

医学統計セミナー2019

記述統計・グラフ表示・統計ソフトウェア

臨床研究センター
谷岡健資

2019年度 医学統計セミナー

開催日	講義内容
5月21日 (火)	記述統計・グラフ表示・統計ソフトウェア
6月18日 (火)	単群・2群比較のための諸種の検定
7月25日 (木)	分散分析と多重比較
8月22日 (木)	相関分析と回帰分析
9月26日 (木)	クロス集計表の解釈と諸種の検定
10月17日 (木)	ロジスティック回帰分析
11月14日 (木)	生存曲線の推定と比較
12月 3日 (火)	Cox比例ハザードモデル
1月23日 (木)	感度・特異度・ROC曲線とその比較
2月27日 (木)	繰り返し測定分散分析、混合効果モデル

目次

□ はじめに

- 解析全体における記述統計・グラフ表示の位置づけ
- 記述統計の概観
- データの尺度と種類
- グラフ表示の概観

□ データ尺度別の集計方法(記述統計とグラフ表示)

- 量的データの一変量に関する要約
- 量的データの二変量間の関係に関する要約
- 質的データの一変量に関する要約
- 質的データの二変量間の関係に関する要約

□ 統計ソフトウェアについて

1.はじめに

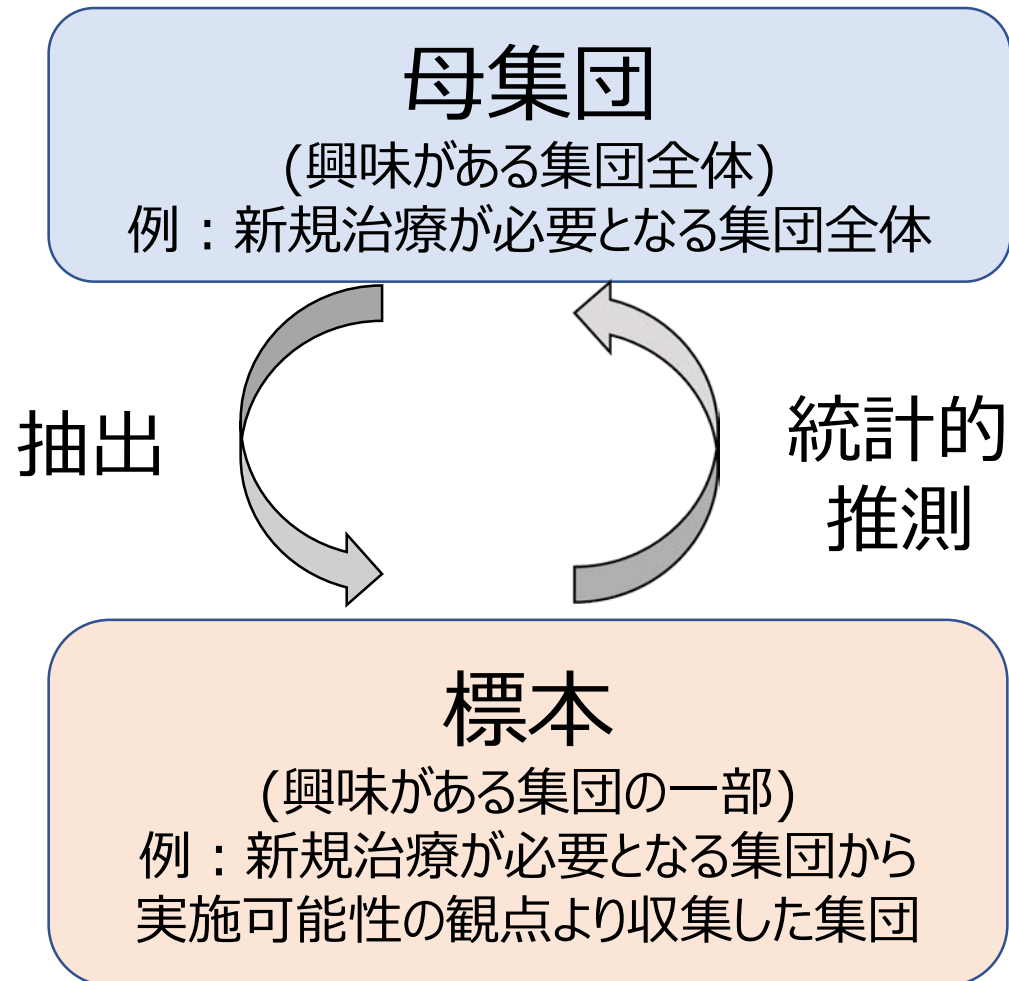
□解析全体における記述統計・グラフ表示の位置づけ

➤問題設定

統計的推測の結果を持って
結果を主張することが多いが…

記述統計・グラフ表示の意味

1. 標本の特徴を把握することは推測を行う上でも重要
2. 統計的推測の結果をグラフ等で補強することは重要
3. 統計的推測が適切でない場合も存在



1.はじめに

□解析全体における記述統計・グラフ表示の位置づけ

➤記述統計とは

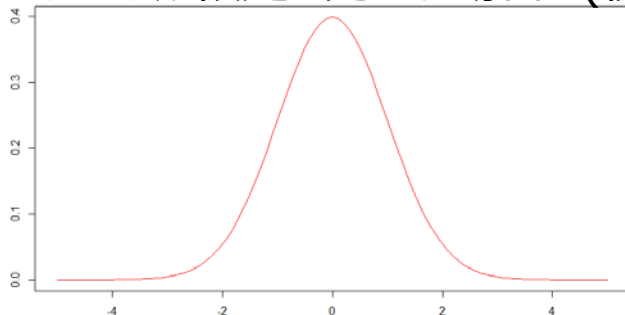
標本に対応するデータの集計・要約

➤1.標本の特徴を把握することは推測を行う上でも重要

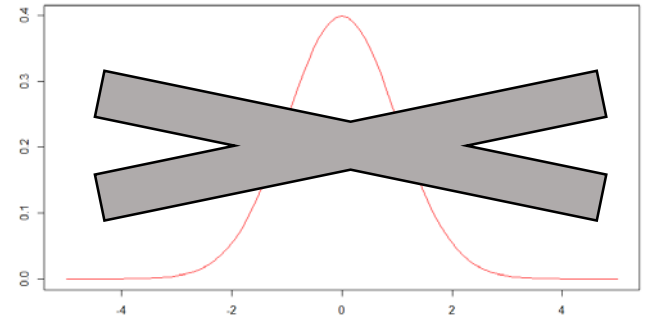
標本の特徴（例：平均）から母集団の特徴（例：平均）を推測する上で「**推測する方法**」を決定するために必要となる

推測する方法は標本データに仮定をおいており、その仮定をみたすかの確認が重要

例： データが仮定を充たす場合（横軸：データ）



データが仮定を充たさない場合



利点：仮定を充たす場合，必要標本数が少なくて済む（記述統計の必要性）

1.はじめに

□解析全体における記述統計・グラフ表示の位置づけ

➤**2.統計的推測の結果をグラフ等で補強することは重要**

統計的推測により試験治療が既存治療と比較して優越性を示せた場合

例：グラフにより視覚的な観点からも試験治療群と既存治療群で差があることを提示できる場合がある

統計的推測により試験治療が既存治療と比較して優越性を示せなかった場合

例：試験治療群と既存治療群で本来差があると考えられるが…

- ・「臨床的に意味がある差」を示すのにサンプルサイズが足りない
 - ・観測方法によってデータの「バラツキ」が大きくなってしまった
- 等の理由で当該有効性を示せなかった場合

グラフにより、今後改善することで当該有効性の可能性を提示

1.はじめに

□解析全体における記述統計・グラフ表示の位置づけ

➤**3.統計的推測が適切でない場合**

希少疾患や第I相試験でサンプルサイズが小さく，統計的推測に関する方法の使用が適切でない場合

対策

記述統計やグラフ等を用いて，安全性やその有効性を示すような集計結果を多数提示する.

1.はじめに

□解析全体における記述統計・グラフ表示の位置づけ

➤論文では…

必ずといっていいほど、最初に基礎集計に関するTableを作成して、被験者の背景情報を集計、提示する必要がある。

2群比較（試験治療 vs 既存治療）の場合、比較可能性が担保されていることを提示するためにも必要となる。

	Control group (n=33)	SABR group (n=66)
Age	69 (64-75)	67 (59-74)
Sex		
Men	19 (58%)	40 (61%)
Women	14 (42%)	26 (39%)
Site of original primary tumour		
Breast	5 (15%)	13 (20%)
Colorectal	9 (27%)	9 (14%)
Lung	6 (18%)	12 (18%)
Prostate	2 (6 %)	14 (21%)
Other	11 (33%)	18 (27%)
Time from diagnosis of primary tumour to randomisation (years)	2-3 (1-3-4-5)	2-4 (1-6-5-3)
Number of metastases		
1	12 (36 %)	30 (46%)
2	13 (40%)	19 (29%)
3	6 (18%)	12 (18%)
4	2 (6%)	2 (3%)
5	0 (0%)	3 (5%)
Location of metastases		
Adrenal	2/64 (3%)	7/127 (6%)
Bone	20/64 (31%)	45/127 (35%)
Liver	3/64 (5%)	16/127 (13%)
Lung	34/64 (53%)	55/127 (43%)
Other*	5/64 (8%)	4/127 (3%)

Data are n (%), n/N (%), or median (IQR). SABR=stereotactic ablative radiotherapy. *Other comprises brain (n=3 lesions in control group; n=1 lesion in SABR group), lymph nodes (n=1 lesion in control group; n=3 lesions in SABR group), and para-renal (n=1 lesion in control group).

Table 1: Baseline characteristics

1.はじめに

□記述統計の概念

➤被験者のデータが与えられた際に一つの指標で特徴を解釈する

入力

頭部外傷症データ(Zhou, et al.,2002)
重篤度をCK-BB(クレアチン・キナーゼBB)識別することが目的

重症群					非重症群		
140	740	543	490	523	136	60	46
1087	126	913	156	76	286	17	
230	153	230	356	303	281	27	
183	283	463	350	353	23	126	
1256	90	60	323	206	200	100	
700	303	509	1560		146	253	
16	193	576	120		220	70	
800	76	671	216		96	40	
253	1370	80	443		100	6	

出力：データの特徴を表す指標(主に下記の3種類の指標)

- (1)「データの中心」：データの代表値
- (2)「データのバラツキ」：データがどれだけばらつきを表した値
- (3)「データ間の関係」：2変量間の関係を表現した値

1.はじめに

□データの尺度（東京大学教養学部統計学教室, 1994）

質的データ

名義尺度

観測値が単に対象のカテゴリを示しているデータ

順序尺度

観測値が序数としての意味をもち、対象間の順序付けを示しているデータ

間隔尺度

観測値が任意の単位の何倍という形で示され、測定値の差が意味をもつデータ

比例尺度

観測値が任意の単位の何倍という形で示され、さらに原点0が絶対的な意味をもっていて、差や比率も意味を持つデータ

量的データ

1.はじめに

□データの尺度（東京大学教養学部統計学教室, 1994）

質的データ

名義尺度の例

例：名前，性別

順序尺度の例

例：癌のstage，ランキング

量的データ

間隔尺度の例

例：温度，西暦

比例尺度の例

例：価格，身長

1.はじめに

□データの尺度

- データの尺度によって、データの集計方法が異なる
- 実際に統計解析ソフトを用いて集計や統計解析を実施する際、データの尺度を統計解析ソフトにおいて指定する必要がある
 - * 質的データの集計等がしたいのに、集計するためのボタンがでてこないときは統計ソフトで量的データ（例：1, 0の2値）として認識されていることが多い
- 情報量という観点から下記の関係が成り立つ.

比例尺度 > 間隔尺度 > 順序尺度 > 名義尺度

- * 情報量が多いほど検定等でサンプルサイズを少なくなる傾向にあることからデータを設計する際に留意

1.はじめに

□データの種類（応用に近い分け方）

➤量的データ

血圧, 腫瘍径といった数値に単位があるようなデータ

➤カウントデータ

ポリープの個数などの「個数」を意味するデータ

➤2値データ

結果が(0,1) (成功,失敗)のように2値で表現されるデータ

➤質的データ

順序データ：疾患のグレード(軽度・中程度・高度)

名義データ：血液型や疾患の種類

➤生存時間データ

「生存期間」と「中途打ち切り」の対で表現されるデータ

1.はじめに

□データの種類（応用に近い分け方）

ホジキン病データ(Pintilie, 2006)

年齢	性別	治療	縦隔腫瘍	結節外疾患	ステージ	生存期間	ステータス
64	F	RT	N	Y	1	3.1	2
63	M	RT	S	N	1	15.9	2
17	M	RT	N	N	2	0.9	1
63	M	CMT	N	N	2	13.1	2
21	M	RT	L	N	2	35.9	0
37	M	RT	N	N	1	1.1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

F : 女性
M : 男性

RT : 放射線治療
CMR : 放射線 + 抗癌剤

L : 大きい
S : 小さい
N : なし

Y : あり
N : なし

0 : 中途打ち切り
1 : 再発
2 : 死亡

1.はじめに

□データの種類（応用に近い分け方）

年齢	性別	治療	縦隔腫瘍	結節外疾患	ステージ	生存期間	ステータス
64	F	RT	N	Y	1	3.1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
量的	2値	2値	順序 カテゴリカル	2値	順序 カテゴリカル	← 生存時間 →	

ここに出てくる統計用語

- ・ **標本サイズ**：データ(例えば患者)の数を表す.
- ・ **変数の数(変量数)**：試験などで得られた項目数を表す. 変数と変量は, 厳密には異なるが, ほぼ同じ.

尺度により, 要約およびグラフ表示の方法が異なる

1.はじめに

□ グラフ表示の概観

記述統計ではデータを目的に応じて一つの指標によって把握する

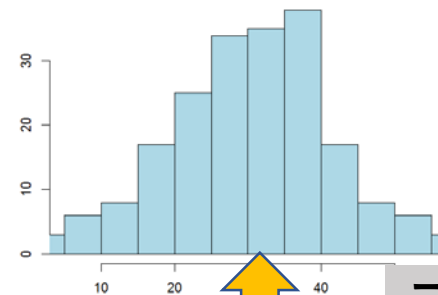
データを1つの指標によって表現することには限界があることから、グラフで視覚化することには意味がある。

困る例：例え記述統計の指標が一緒だったとしても、「データの特徴全て」が一緒ではないため、グラフが重要

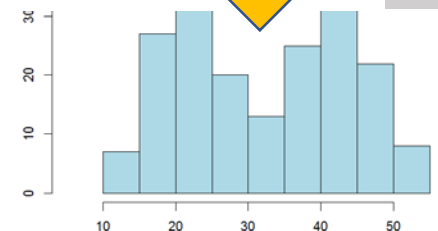
グループA
データの中心: 31.5

ほぼ同じ

グループB
データの中心: 32.8



データの特徴が違う



*グラフの説明は後述

2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約（量的データの中心値）

	データ値(検査値)
被験者1	23.4
被験者2	15.7
被験者3	25.2
⋮	⋮
被験者n	19.4

平均値



$$\text{データ値の平均値} = \frac{\text{データ値の合計}}{\text{被験者の総数}}$$

最も用いられるデータの中心値の指標
解釈が容易

注意点

- ・「データ値の合計」を割る数の定義について確認
- ・「外れ値」の影響を受ける

上の表に外れ値を加えて、平均をとると・・・ ⇒ 非常に影響をうける

	データ値(検査値)
被験者n+1	128.2

平均値 : without 外れ値
20.93

平均値 : with 外れ値
42.38

2. データ尺度別の集計方法(記述統計とグラフ表示)

□ 量的データの一変量に関する要約 (量的データの中心値)

Step 1: データ値の大きさに基づいて並びかえる

$$15.7 < 19.4 < \dots < 23.4 < \dots < 25.2$$

Step 2: Step1で並び替えたデータ値の中央値は下記のように定義される.

$$\left\{ \begin{array}{ll} \frac{\frac{n}{2} \text{ 番目のデータ値} + \frac{n}{2} + 1 \text{ 番目のデータ値}}{2} & \text{(被験者数が偶数)} \\ \frac{n}{2} + 1 \text{ の整数 番目のデータ値} & \text{(被験者数が奇数)} \end{array} \right.$$

例)

外れ値なし	被験者1	被験者2	被験者3	被験者4
	23.4	15.7	25.2	19.4

平均値 (without 外れ値) 20.93 中央値 (without 外れ値) 23.4

外れ値あり	被験者1	被験者2	被験者3	被験者4	被験者5
	23.4	15.7	25.2	19.4	128.2

平均値 (with 外れ値) 42.38 中央値 (with 外れ値) 23.4

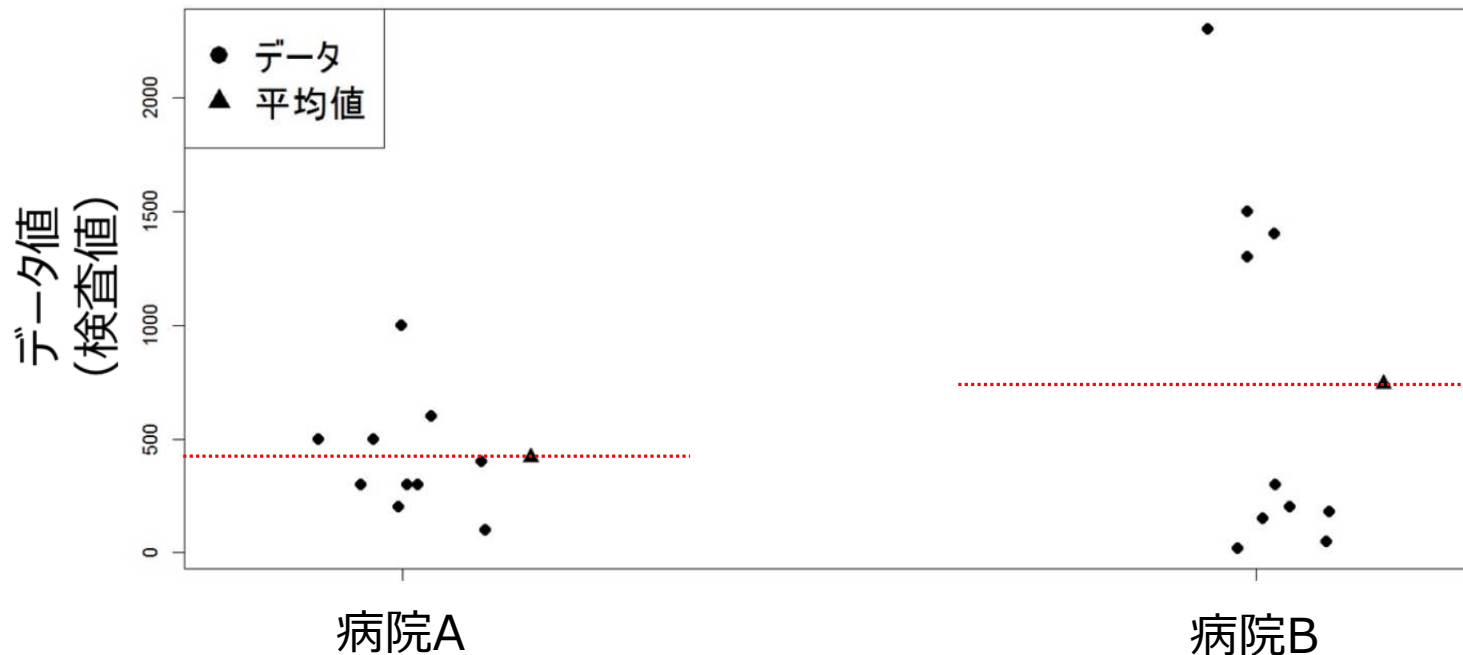
2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (分散の必要性)

バラツキを測ることの必要性

「平均値が2群で同じ」 ≠ 「2群のデータの特徴が同じ」

図：ドットチャート



平均値は2病院であまり差がないが、

病院A：検査値は病院Bの検査値と比較して平均値に近い値である傾向

病院B：検査値は病院Aと比較して平均値より高い値や低い値である傾向

⇒ 病院Bの方が検査値の値がばらついている

2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (分散の必要性)

バラツキを測る指標

分散

$$\frac{(\text{被験者1のデータ値} - \text{平均値})^2 + \dots + (\text{被験者nのデータ値} - \text{平均値})^2}{\text{被験者数}}$$

意味

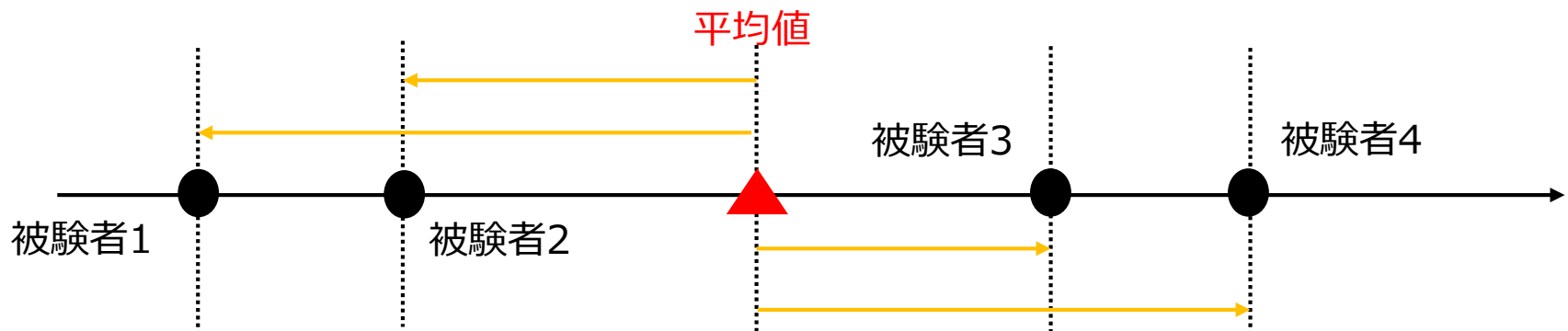
平均値を原点としてデータがどれくらいばらついているかを表現した指標

ばらついているとき : 分散の値が大きくなる

データ値が平均値より離れている傾向にある

ばらついていないとき : 分散の値が小さくなる

データ値が平均値と似ている値である傾向にある



2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (分散の必要性)

バラツキを測る指標

分散

$$\frac{(\text{被験者1のデータ値} - \text{平均値})^2 + \cdots + (\text{被験者nのデータ値} - \text{平均値})^2}{\text{被験者数}}$$

問題点

分散はバラツキを測る解釈容易な指標であるが・・・
2乗していることから元のデータの単位と異なる

標準偏差

$$\left[\frac{(\text{被験者1のデータ値} - \text{平均値})^2 + \cdots + (\text{被験者nのデータ値} - \text{平均値})^2}{\text{被験者数}} \right]^{\frac{1}{2}}$$

分散

標準偏差では元のデータの単位であることからバラツキを解釈しやすい

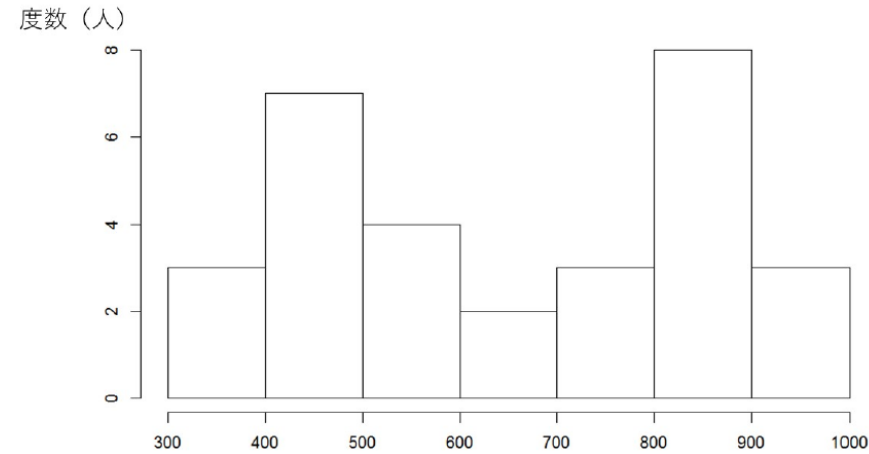
2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (ヒストグラム)

・Step1 : データ表

ID	data	ID	data	ID	data	ID	data
1	421	9	447	17	453	25	853
2	464	10	373	18	491	26	799
3	562	11	551	19	352	27	677
4	398	12	540	20	635	28	755
5	564	13	479	21	445	29	881
6	901	14	808	22	821	30	902
7	828	15	990	23	898		
8	815	16	885	24	734		

・Step3:ヒストグラム



横軸：データ値

データ値の分布を解釈することができる
例：2グループのデータが混じってそう

・Step2 : 度数分布表

階級	階級値 y_j	度数 f_j	累積度数 F_j	比率 p_j	累積比率 P_j
[300, 400)	350	3	3	0.10	0.10
[400, 500)	450	7	10	0.23	0.33
[500, 600)	550	4	14	0.13	0.47
[600, 700)	650	2	16	0.07	0.53
[700, 800)	750	3	19	0.10	0.63
[800, 900)	850	8	27	0.27	0.90
[900, 1000)	950	3	30	0.10	1.00

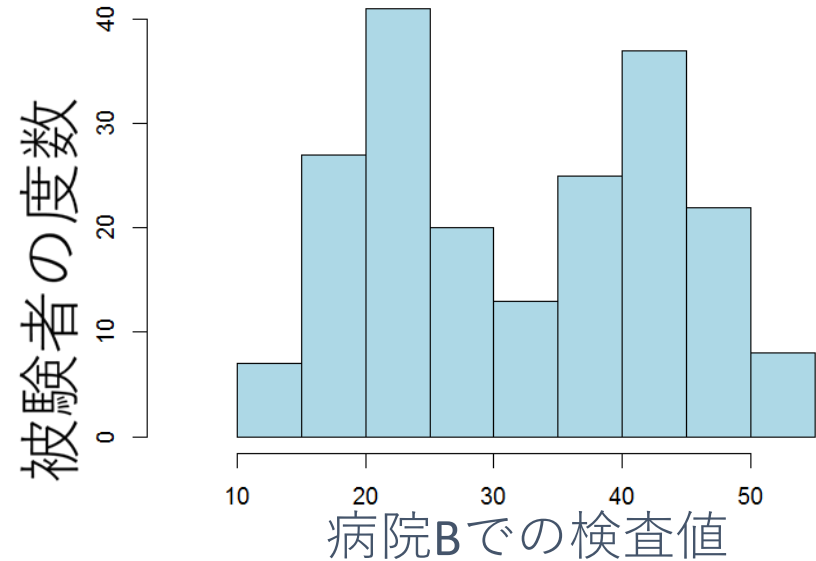
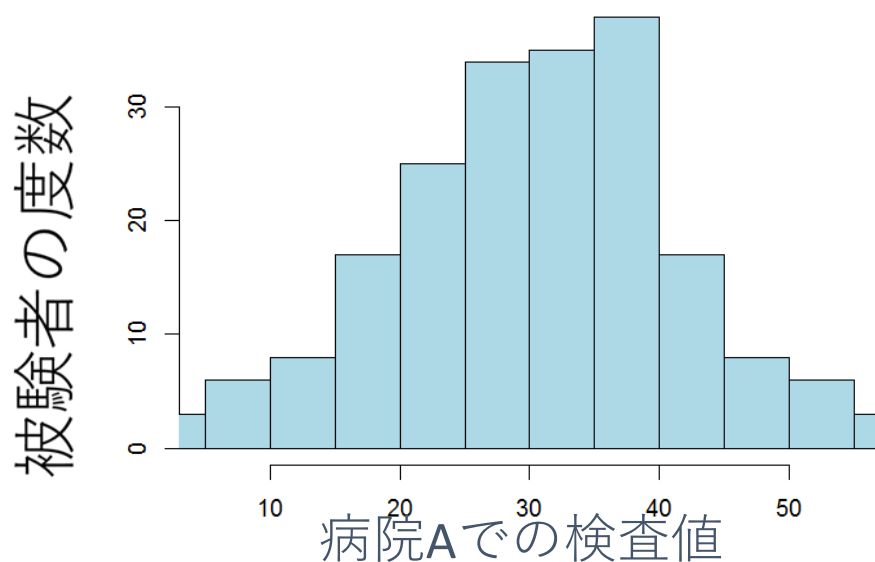
2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (ヒストグラム)

何故ヒストグラムが必要か？

病院Aでの検査値
平均値 : 31.56
標準偏差 : 11.01

病院Bでの検査値
平均値 : 32.80
標準偏差 : 11.43



平均値および標準偏差はほとんど一緒であるが、データの分布が異なる

統計的推測を使う場合は左図の場合はパラメトリック法（2回目以降説明）が適切で、
右図の場合はノンパラメトリック法（2回目以降説明）が適切

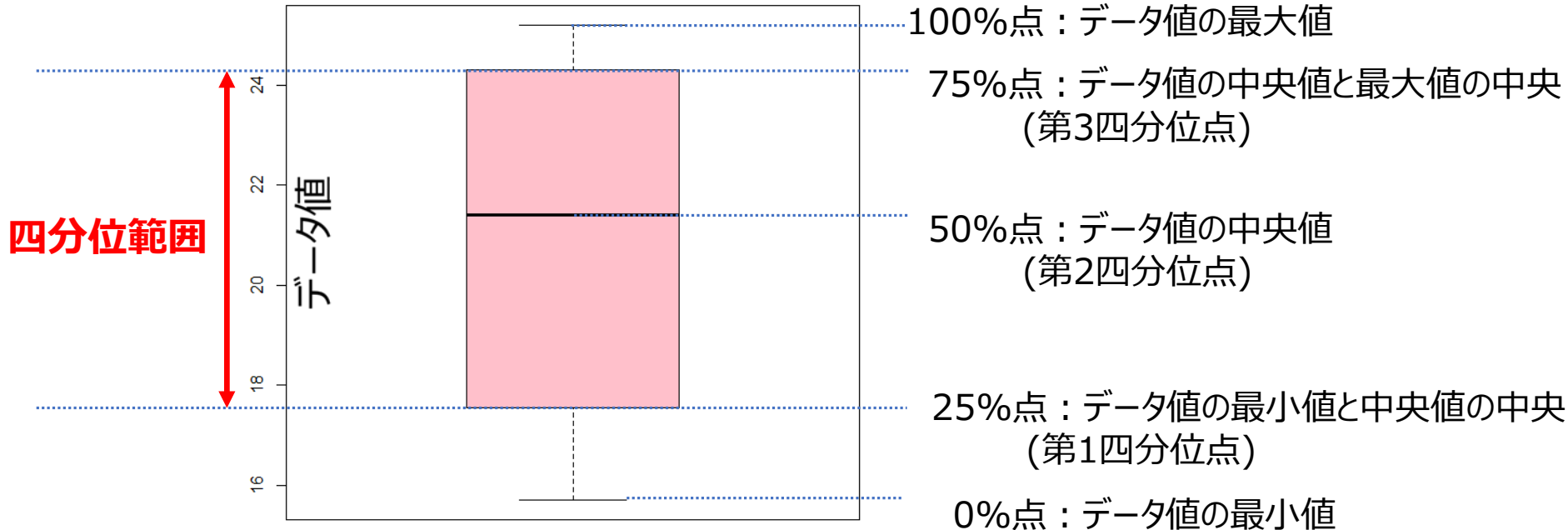
2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの一変量に関する要約 (箱ひげ図)

箱ひげ図(Box-plot)

データ値の下記を視覚化したグラフ

0%点, 25%点, 中央値, 75%点, 100%点



四分位範囲 : 「データ値の75%点と25%点の値の差」でデータのバラツキを表現
大きいとばらついている vs 小さいとばらついていない

箱ひげ図は「データの中心」と「データのバラツキ」を同時に表現したグラフ

2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの二変量に関する要約

共分散:データの線型関係が仮定されるときその強さを測る指標

$$\frac{1}{\text{被験者数}} \times$$



$$\left\{ (\text{被験者1のデータX} - \text{データXの平均値}) (\text{被験者1のデータY} - \text{データYの平均値}) + \dots + (\text{被験者nのデータX} - \text{データXの平均値}) (\text{被験者nのデータY} - \text{データYの平均値}) \right\}$$

2変量のデータ間の関係に線形関係が仮定できるとき,

共分散の値が大きい: データXが大きければ大きいほどデータYも大きくなる傾向
データYが大きければ大きいほどデータXに大きくなる傾向

共分散の値が小さい: データXが大きければ大きいほどデータYも小さくなる傾向
データYが大きければ大きいほどデータXに小さくなる傾向

2.データ尺度別の集計方法(記述統計とグラフ表示)

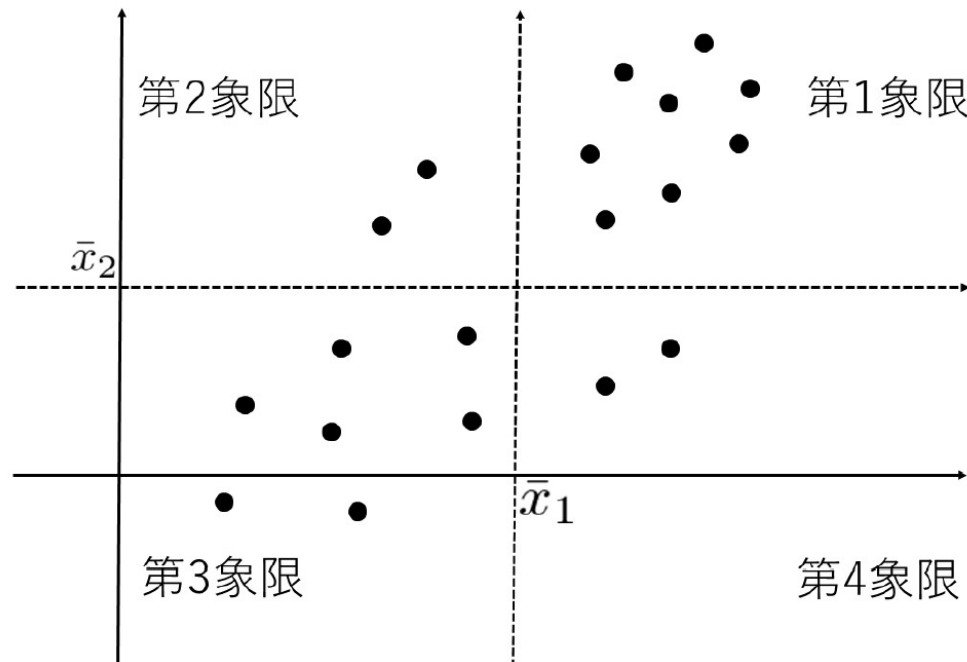
□量的データの二変量に関する要約

共分散:データの線型関係が仮定されるときその強さを測る指標

$$\frac{1}{\text{被験者数}} \quad \times$$

$$\left\{ (\text{被験者1のデータX} - \text{データXの平均値}) (\text{被験者1のデータY} - \text{データYの平均値}) + \dots + (\text{被験者nのデータX} - \text{データXの平均値}) (\text{被験者nのデータY} - \text{データYの平均値}) \right\}$$

第1象限と第2象限にデータ値が集まると共分散は高くなる！



2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの二変量に関する要約

共分散はデータ値の単位に値が影響されることから値がどれほど大きいと線形性があるのかわからないのかという議論が困難

データの線型性の度合を測る指標として相関係数が有用

$$\text{データXとデータYの相関係数} = \frac{\text{データXとデータYの共分散}}{\text{データXの標準偏差} \times \text{データYの標準偏差}}$$

$$-1 \leq \text{相関係数} \leq 1$$

− 1 : 負の相関関係

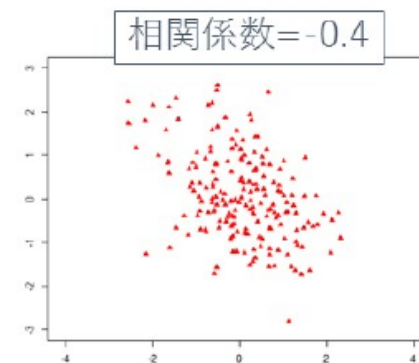
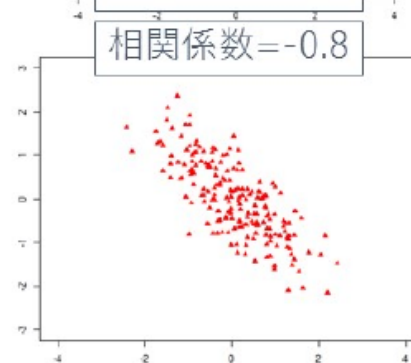
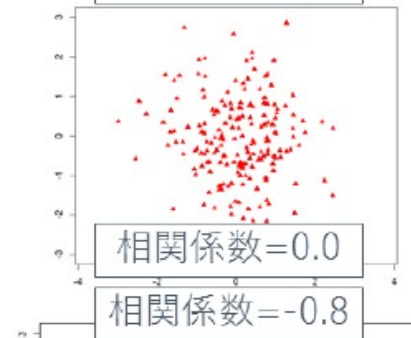
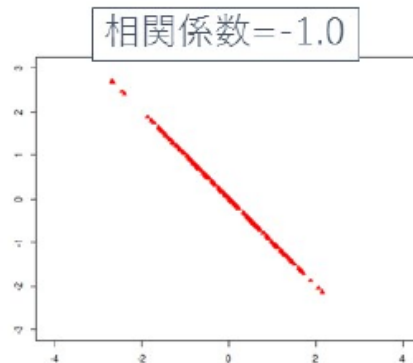
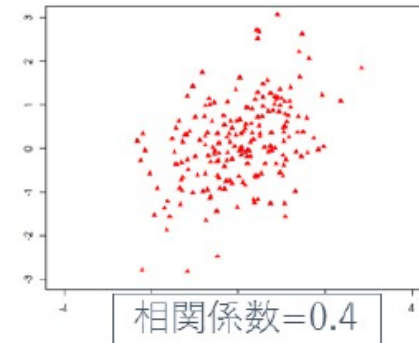
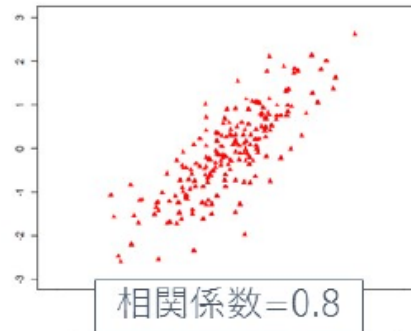
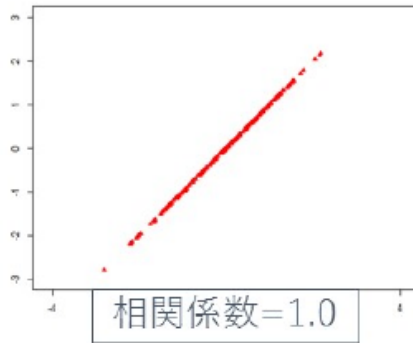
0 : 相関がない

1 : 正の相関関係

2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの二変量に関する要約

相関関係と散布図



2.データ尺度別の集計方法(記述統計とグラフ表示)

□量的データの二変量に関する要約

相関係数について問題

下記の文章は正しいか？

「相関係数の値が1に近ければ，正の相関関係がある」



2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの一変量に関する要約

縦隔腫瘍

縦隔腫瘍	度数	相対度数
Large	113	0.131
Small	288	0.333
No	464	0.536
計	865	1.000

度数：質的変数の項目に該当する個体(被験者)数

相対度数：項目に該当する個体(被験者数)の比率(度数/合計数), 比率とも呼ばれる.

分布：項目とその相対度数(比率)の関係をいう. それを表にしたものが度数分布表である.

2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの一変量に関する要約

これは、胃癌患者に対するある臨床試験の患者背景のうち、組織型のみを抽出したものである。

組織型	度数	相対度数
乳頭腺癌(pap)	2	0.016
管状腺癌(tub)	70	0.565
低分化腺癌(por)	46	0.371
印環細胞癌(sig)	4	0.032
粘液癌(muc)	1	0.008
その他	1	0.008
計	124	1.000



組織型	度数	相対度数
管状腺癌(tub)	70	0.565
低分化腺癌(por)	46	0.371
印環細胞癌(sig)	4	0.032
乳頭腺癌(pap)	2	0.016
粘液癌(muc)	1	0.008
その他	1	0.008
計	124	1.000

各項目に順序関係がない(名義カテゴリーカル・データ)での度数分布表の順番を替えることはできる。

並べ替えを行うことで、省察することが平易になる。

2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの二変量に関する要約

治療法			結節外疾患		
治療法	度数	相対度数	結節外疾患	度数	相対度数
RT	616	0.712	Yes	79	0.091
CMT	249	0.288	No	786	0.909
計	865	1.000	計	865	1.000

治療法と結節外疾患には関係がある(例えば、治療法によって結節外疾患に違いがある?)



2つないしそれ以上の質的変数を組み合わせて、同時に集計することで2変数の関連性を調べるための表を**クロス集計表(分割表)**という

		結節外疾患		
		Y	N	計
治療法	RT	29	587	616
	CMT	50	199	249
	計	79	786	865

2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの二変量に関する要約

行パーセント

列毎でパーセントをとった場合をいう(RT：放射線, CMT：放射線+抗癌剤)

		結節外疾患		
		Yes	No	計
治療法	RT	29 (4.7%)	587 (95.3%)	616 (100.0%)
	CMT	50 (20.1%)	199 (79.9%)	249 (100.0%)
	計	79 (9.1%)	786 (90.9%)	865 (100.0%)

Pr(結節外疾患|治療法)がわかる.

- ・ $\text{Pr}(\text{RT}|\text{Yes})=4.7\%$ (治療法がRTの患者で結節外疾患の確率は4.7%)
- ・ $\text{Pr}(\text{RT}|\text{No})=95.3\%$ (治療法がRTの患者で結節外疾患でない確率は95.3%)
- ・ $\text{Pr}(\text{CMT}|\text{Yes})=20.1\%$ (治療法がCMTの患者で結節外疾患の確率は20.1%)
- ・ $\text{Pr}(\text{CMT}|\text{No})=79.9\%$ (治療法がCMTの患者で結節外疾患でない確率は79.9%)

治療法の群間で結節外疾患の罹患割合に約15%の違いが認められる

2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの二変量に関する要約

列パーセント

行毎でパーセントをとった場合をいう(RT：放射線, CMT：放射線+抗癌剤)

		結節外疾患		
		Yes	No	計
治療法	RT	29 (36.7%)	587 (74.7%)	616 (71.2%)
	CMT	50 (63.3%)	199 (25.3%)	249 (28.8%)
	計	79 (100.0%)	786 (100.0%)	865 (100.0%)

Pr(結節外疾患|治療法)がわかる。

- $\text{Pr}(\text{Yes}|\text{RT})=36.7\%$ (結節外疾患患者の治療法がRTである確率は36.7%)
- $\text{Pr}(\text{Yes}|\text{CMT})=63.3\%$ (結節外疾患患者の治療法がCMTである確率は63.3%)
- $\text{Pr}(\text{No}|\text{RT})=74.7\%$ (結節外疾患でない患者の治療法がRTである確率は74.7%)
- $\text{Pr}(\text{No}|\text{CMT})=25.3\%$ (結節外疾患でない患者の治療法がCMTである確率は25.3%)

結節外患者において、治療法に37.7%の違いがあった

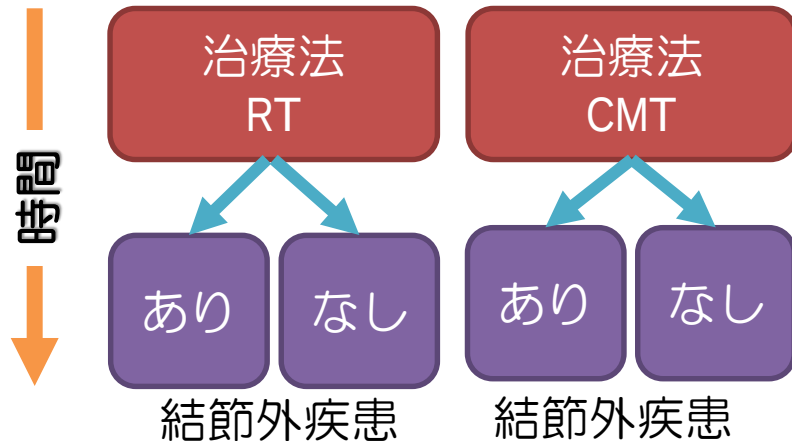
2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの二変量に関する要約

行パーセント

		結節外疾患		
		Y	N	計
治療法	RT	29	587	616
	CMT	50	199	249
	計	79	786	865

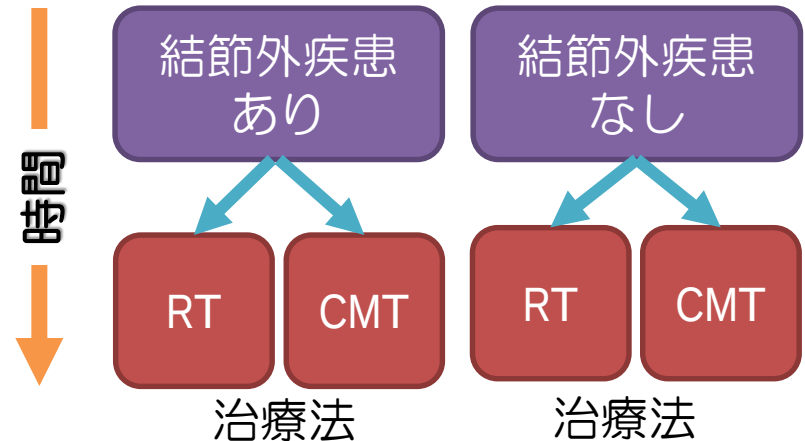
パーセントをとる方向



列パーセント

		結節外疾患		
		Y	N	計
治療法	RT	29	587	616
	CMT	50	199	249
	計	79	786	865

パーセントをとる方向

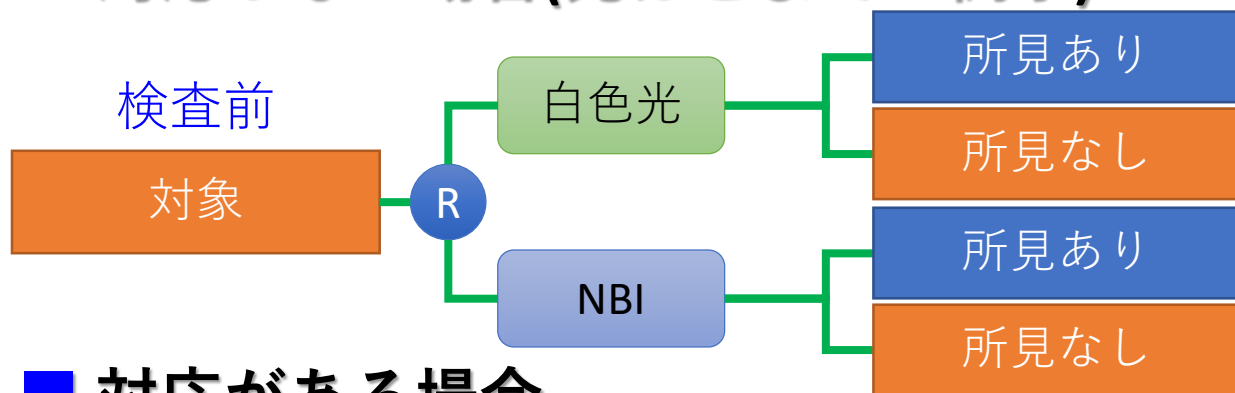


2.データ尺度別の集計方法(記述統計とグラフ表示)

□質的データの二変量に関する要約

いま、ある消化器癌に対する内視鏡検査において、NBIと白色光の比較を行ったときの所見の有無の比較を行う仮想的な試験があったとする。

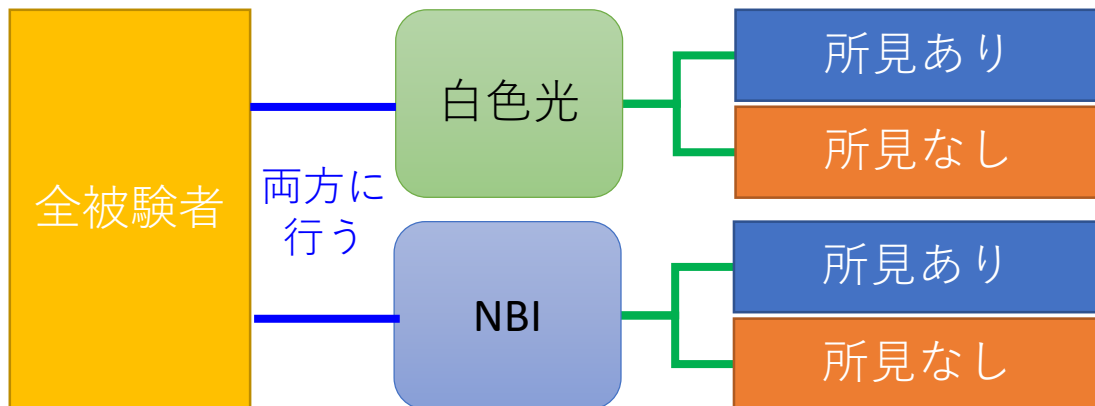
■対応がない場合(先ほどまでの例示)



クロス集計表

		所見	
		あり	なし
群	白色光	a	b
	NBI	c	d

■対応がある場合



クロス集計表

		NBI	
		陽性	陰性
白色光	陽性	a	b
	陰性	c	d

3.統計ソフトウェアについて

□当センターで貸出を行っているソフトについて

- JMPproを本学に籍がある方を対象として臨床研究センターでは無料で貸し出ししている

<https://waidai-csc.jp/gakunai/jmppro/>



もし、使用方法で不明な点がございましたら臨床研究センターまで

ご清聴ありがとうございました

