデル」を構築するのですが、R コ マンダーではここでいよいよ、今まで赤いままだった 「アクティブモデル」の出番となります。

thuesen をデータセットに指定後、「統計量」→「モデルへの適合」→「線形回帰…」を選びま しょう。

7% 線形回帰	
モデル名を入力: Model1	
目的変数(1 つ選択)	説明変数(1つ以上選択)
blood.glucose short.velocity	blood.glucose short.velocity
部分集合の表現	
<全ての有効なケース>	
₹ Þ	2
OK キャンセ	·ル ヘルフ

数理モデルは、変数と同様に名前がつけられ、作業スペースに保存されます。「線形回帰」ウィ ンドウが出たら、モデル名を「Model1」として名付け、目的変数(従属変数)に short.velocity, 説明変数(独立変数)に blood.glucose を指定して Ok を押してください。





すると、出力ウィンドウに構築されたモデルの詳細が表示されるとともに、コマンダーの「モデ ル」欄にも、いま作成した「Model1」が入ります。

```
> summary(Model1)
Call:
lm(formula = short.velocity ~ blood.glucose, data = thuesen)
Residuals:
    Min
              1Q Median
                                3Q
                                        Max
-0.40141 -0.14760 -0.02202 0.03001 0.43490
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)
             1.09781
                        0.11748 9.345 6.26e-09 ***
blood.glucose 0.02196
                        0.01045 2.101 0.0479 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2167 on 21 degrees of freedom
 (1 observation deleted due to missingness)
Multiple R-Squared: 0.1737,
                              Adjusted R-squared: 0.1343
F-statistic: 4.414 on 1 and 21 DF, p-value: 0.0479
```

Modell は、このような数式であらわされます。

$short.velocity = 1.10 + 0.0220 \times blood.glucose$

出力結果の「Coefficients」のところで、回帰係数の推定値(Estimate)をみると、1.10 と 0.020 がでていますね。(Intercept)が α の値、すなわち切片で、blood.glucose のところが β となります。

 α 、 β には、Estimate のほかにも、Standard Error、つまり推定値のばらつきも表示されます。 推定値があまり大きくばらついているようなら、あまり信用のできない推定値ということになるで しょう。そして、t value と Pr(>|t|)は、それぞれの回帰係数が0である、という帰無仮説をt検 定で検定した結果のt統計量とp値を示しています。Pr(>|t|)のところが0.05未満であれば、帰無 仮説は棄却され、回帰係数がゼロ(すなわち、その変数はまったく説明に寄与していない)とはい えないことになります。なお、切片のところでも一応p値がでていますが、切片がゼロかどうかは あまりモデルの解釈に影響しないので、それほど注目はされません。

このモデルがデータにどの程度よくあてはまっているのか?ということをあらわすのが、 Residuals, すなわち「残差」です。Min, 1Q, Median, 3Q, Max とありますが、これはデータの要約のところでもでてきた、最小、第1四分位点、中央値、第3四分位点、最大値です。回帰モデルは、この残渣の平均値がゼロになるように計算されているので、平均値は必ずゼロですからここには出てきません。モデルがよくあてはまるならば、残差はなるべく小さくなるべきであり、そして 2007/8/8

R Commander を用いた統計解析の基礎



正規分布するべきですが、そうなっているかをここで簡単ではありますが確認することができま す。

Residual Standard error のところは、残差のばらつきですから、実際の観測値が回帰直線のま わりにどの程度ばらついているか、ということを示しています。これはあてはめた測定値(short. velocity)の単位に依存する値なので、一般的なモデル適合度の指標値ではありませんが、測定値に 比べてあまり大きいとあてはまりはよくないことがわかります。

Multiple R-squared は、説明変数と目的変数の間のピアソンの相関係数を2乗したものですが、 これは回帰によって説明される偏差平方和が、総平方和の何%であるか、という値と同じことにな ります。分散分析のところでもう少し詳しく説明しますが、実は分散分析の考え方と回帰分析の考 え方は統一的にとらえることができるのです。ここでは、この Multiple R-squared(R²)は全ての データの変動のうちモデルによって説明できる割合を示していると理解してください。