

# データの記述

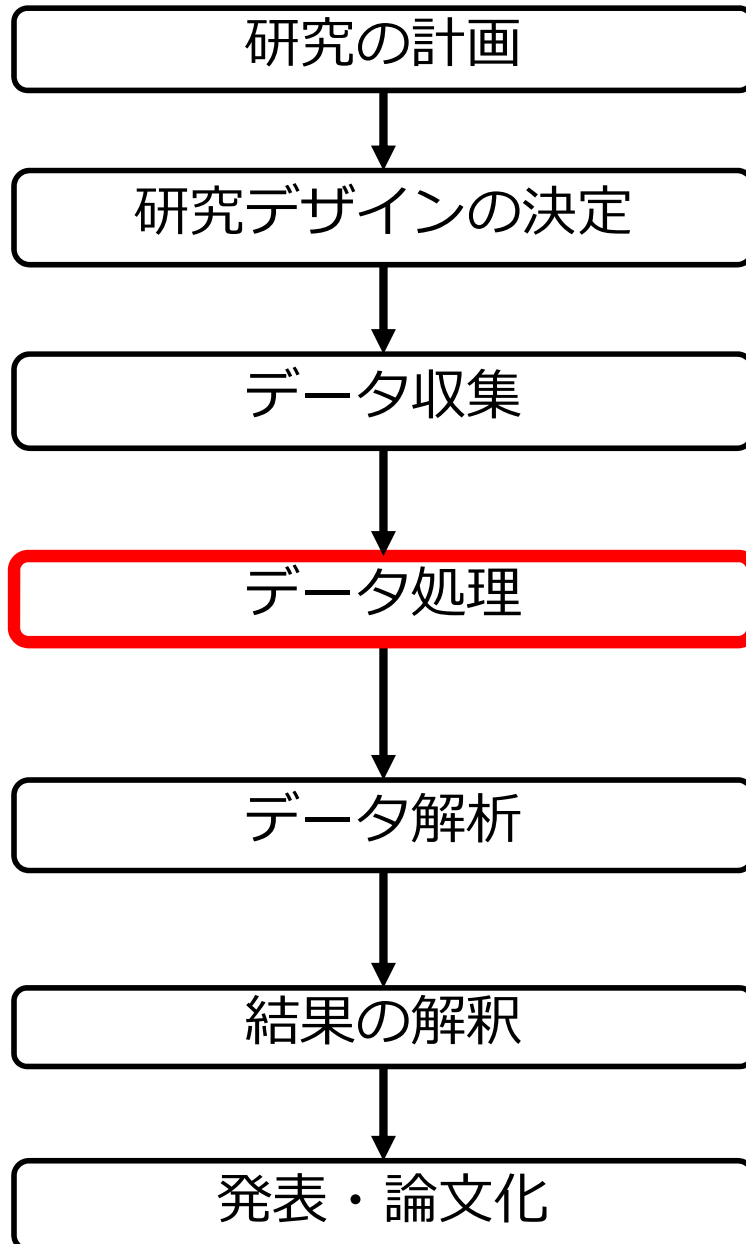
生物統計部/JCOGデータセンター

水澤 純基

生物統計基礎セミナー

2016. 4.11

# 一般的な研究の流れ



臨床的仮説は？

前向き？後ろ向き？単群？ランダム化？

どの患者のどのデータを収集？

**おかしいデータはないか？**  
**データの示し方は？**

群間で差があるか？XXは予後因子？

臨床的仮説は解決した？研究の限界は？  
次の研究は？

# 講義の内容

- データ解析をする前の注意事項
  - 入力されたデータを疑う
  - エクセルデータの罫
- 1変数の連続量データのまとめ方
  - ヒストグラム
  - 要約統計量
- 2変数の連続量データのまとめ方
  - 散布図と相関係数
  - 連続量データのカットオフ
- カテゴリカルデータのまとめ方
  - 分割表
- 解析対象集団とpatients flow diagram

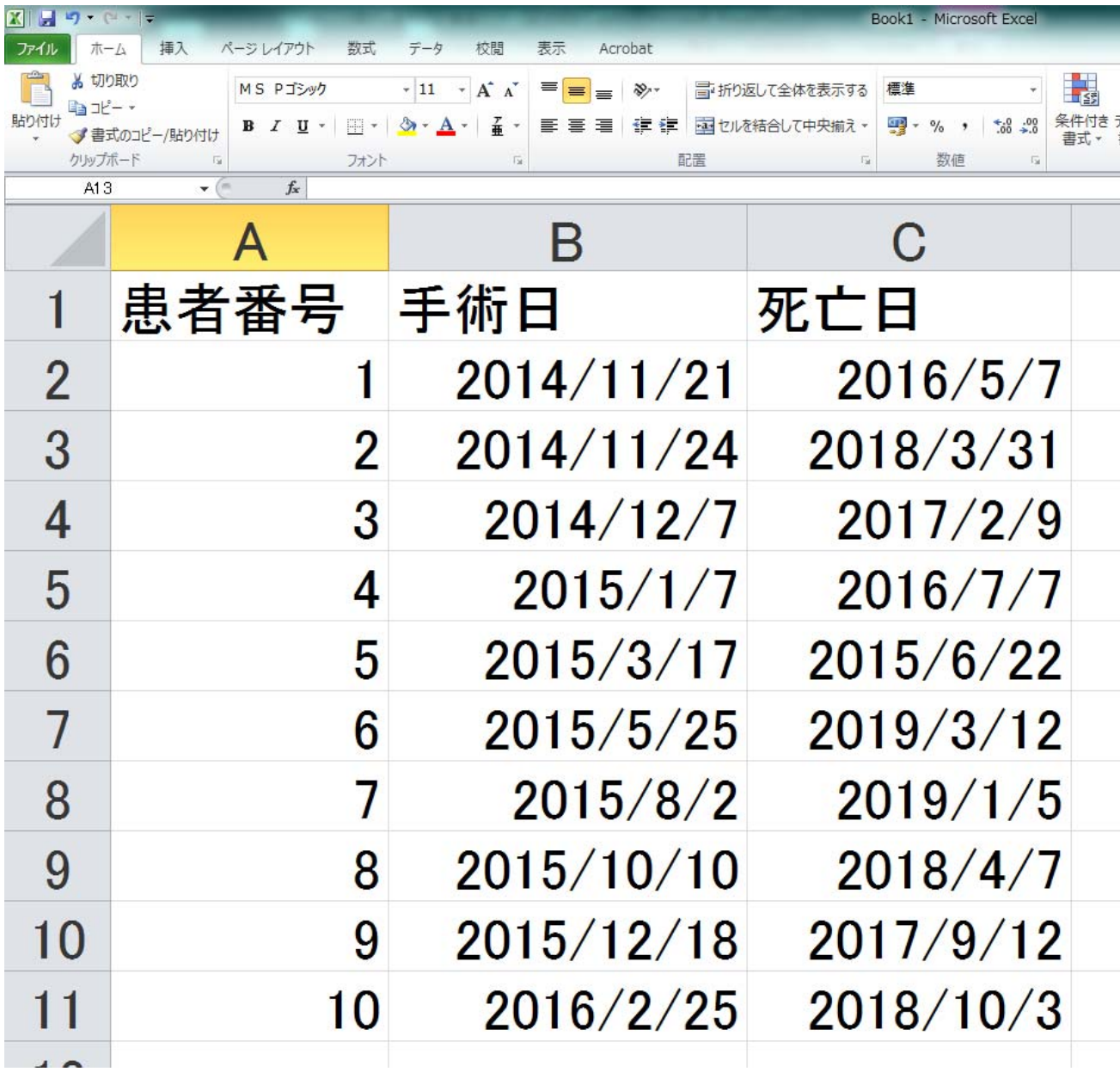
# データ解析をする前の注意事項

統計解析する以前の落とし穴

# そもそも入力されたデータは正しい？

- 研究者がコンピュータにデータ入力した場合、入力ミスが0%ということはありません
  - 1フィールドあたりの入力ミスは0.5% - 6.9%との報告
    - Neaton, James D., et al. *Statistics in medicine* 9.1-2 (1990): 115-124.
    - Norton, Susan L., et al. *Computers and Biomedical Research* 14.2 (1981): 179-198.
    - Kronmal, Richard A., et al. *Computers and Biomedical Research* 11.6 (1978): 553-566.
  - ベテランのデータマネージャーと一般の臨床医の間でも、かなりの差あり
    - by Dr.福田@JCOG DC
- 臨床研究のデータセンターで行っていること
  - ダブルエントリー（2人で別々に入力して一致するか確認）
  - 読み合わせ（1人で入力し、入力したデータを声に出して読み上げ、元データ[カルテやCRF]のデータが一致しているか確認）
  - ダブルエントリーの方が入力ミスは少なくなるが時間がかかる
    - Kawado, Miyuki, et al. *Controlled clinical trials* 24.5 (2003): 560-569.

# Macユーザーの私がエクセルでデータを貰いました



	A	B	C
1	患者番号	手術日	死亡日
2	1	2014/11/21	2016/5/7
3	2	2014/11/24	2018/3/31
4	3	2014/12/7	2017/2/9
5	4	2015/1/7	2016/7/7
6	5	2015/3/17	2015/6/22
7	6	2015/5/25	2019/3/12
8	7	2015/8/2	2019/1/5
9	8	2015/10/10	2018/4/7
10	9	2015/12/18	2017/9/12
11	10	2016/2/25	2018/10/3

- 多くのデータが未来の日付になっている！
- データ提供元に確認しても、「正しく入力している」との返答
- 一体何故？

# MacとWindowsは日付の起算日が異なる

## Windowsの設定

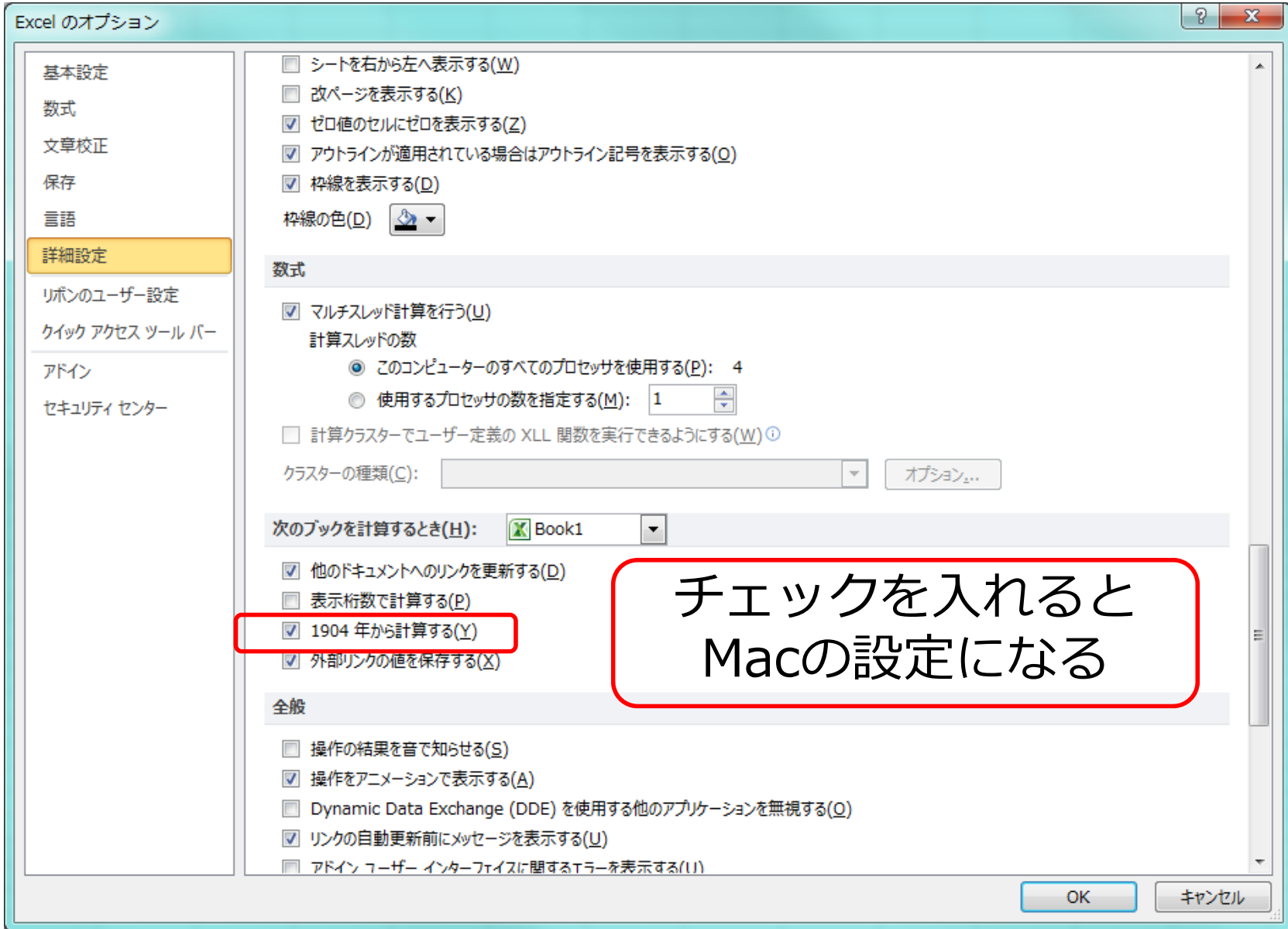
基準日：1900/1/1を1としてカウント

## Macの設定

基準日：1904/1/1を0としてカウント

A	B	C		A	B	C
患者番号	手術日	死亡日		患者番号	手術日	死亡日
1	2010/11/20	2012/5/6		1	2014/11/21	2016/5/7
2	2010/11/23	2014/3/30		2	2014/11/24	2018/3/31
3	2010/12/6	2013/2/8		3	2014/12/7	2017/2/9
4	2011/1/6	2012/7/6		4	2015/1/7	2016/7/7
5	2011/3/16	2011/6/21		5	2015/3/17	2015/6/22
6	2011/5/24	2015/3/11		6	2015/5/25	2019/3/12
7	2011/8/1	2015/1/4		7	2015/8/2	2019/1/5
8	2011/10/9	2014/4/6		8	2015/10/10	2018/4/7
9	2011/12/17	2013/9/11		9	2015/12/18	2017/9/12
10	2012/2/24	2014/10/2		10	2016/2/25	2018/10/3

# [ファイル]→[オプション]→[詳細設定]





# 1変数の連続量データのまとめ方

t検定、Wilcoxon順位和検定をする前に

# アウトカムの型と統計手法

	連続量	2値(0 / 1)	生存時間
アウトカムの例	血圧値、 臨床検査値	奏効割合 有害事象発生割合	全生存期間 無増悪生存期間
データの要約法	ヒストグラム・散布図	分割表	Kaplan-Meier曲線
群間比較(検定)	t検定、 Wilcoxon検定	カイ2乗検定、 Fisherの 直接確率検定	ロジスティック 検定
モデルのあてはめ	重回帰分析	ロジスティック回帰	Cox回帰

# 検定する前にデータの視覚化と要約を！

## 視覚化

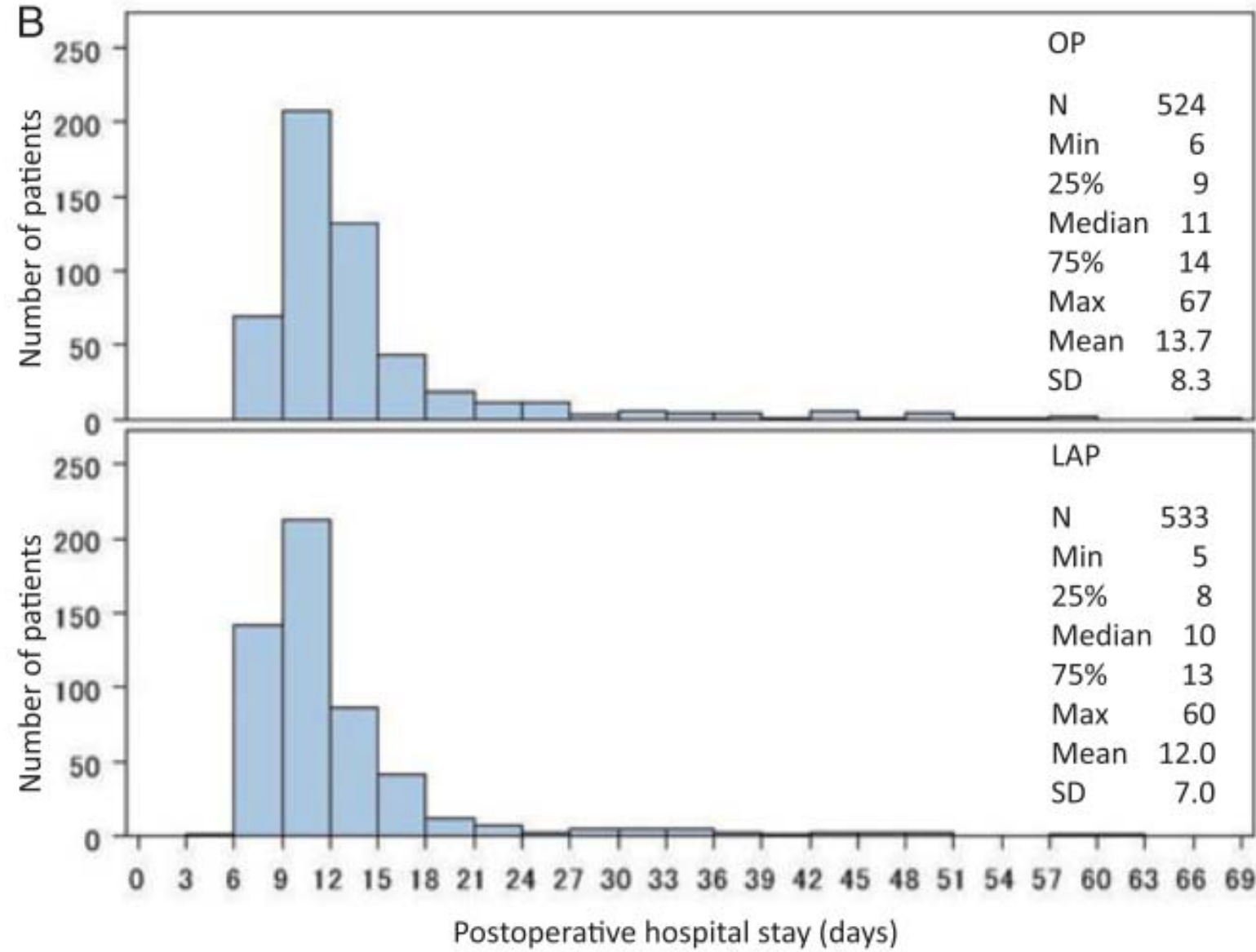
- グラフ
  - ヒストグラム
  - 箱ひげ図

## 要約統計量

- データの位置・大きさ
  - 平均値
  - 中央値
- データのばらつき・広がり
  - 標準偏差
  - 第1四分位点、第3四分位点  
(Interquartile range[IQR])
  - 最小値、最大値 (Range)

# ヒストグラム：1変量に対するグラフ表現

## 大腸がんに対する外科的切除後の入院日数



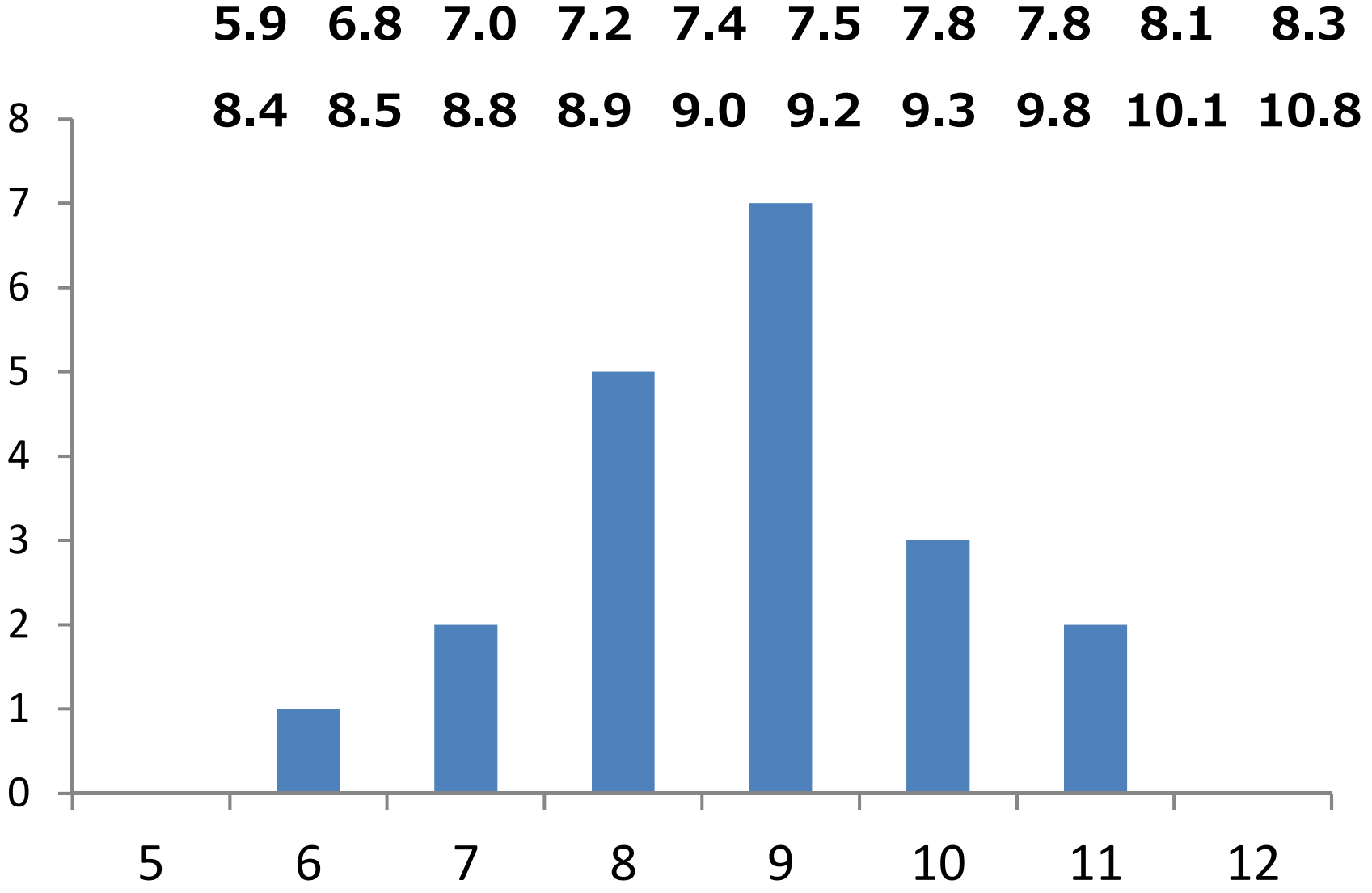
開腹切除術  
(OP)

腹腔鏡下  
切除術  
(LAP)

# ヒストグラムを描く目的

- 外れ値やおかしなデータがないかどうか調べる
  - クレアチニン値=0.02など
  - 臨床研究の適格規準外のデータがないか
- 層別の必要性がないかどうか調べる
  - グループが存在するかどうか
  - 2峰性になっていないかどうか
- 分布の形状を調べる
  - 左右対称か、片方に裾を引いているか

# 20人分のデータ数値例



# データが正規分布に従う場合の要約指標

- 平均値(mean)

$$\text{平均値} = \frac{5.9 + 6.8 + \dots + 10.1 + 10.8}{20} = 8.33$$

- 標準偏差 (standard deviation [SD])

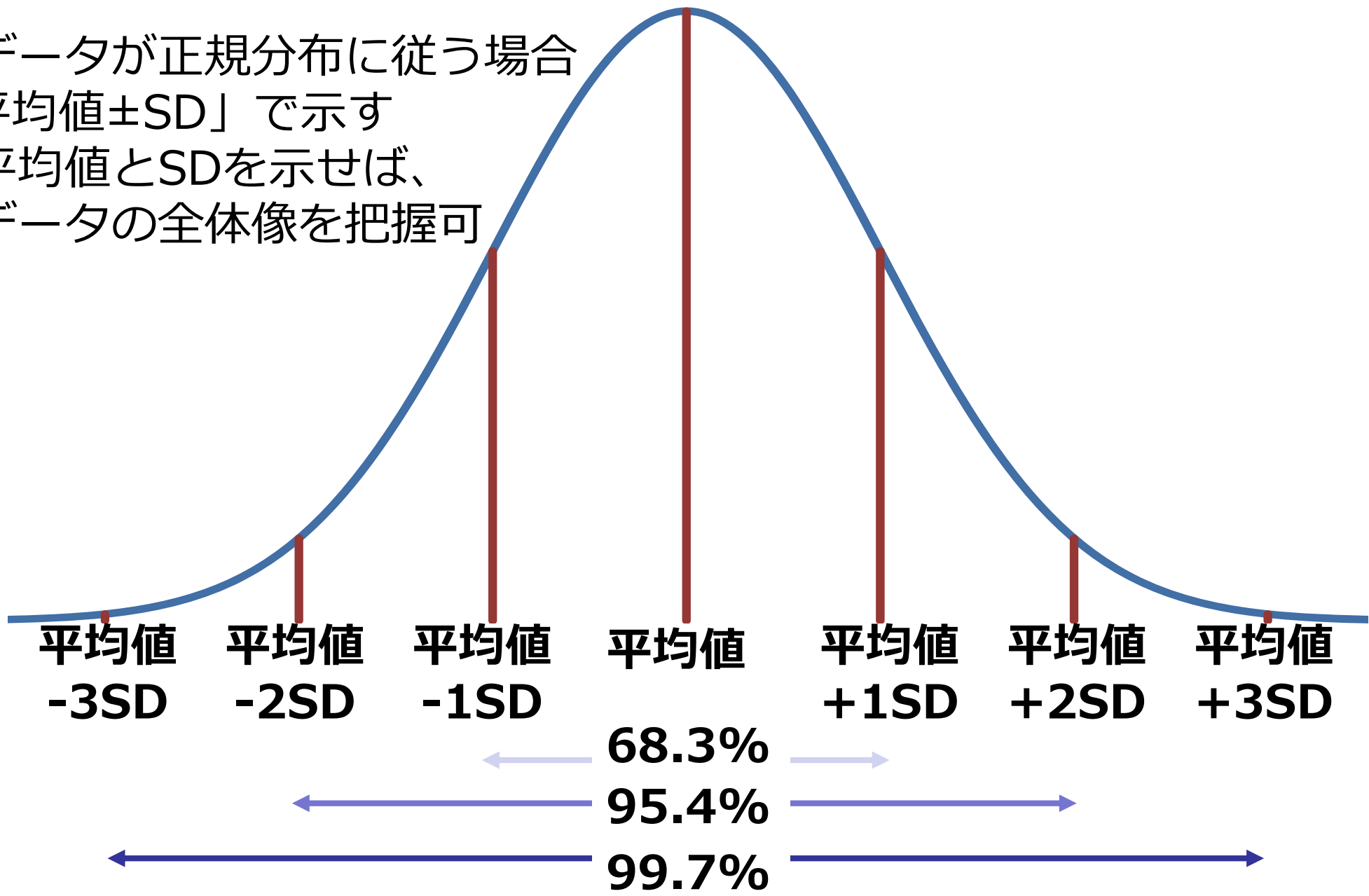
- 分子：各データと平均値との差を2乗して合計したもの
- 分母：データの数 - 1
- 上記の分数の平方根をとったもの

$$SD = \sqrt{\frac{(5.9 - 8.33)^2 + (6.8 - 8.33)^2 + \dots + (10.1 - 8.33)^2 + (10.8 - 8.33)^2}{20 - 1}}$$

= 1.13

# 正規分布の1SD、2SD、3SDのルール

- データが正規分布に従う場合「平均値±SD」で示す
- 平均値とSDを示せば、データの全体像を把握可

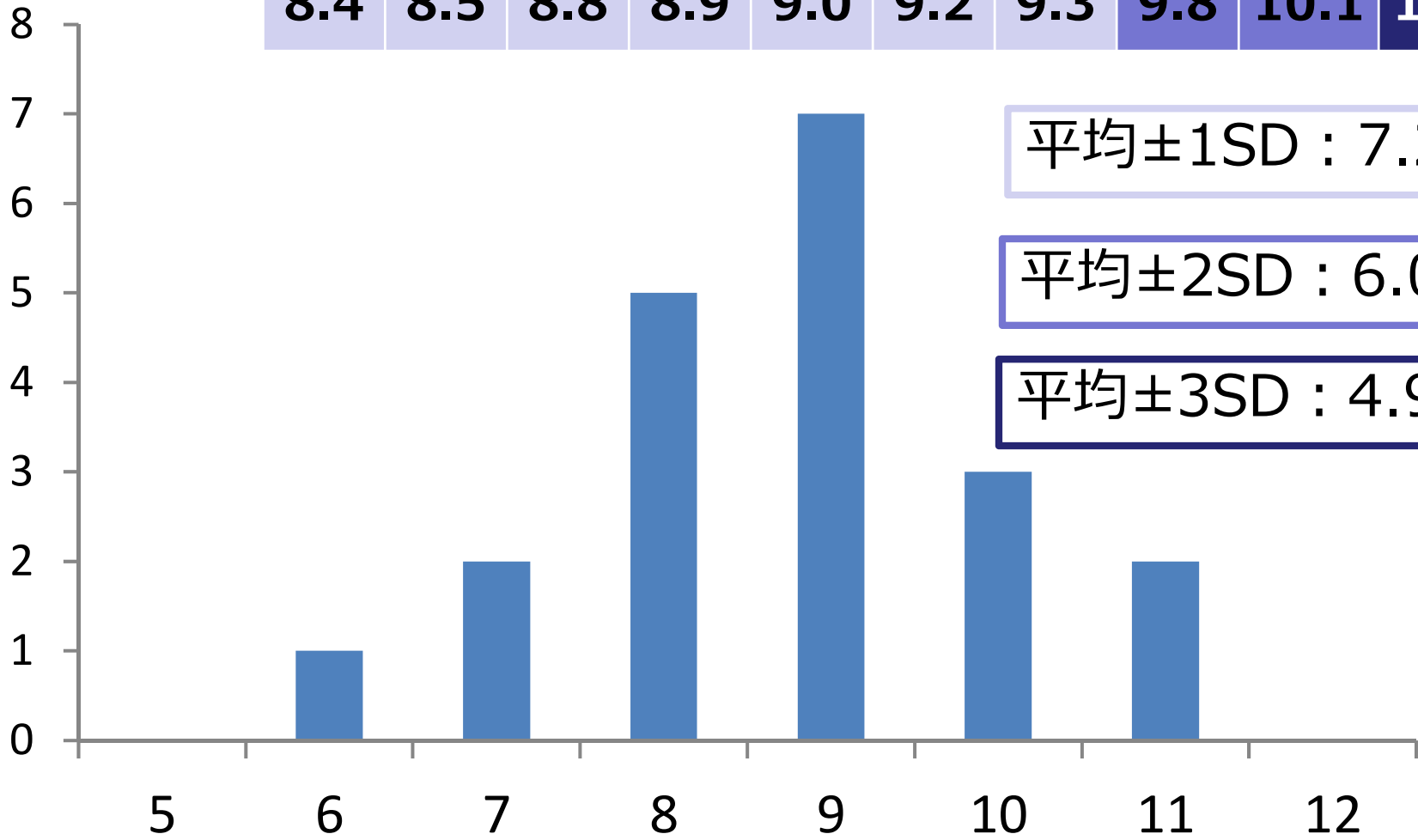




# 20人分のデータ数値例の場合

## 20人分のデータ

5.9	6.8	7.0	7.2	7.4	7.5	7.8	7.8	8.1	8.3
8.4	8.5	8.8	8.9	9.0	9.2	9.3	9.8	10.1	10.8



平均±1SD : 7.20-9.46

平均±2SD : 6.09-10.57

平均±3SD : 4.97-11.69

# 平均値の解釈は？

- 国税庁の「民間給与の実態調査結果」によると、平成22年度において男性の平均年収は507.4万円であった
- このデータから得られる解釈は？
  - 507万円以下、507万円以上の人が50%ずつ
  - 507万円よりも少ない人が50%以上いる
  - 507万円よりも多い人が50%以上いる

# 平均値の特徴を理解しよう

- 数値例

- 1, 3, 5, 7, 9 → 平均値 5、中央値5

- 1, 3, 5, 7, **100** → 平均値 23.2、中央値5

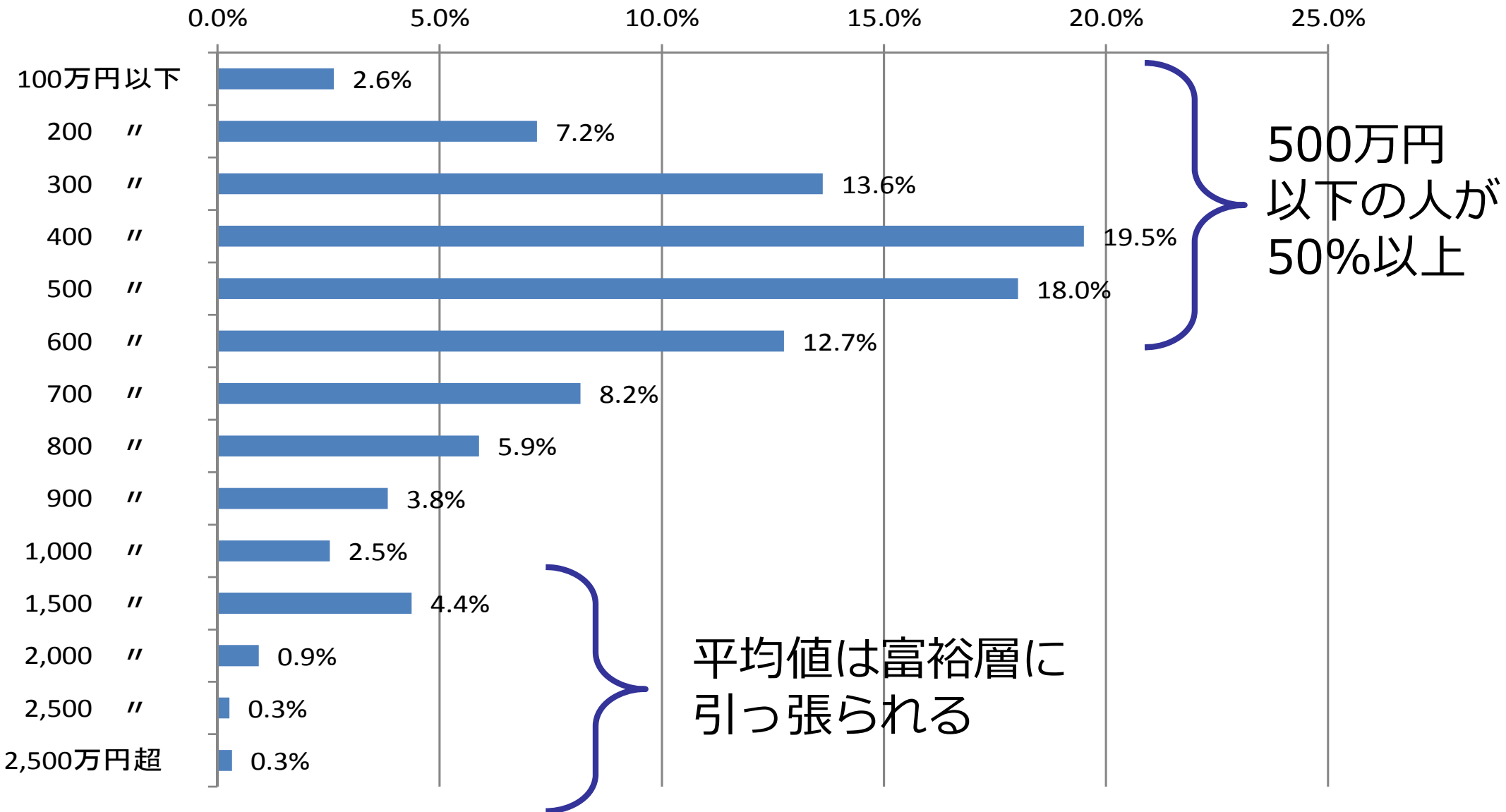
- 平均値は他と比べて極端に大きな(小さな)値に影響されやすい

- 大きな(小さな)値 = 外れ値

- 中央値は常にデータの真ん中の値を示す

- 外れ値の影響を受けにくい = ロバスト

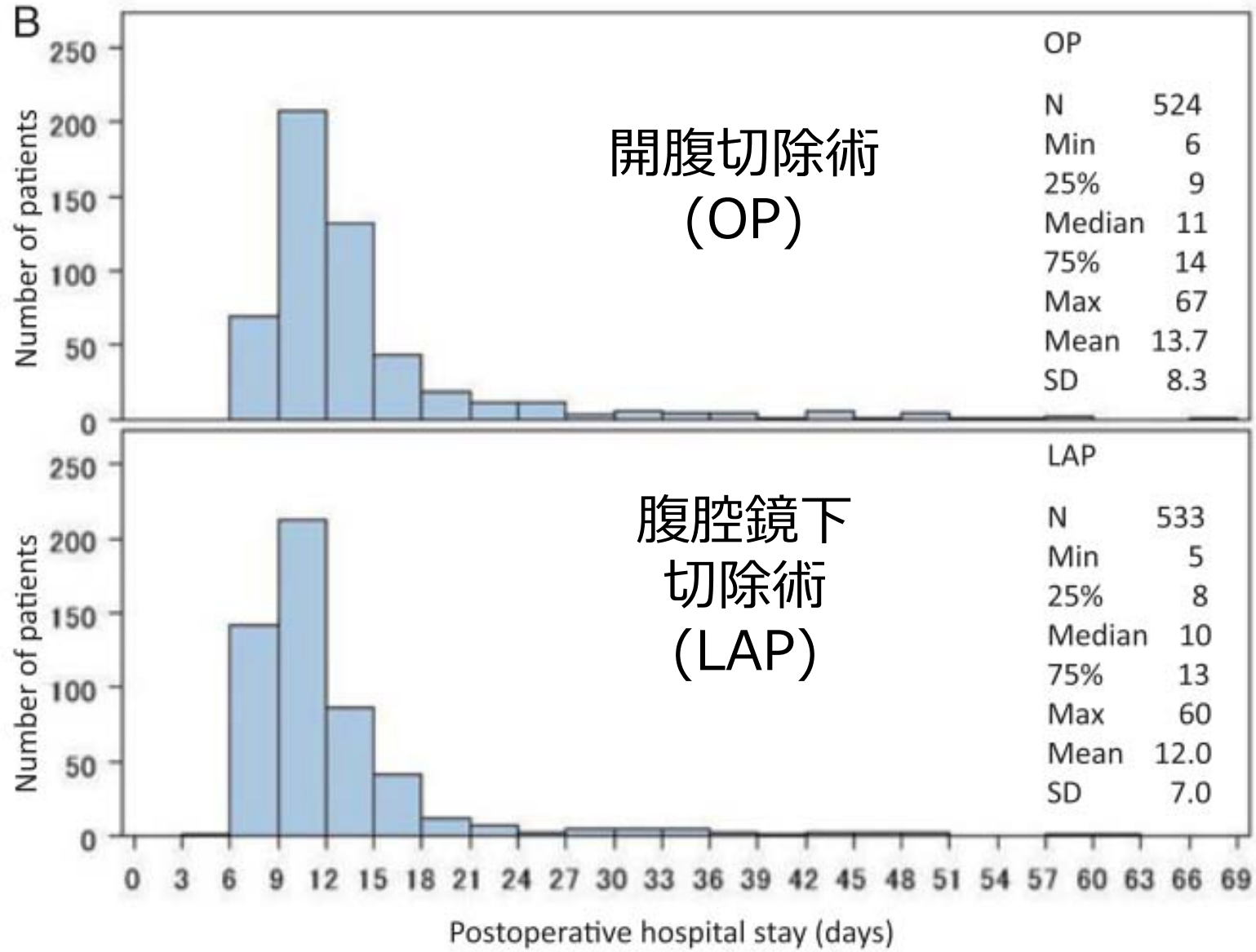
# 平均値は分布の中心ではない



<http://www.nta.go.jp/kohyo/tokei/kokuzeicho/minkan2010/minkan.htm> より作成

# データ正規分布ではない場合

## 大腸がんに対する外科的切除後の入院日数



- ヒストグラムが右に裾を引いている
- 平均値±SDではデータの全体像を把握できない
- 平均値、SD以外の要約指標が必要

# データが右に裾を引くことが疑われる例

- 肝切除後の合併症と背景因子の検討

**Table 1.** Comparison of Comorbidities: Overall and by Morbidity Status

Variable	Overall (n = 783)	With complication (n = 177)	Without complication (n = 606)	p Value
Male gender (%)	48.5	55.4	46.5	0.041
Age (y), mean ± SD	57.5 ± 13.7	58.3 ± 13.8	57.2 ± 13.7	0.340
Caucasian race (%)	80.1	76.8	81.0	0.240
Cardiovascular (%)				
General	<b>14.2±25.1</b>	<b>16.6±24.7</b>	<b>13.5±25.1</b>	
ASA class 3 or higher (%)	54.5	65.0	39.9	0.002
Pack-years of tobacco, mean ± SD	14.2 ± 25.1	16.6 ± 24.7	13.5 ± 25.1	0.167
Smoker (%)	18.4	21.5	17.5	0.227
Alcohol use (%)	4.9	5.7	4.6	0.554
Emergency status (%)	0.8	1.1	0.7	0.622
DNR status (%)	0.0	0.0	0.0	1.000

- 正の値しか取らない状況で、**平均値 ≤ SD** の場合、分布の形状が左右非対称な裾を引いた分布になっていると考えて良い
- 平均値 ≤ SD の場合、平均値とSDによる要約は不適切
  - 大雑把に「平均値 > 2SD」が平均値 ± SDを示す目安

# データが正規分布に従わない場合の要約指標①

- 中央値(median)
  - データの大きさの順に並べて
    - データが奇数個なら中央に位置するデータの値
    - データが偶数個なら中央に位置する2つのデータの平均

$$\text{中央値} = \frac{8.3 + 8.4}{2} = 8.35$$

- 第1四分位点 (25%点)
  - データを大きさの順に並べて最小値から数えて1/4番目にあたるデータ
- 第3四分位点 (75%点)
  - データを大きさの順に並べて最小値から数えて3/4番目にあたるデータ

$$\text{第1四分位点} = \frac{7.4 + 7.5}{2} = 7.45, \quad \text{第3四分位点} = \frac{9.0 + 9.2}{2} = 9.10$$

## データが正規分布に従わない場合の要約指標②

- 四分位範囲 (interquartile range [IQR])
  - 第1四分位点(25%点)～第3四分位点(75%点)
  - データの半分がばらついている範囲を示す
- 範囲 (range)
  - 最小値～最大値
  - データの全範囲を示す



# データ正規分布ではない場合の示し方

TABLE 3. Operative Results

	Open Surgery (n = 524)	Laparoscopic Surgery (n = 533)	P
<b>Procedures</b>			
Ileocecal resection	54 (10.3%)	37 (6.9%)	
Right hemicolectomy	100 (19.1%)	110 (20.6%)	
Sigmoidectomy	208 (39.7%)	240 (45.0%)	
Anterior resection	155 (29.6%)	137 (25.7%)	0.1534
Partial resection	4 (0.8%)	7 (1.3%)	
Hartmann	1 (0.2%)	0	
Others	2 (0.4%)	2 (0.4%)	
<b>Lymphadenectomy</b>			
D3	521 (99.4%)	529 (99.2%)	1
<b>Estimated blood loss, mL</b>			
Median	85	30	<0.001
Range	0-3395	0-4080	
Interquartile range	49.5-180	10-70	
<b>Operative time, min</b>			
Median	159	211	<0.001
Range	68-710	80-616	
Interquartile range	130-188.5	179-256	
<b>Blood transfusion within 3 postoperative days</b>			
Yes	6 (1.1%)	4 (0.8%)	0.5437

- 3つセットで報告するのが最も良い

– Median

– Interquartile range

– Range

- データのばらつきを解釈するためには、範囲よりも四分位範囲の方がよい

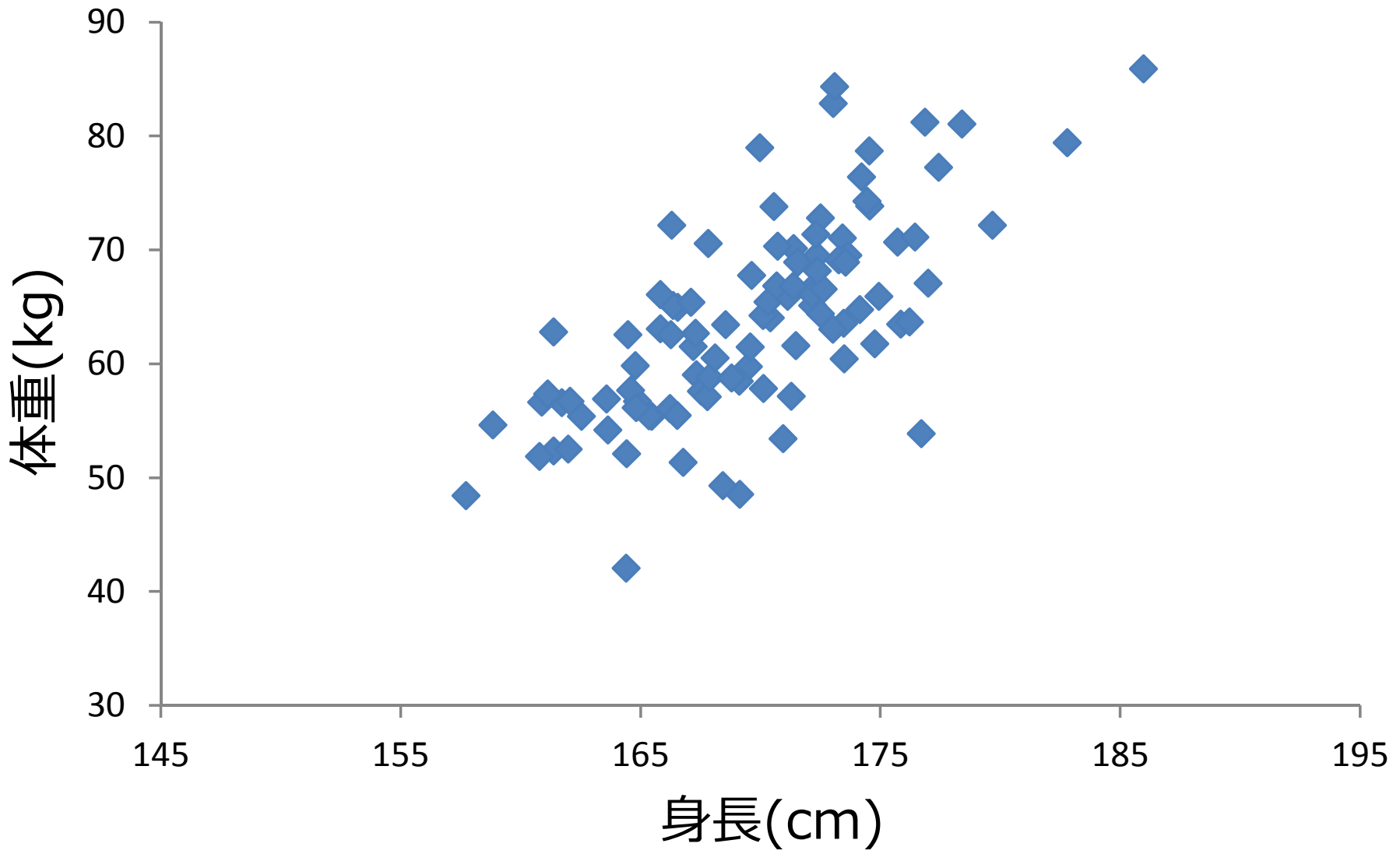
# 臨床データについては平均値よりも中央値

- 臨床のデータは右に裾を引いた分布が多い
  - 臨床検査値
  - バイオマーカー
  - 手術時間、出血量
- 臨床のデータは中央値を用いる方が無難
- 左右対称ではない分布の位置のずれを検定したいならt検定ではなく、Wilcoxon順位和検定
  - 詳しくはICRwebの「生物統計基礎セミナー 仮説検定2」をご覧ください
  - [http://int.east.ncc.go.jp/rinsho/semi\\_seibutsu/kiso/20150428\\_kuchiba.pdf](http://int.east.ncc.go.jp/rinsho/semi_seibutsu/kiso/20150428_kuchiba.pdf)

# 2つの連続量データの関係を 確認する方法

# 散布図 (scatter plot)

- 100人の男性の身長[横軸/cm]と体重[縦軸/kg]をプロットしたもの



# 散布図を描く目的

- 外れ値やおかしなデータがないかどうか調べる
  - 身長190cmで体重45kg
  - 個々のデータは正常だが組み合わせると奇妙なデータ
- 層別の必要性がないかどうか調べる
  - グループが存在するかどうか
  - グループごとに2変数間の関係は異なるか
- 相関のおおよその程度をつかんでおく
  - 直線的 or 曲線的？、その強さは？

# 相関 (correlation)

- 元々の定義は「連続型の2変数間の直線的な関連」
  - 関連 (association) はあっても直線関係でないものは本来「相関がある」とは言わない
    - 「遺伝子発現量のhighとlowと予後が“相関”している」は2つのカテゴリーしか確認していないので、適切な表現とはいえない
- 直線的な関連以外は単に「関連がある (associate) 」とした方がよい

# 相関係数

- 相関を測る尺度の一つ。
  - 範囲は-1 ~ 1
    - -1 : Xが増加するとYが減少する。完全な負の相関
    - 0 : 相関なし。ただし、「関連がない」という意味ではない
    - 1 : Xが増加するとYが増加する。完全な正の相関.
- Pearsonの積率相関係数( $r$ )
  - 2つの変数間に直線的な関係があるかどうかを測る指標
- Spearman順位相関係数( $\rho$ )
  - 一方が増加(減少)すれば、もう一方も増加(減少)するかを測る
    - 直線的な関係かどうかは関係ない

# 相関係数の式（覚える必要ありません）

- Pearsonの積率相関係数(r)

$$r = \frac{\sum_{i=1}^n (\text{各データの } x \text{ 値} - x \text{ の平均}) (\text{各データの } y \text{ 値} - y \text{ 値の平均})}{\sqrt{\left[ \sum_{i=1}^n (\text{各データの } x \text{ 値} - x \text{ の平均})^2 \right] \left[ \sum_{i=1}^n (\text{各データの } y \text{ 値} - y \text{ 値の平均})^2 \right]}}$$

- Spearman順位相関係数(ρ)

- 全てのデータを順位尺度に変換してPearsonの相関係数の式に

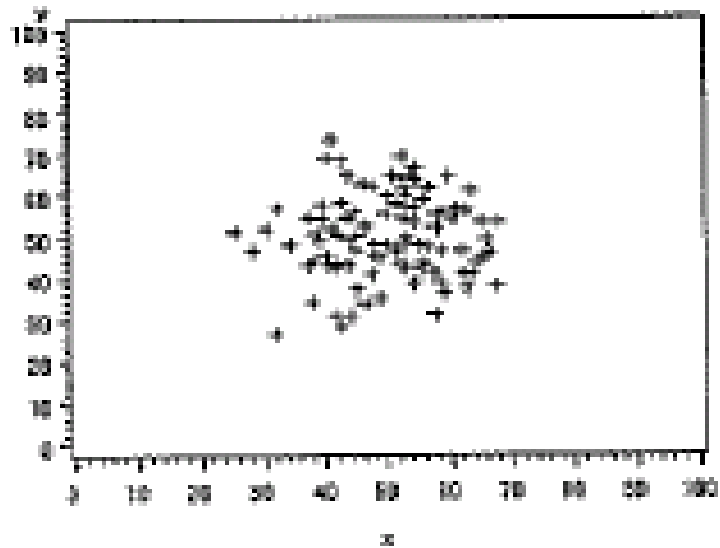
x	y
96	268
86	318
101	289
100	301
98	298



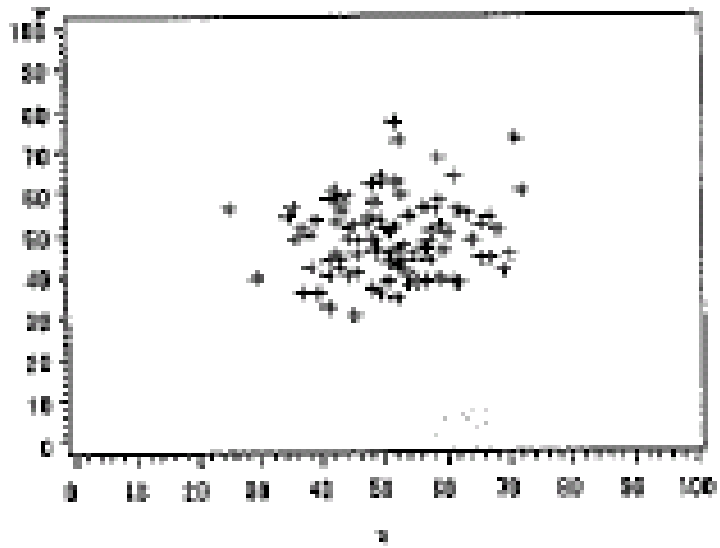
x	y
2	1
1	5
5	2
4	4
3	3



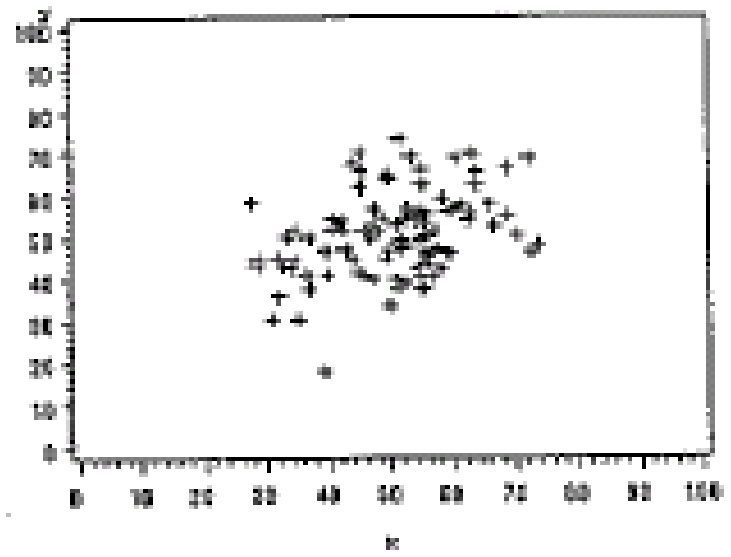
# ピアソンの相関係数と散布図①



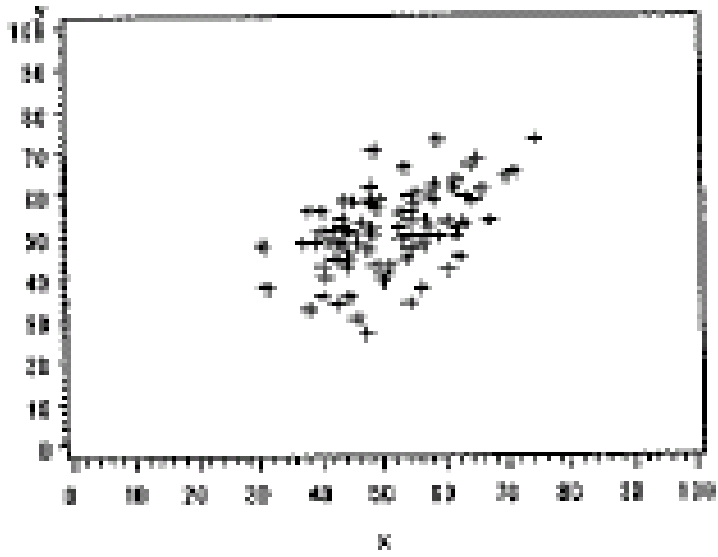
$r=0.0$



$r=0.2$



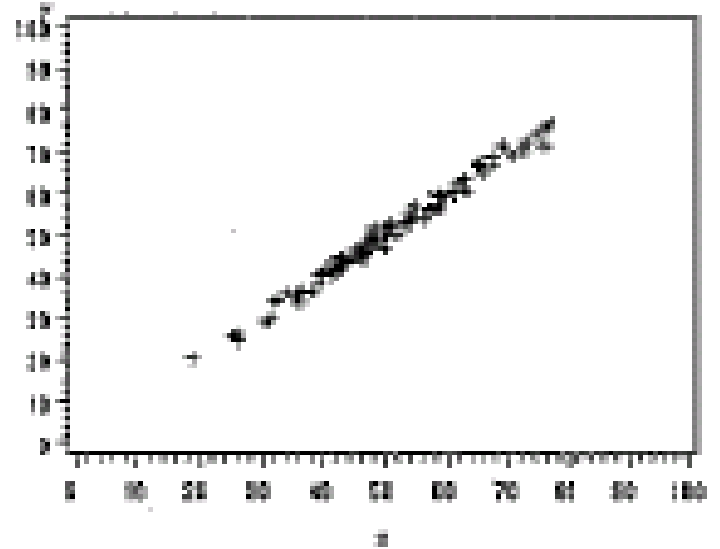
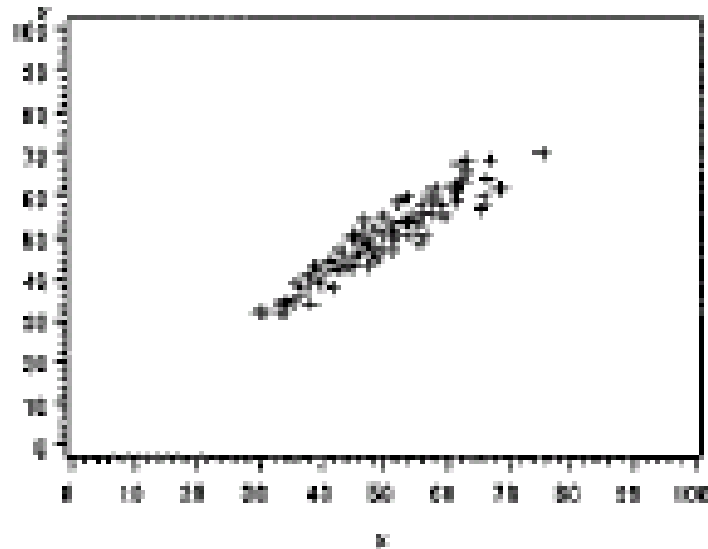
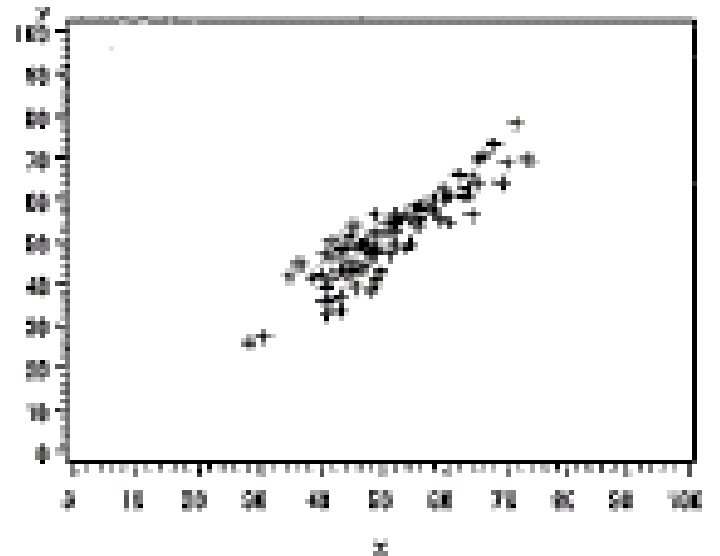
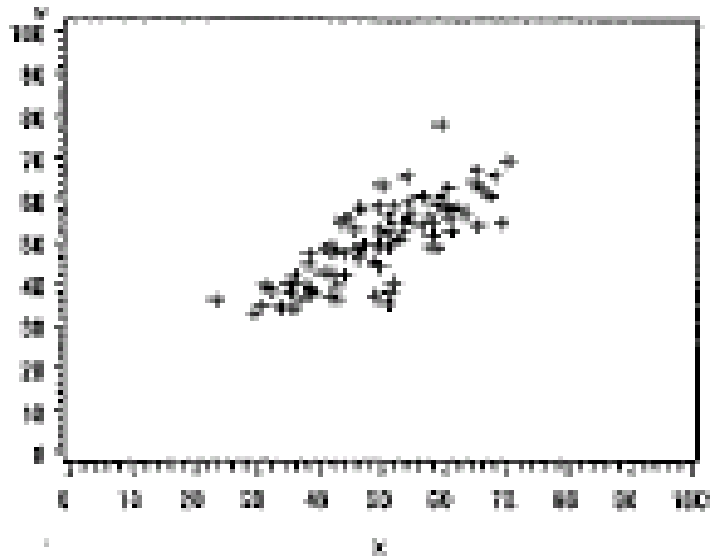
$r=0.4$



$r=0.6$

# ピアソンの相関係数と散布図②

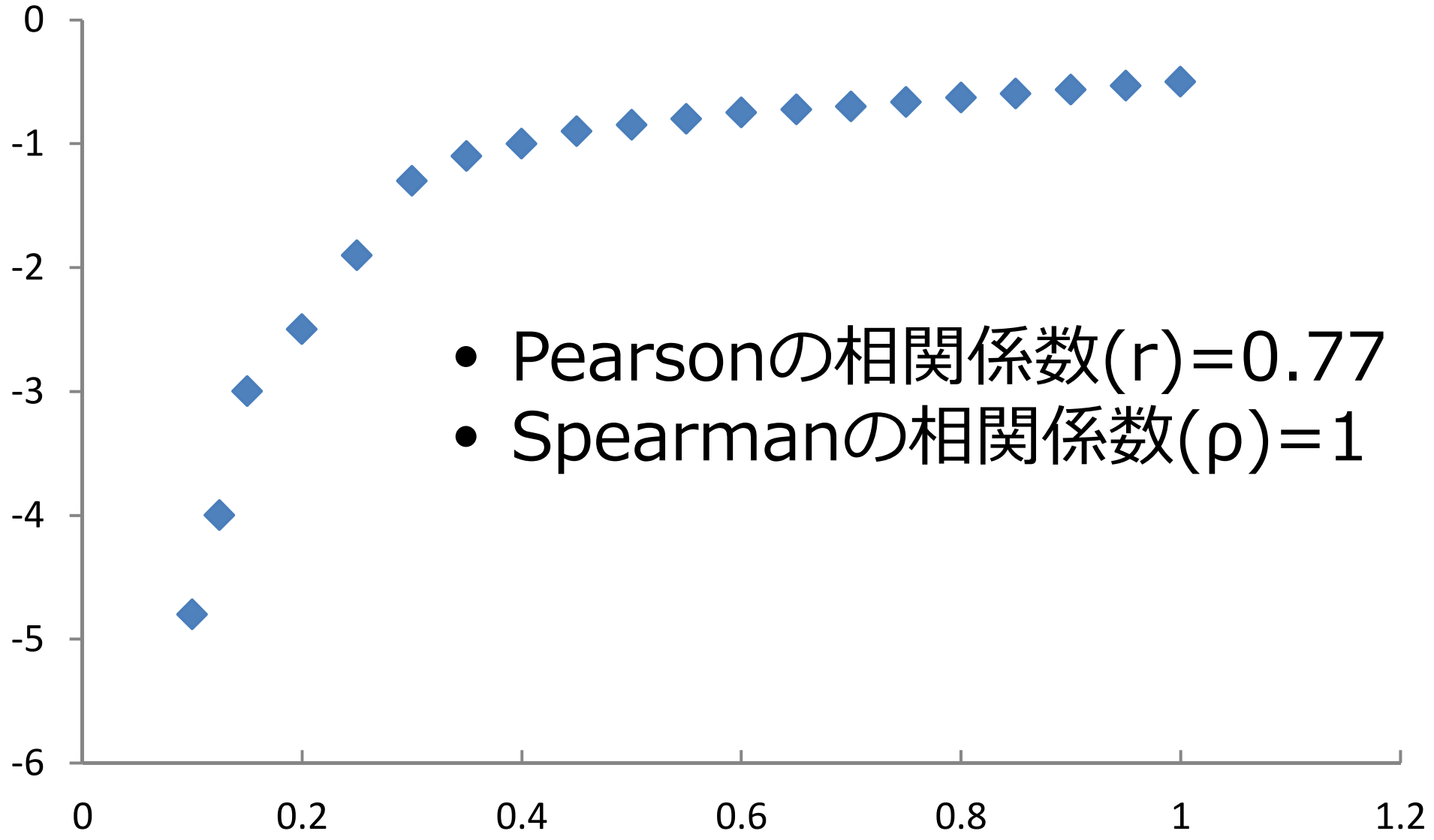
相関係数が0.8でも、相関しているようには見えない



# 相関係数に関する注意

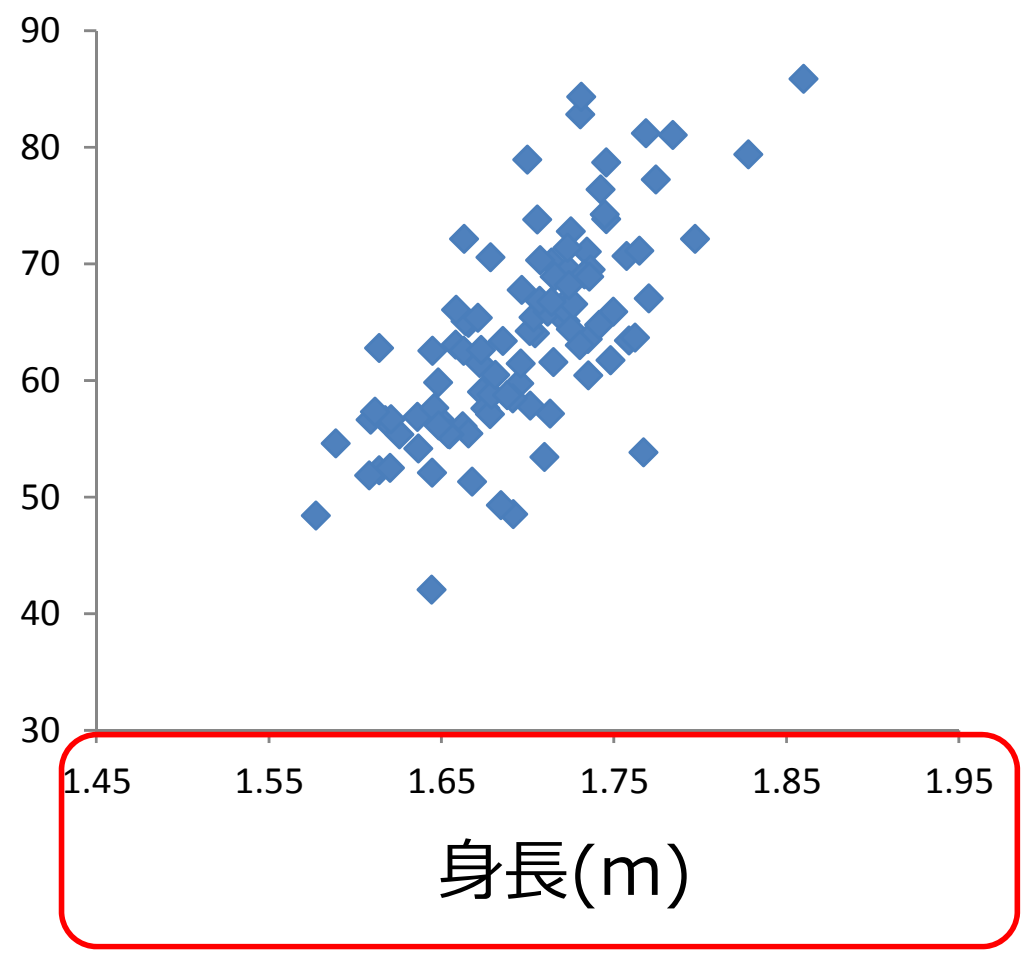
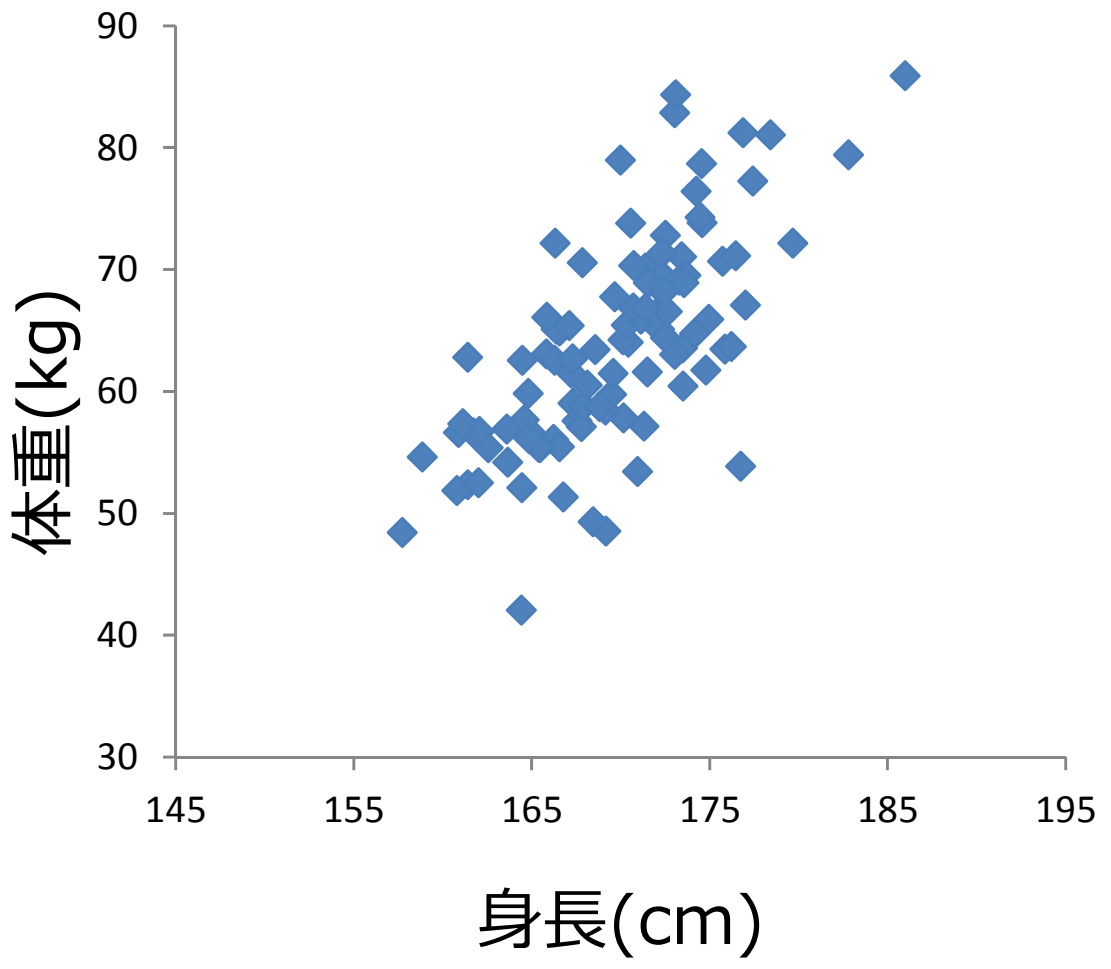
1. Pearsonの相関係数は2変数間の直線的な関連の強さを示す
2. 測定単位や見た目には依存しない
3. Pearsonの相関係数は外れ値に影響を受けやすい
4. グループの存在に注意
5. 相関係数の有意差はNに依存する

# 直線関係ではないが2変数の関連あり



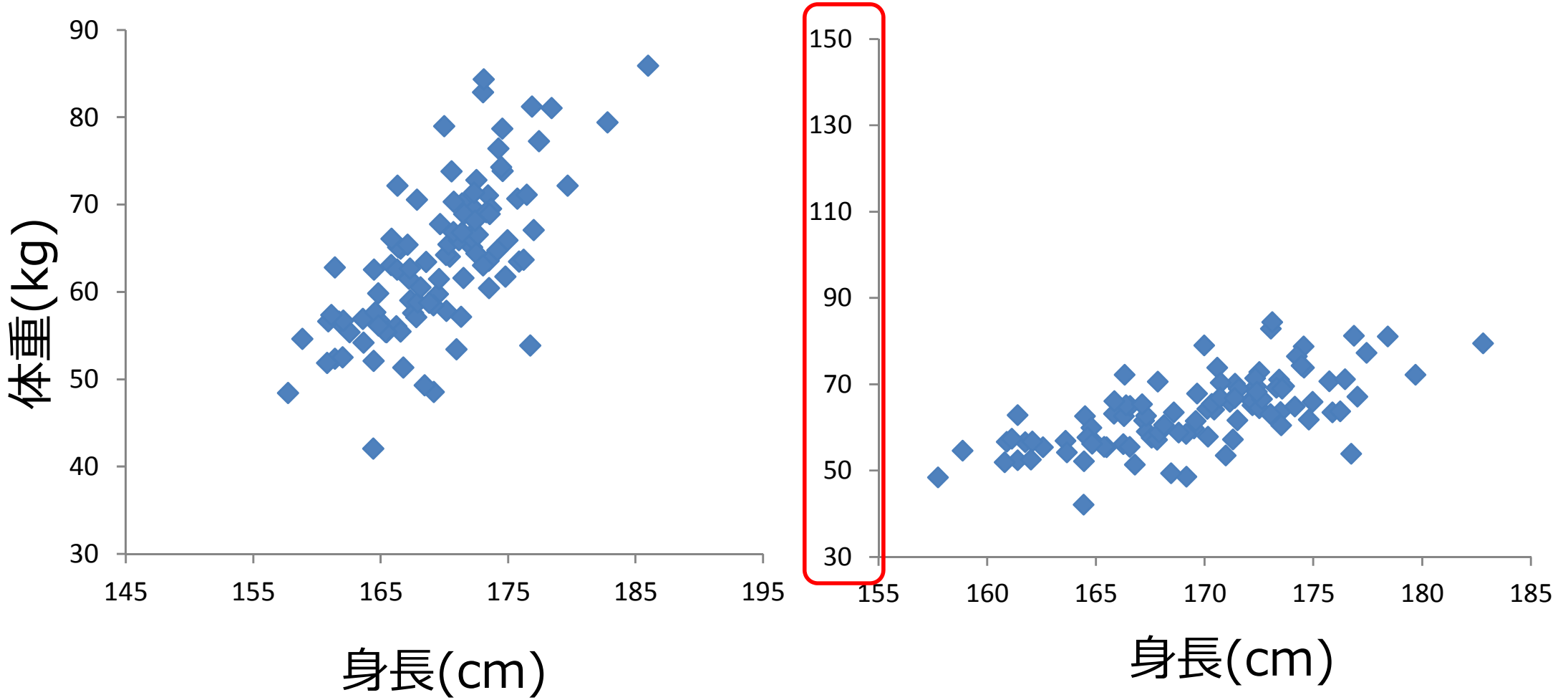
# 測定単位や見た目には依存しない①

- どちらの場合もPearsonの相関係数( $r$ ) = 0.70



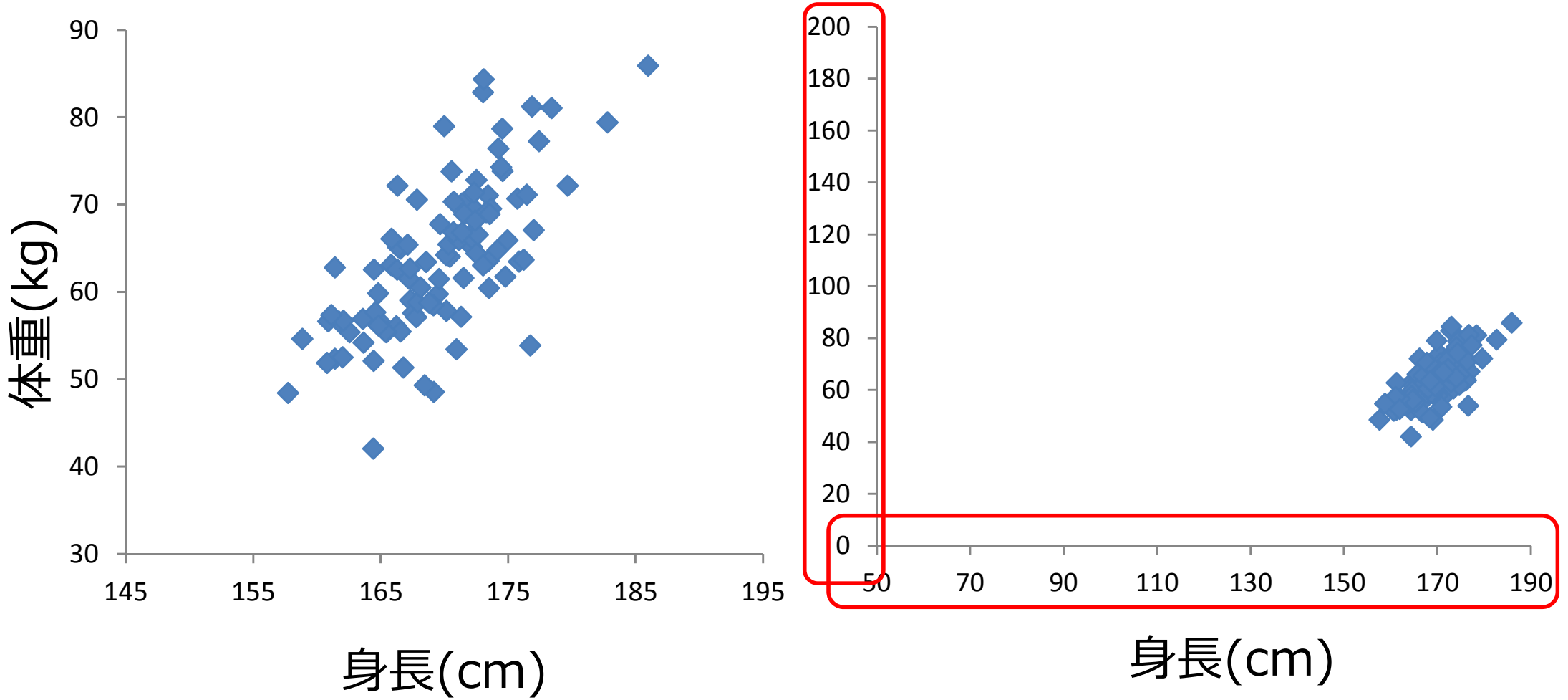
# 測定単位や見た目には依存しない②

- どちらの場合もPearsonの相関係数( $r$ ) = 0.70



# 測定単位や見た目には依存しない③

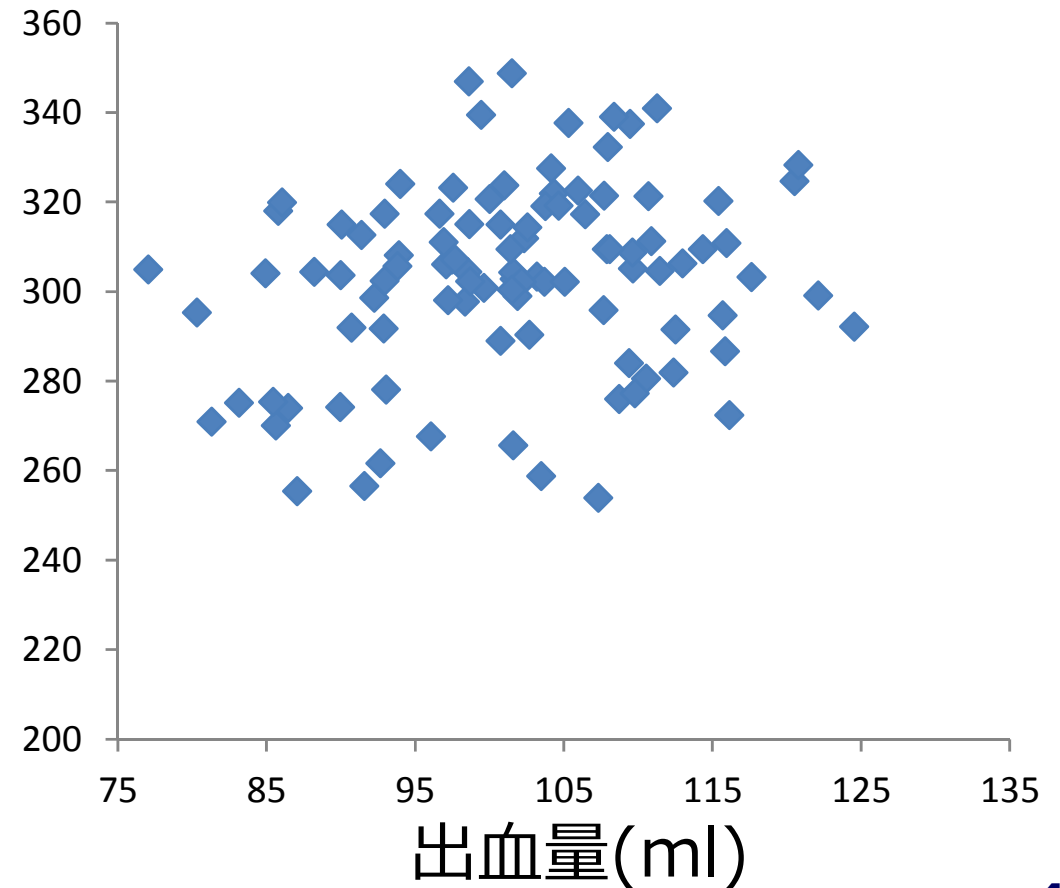
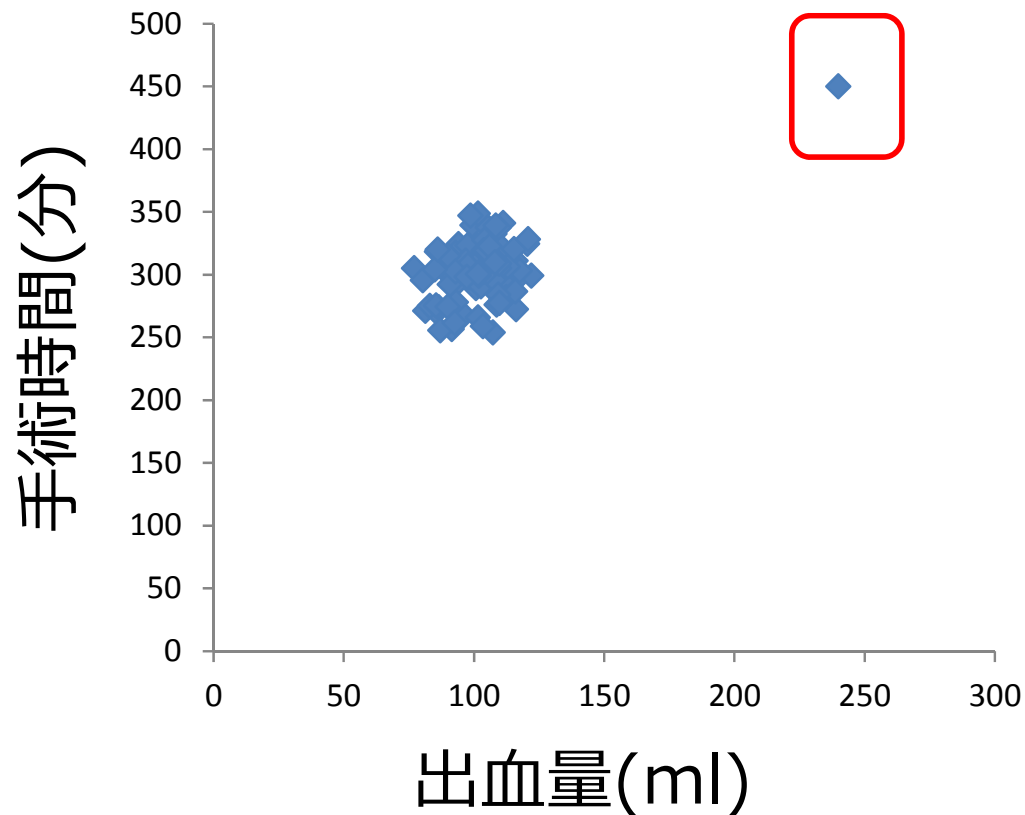
- どちらの場合もPearsonの相関係数( $r$ ) = 0.70



# Pearsonの相関係数は外れ値の影響を受けやすい

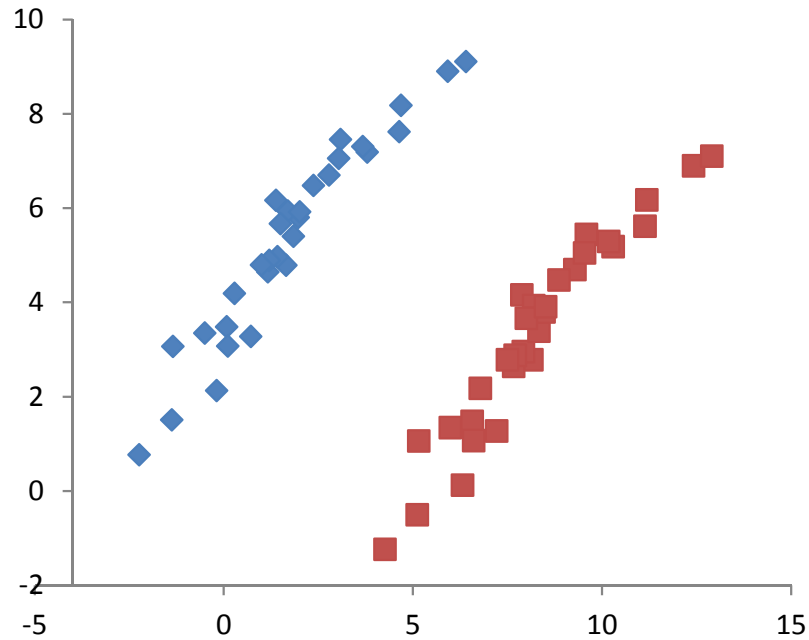
- 外れ値を含めると
  - Pearsonの相関係数( $r$ )=**0.58**
  - Spearmanの相関係数( $\rho$ )=0.23

- 外れ値を除くと
  - Pearsonの相関係数( $r$ )=**0.23**
  - Spearmanの相関係数( $\rho$ )=0.21

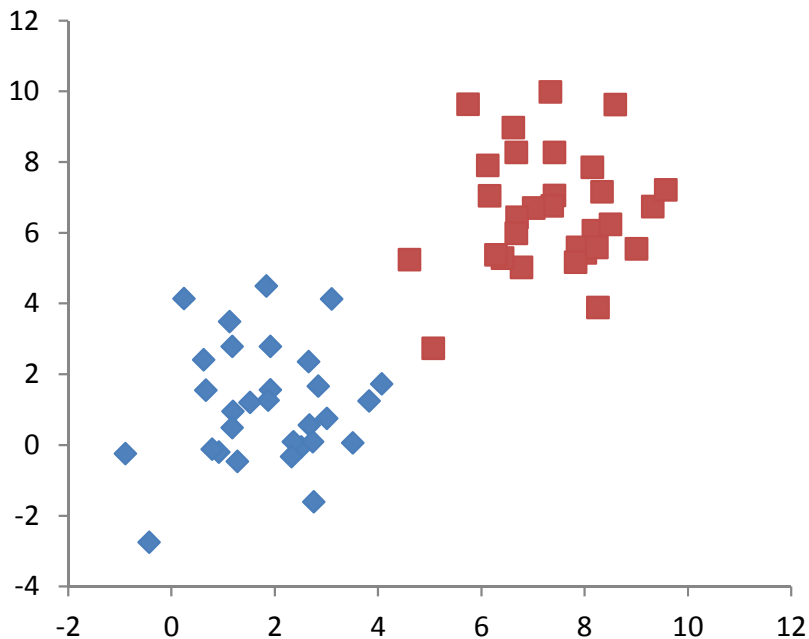




# データにグループがないかどうか確認



- グループ別のPearsonの相関係数(r)
  - 0.96と0.96
- まとめた時のPearsonの相関係数(r)
  - 0.09



- グループ別のPearsonの相関係数(r)
  - 0.1と0.1
- まとめた時のPearsonの相関係数(r)
  - 0.81

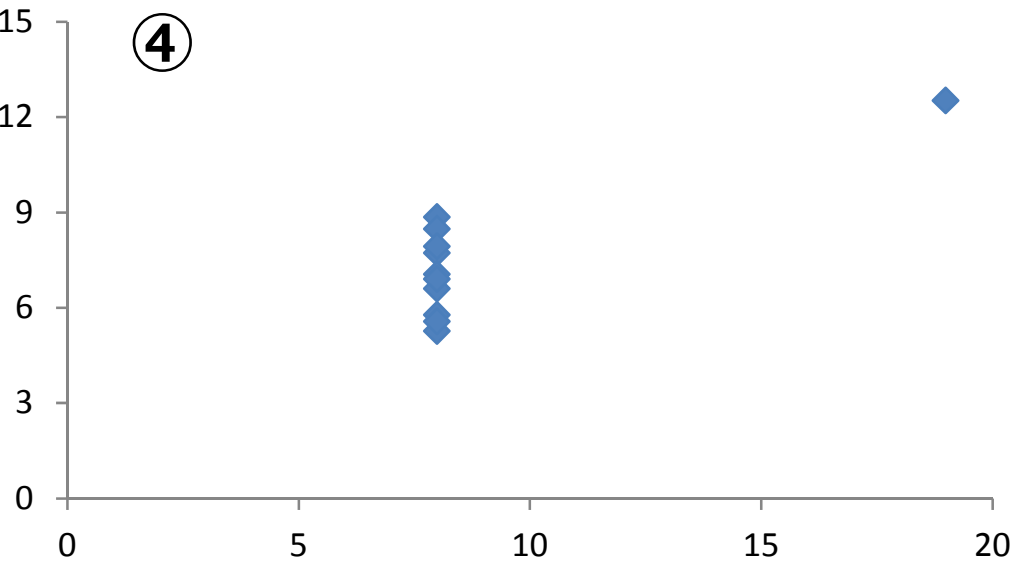
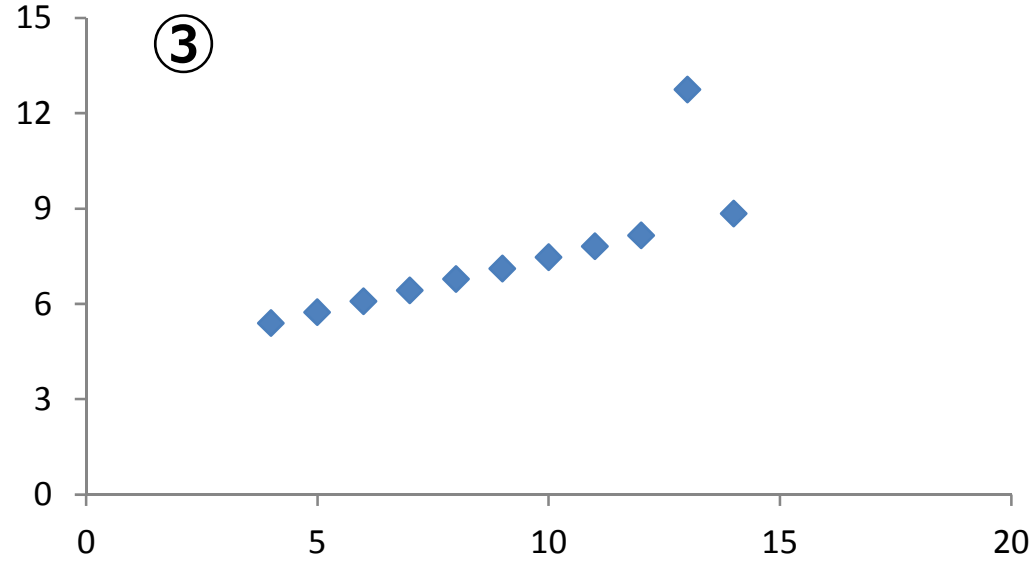
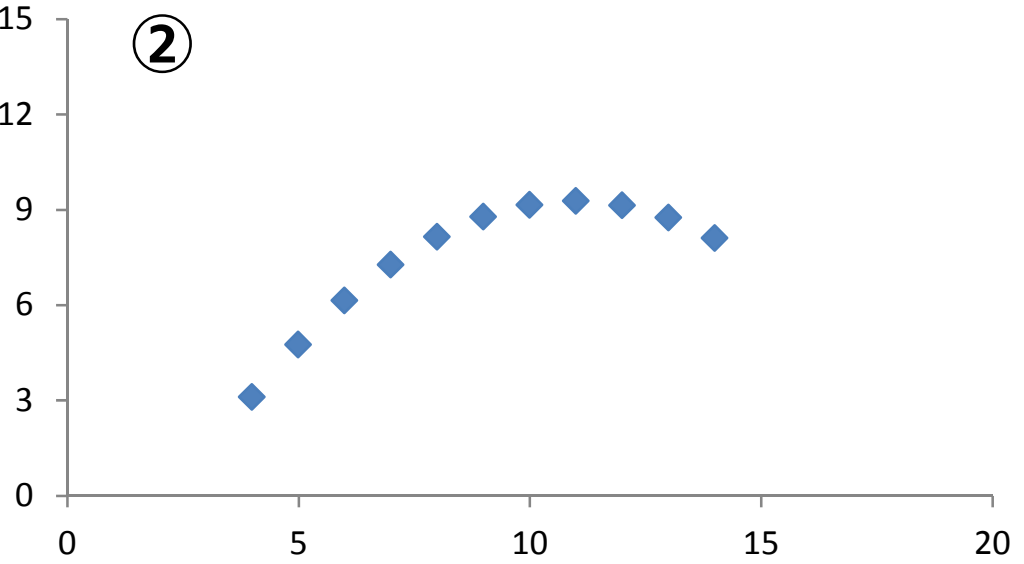
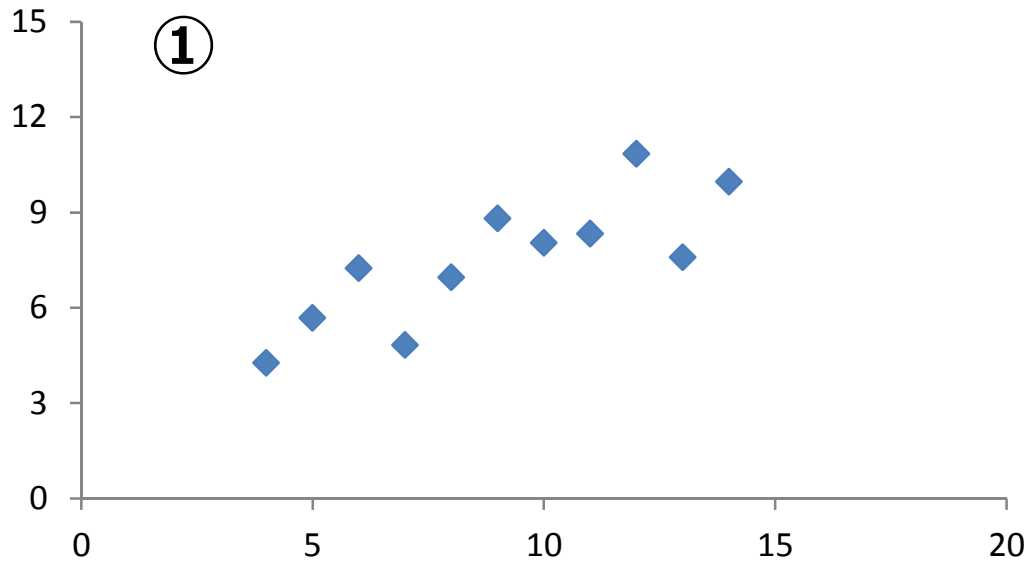
# どのパターンが相関が強そうでしょうか？

①		②		③		④	
$X_1$	$Y_1$	$X_2$	$Y_2$	$X_3$	$Y_3$	$X_4$	$Y_4$
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	<b>13</b>	<b>12.74</b>	8	6.89
14	9.96	14	8.1	14	8.84	<b>19</b>	<b>12.5</b>

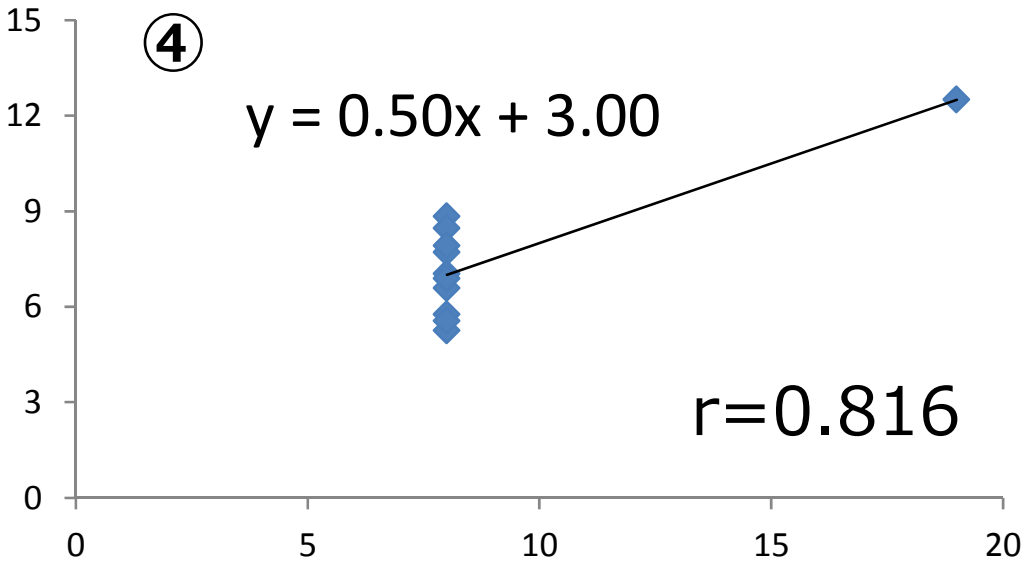
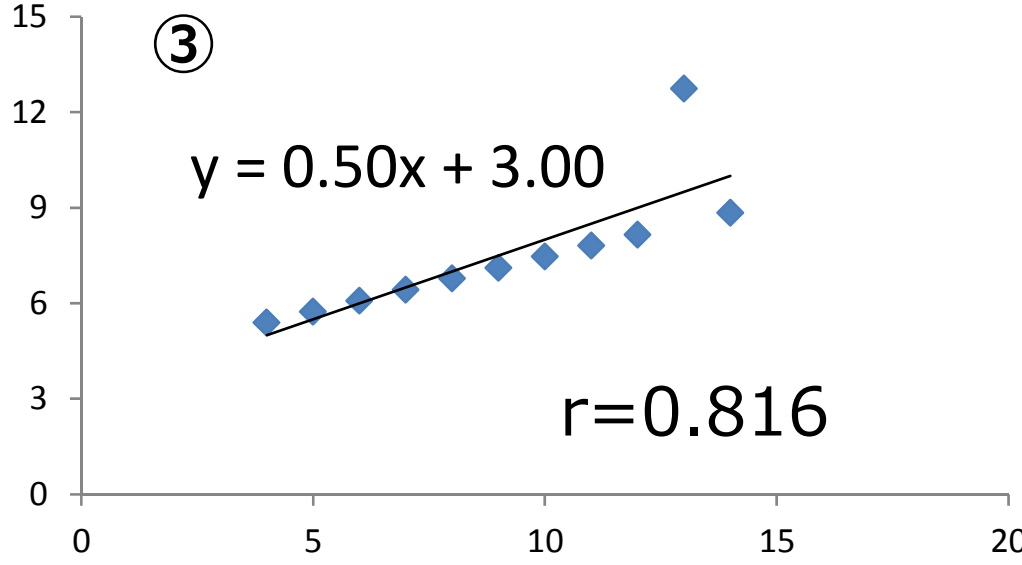
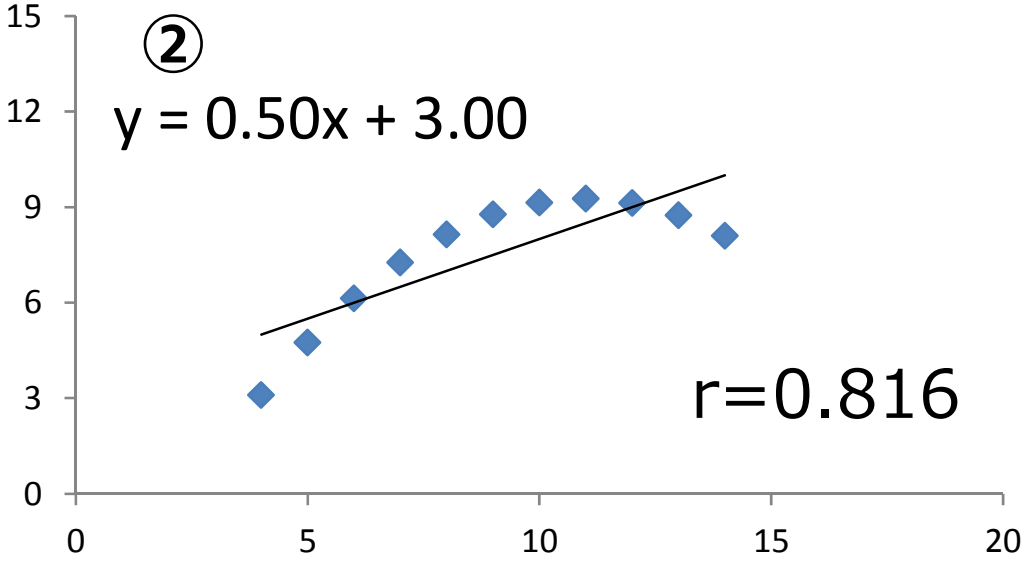
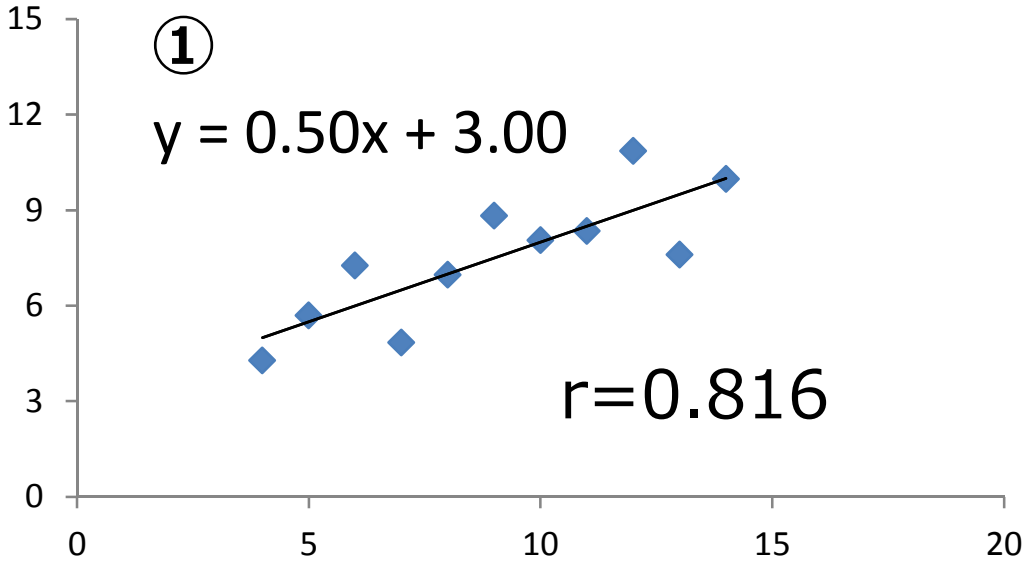
# グラフと相関係数

	N	平均値	標準偏差	相関係数 (p値)	回帰式 (YとXの関係式)
X <sub>1</sub>	11	9.00	3.32	0.816	Y <sub>1</sub> =3.00+0.50X <sub>1</sub>
Y <sub>1</sub>	11	7.50	2.03	(p=0.0022)	
X <sub>2</sub>	11	9.00	3.32	0.816	Y <sub>2</sub> =3.00+0.50X <sub>2</sub>
Y <sub>2</sub>	11	7.50	2.03	(p=0.0022)	
X <sub>3</sub>	11	9.00	3.32	0.816	Y <sub>3</sub> =3.00+0.50X <sub>3</sub>
Y <sub>3</sub>	11	7.50	2.03	(p=0.0022)	
X <sub>4</sub>	11	9.00	3.32	0.816	Y <sub>4</sub> =3.00+0.50X <sub>4</sub>
Y <sub>4</sub>	11	7.50	2.03	(p=0.0022)	

# 散布図の結果 (横軸x, 縦軸y)



# 要約統計量だけで判断せずグラフ化が大切！



## 「相関あり」と判断する規準は？

- Nが増えればp値は小さくなるので、統計的有意差で判断するのは必ずしも適切とはいえない
- 研究領域にもよるのでケースバイケースの判断

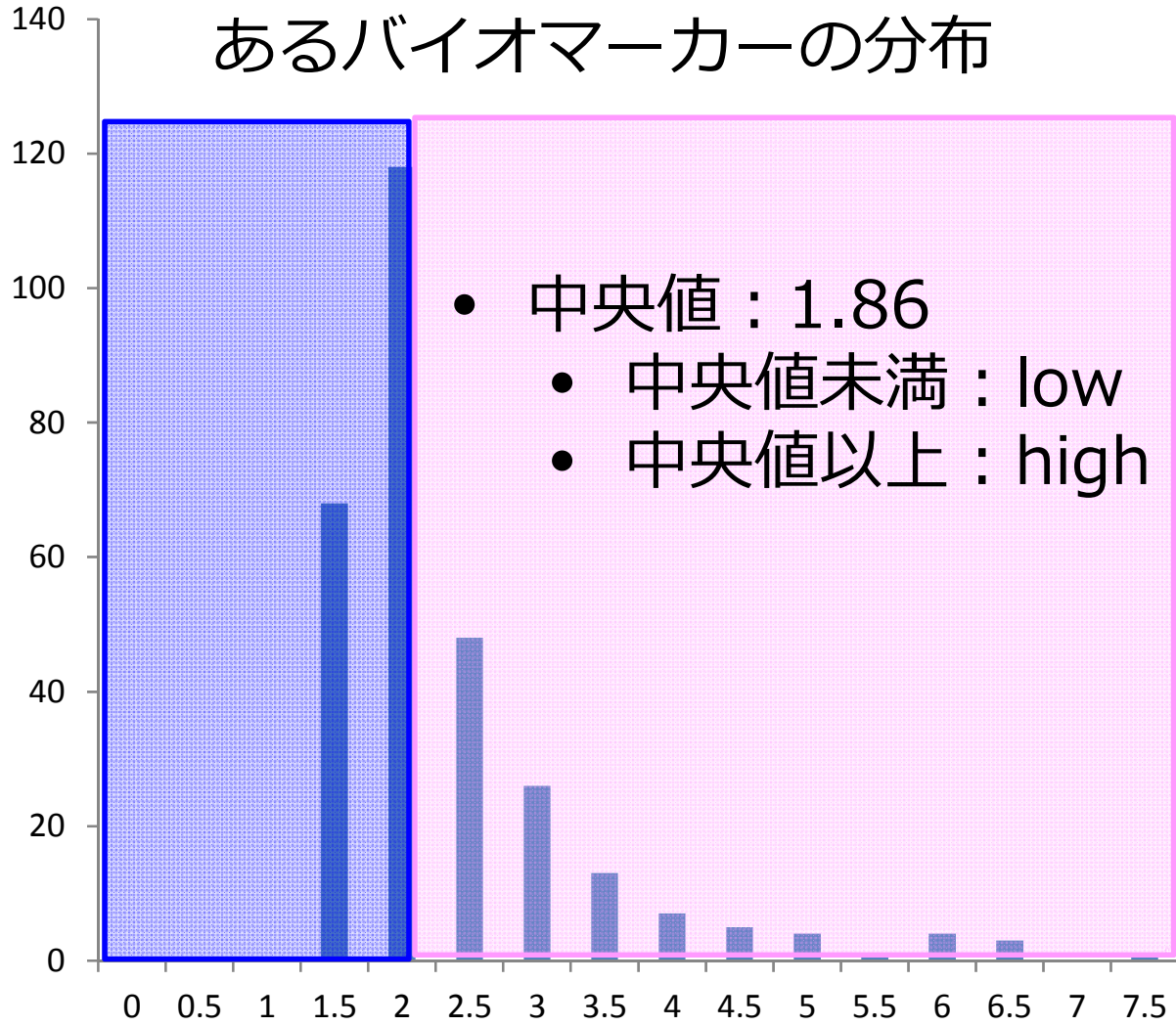
母相関係数=0の 検定に対するP値		Pearsonの相関係数(r)			
		0.1	0.3	0.5	0.7
N	5	0.870	0.615	0.374	0.165
	10	0.783	0.397	0.137	0.022
	20	0.675	0.198	0.024	0.001
	100	0.322	0.002	0.000	0.000
	400	0.045	0.000	0.000	0.000

## 2つの連続量データのまとめ方について

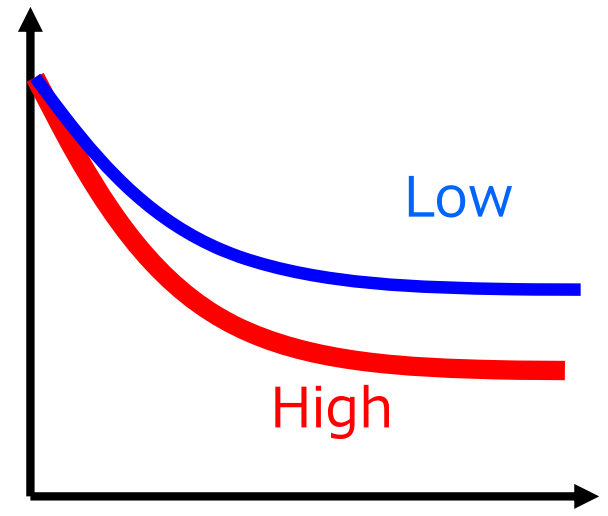
- 散布図 (scatter plot) を用いる
- 外れ値の存在や、おおよその相関関係を確認
  - 直線的 or 曲線的
- 臨床データは右に裾を引いた分布が多いので、Pearsonの相関係数よりもspearmanの相関係数の方が適切な場合が多い
- 相関があるかどうかは検定結果だけで判断しない

# 連続量データは2値で分けなければいけないか？

- バイオマーカーと予後の関係を検討する



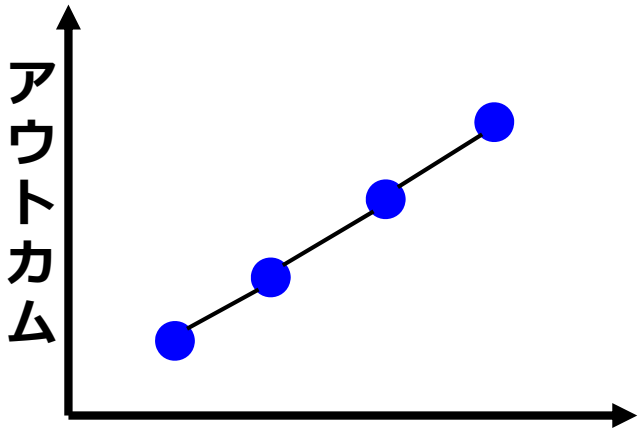
これが本当に  
確かめたかったこと？



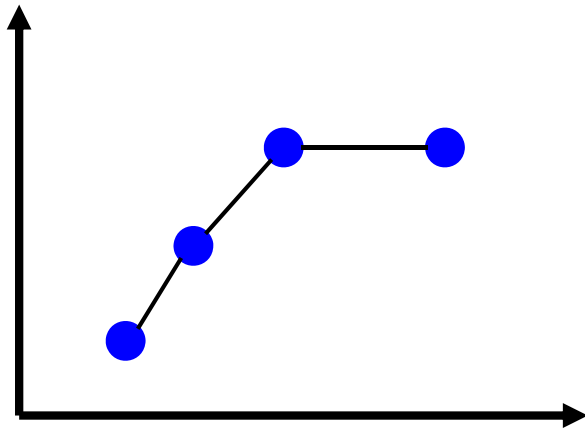


# 連続量データを2値に分ける前に

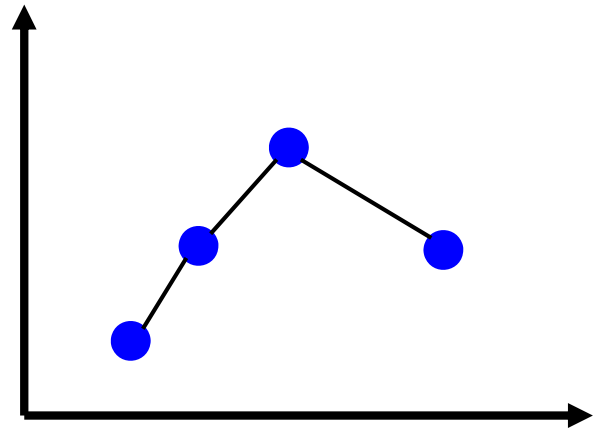
直線関係があるか？



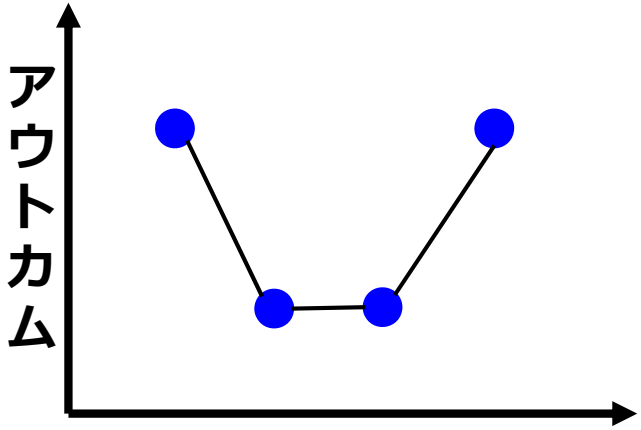
効果がプラトーになるか？



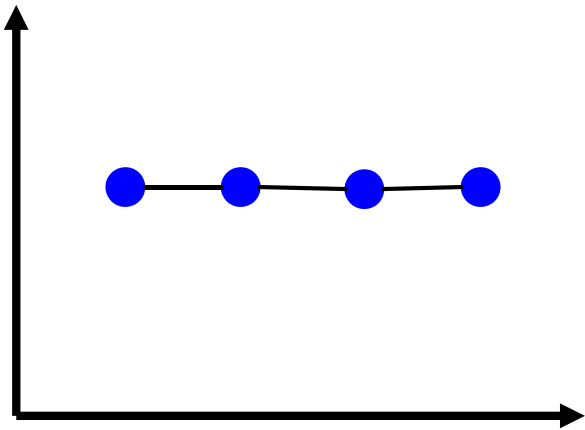
効果が弱くなるか？



バスタブ曲線？



効果に変化なし？



- 2値で分けると、どの関係なのかわからない
- 連続量の検討は3値以上に分けて行う方が適切

# カテゴリカルデータの場合

カイ2乗検定、Fisherの直接確率計算をする前に

# アウトカムの型と統計手法

	連続量	2値(0 / 1)	生存時間
アウトカムの例	血圧値、 臨床検査値	奏効割合 有害事象発生割合	全生存期間 無増悪生存期間
データの要約法	ヒストグラム・散布図	分割表	Kaplan-Meier曲線
群間比較(検定)	t検定、 Wilcoxon検定	カイ2乗検定、 Fisherの 直接確率検定	ロジスティック 検定
モデルのあてはめ	重回帰分析	ロジスティック回帰	Cox回帰

# カテゴリカルデータの確認は集計表 & 分割表！

## データセット

患者番号	T因子	N因子	M因子	病期
1	10	20	10	40
2	10	10	00	10
3	20	20	10	30
4	20	10	00	20
5	30	20	00	30
.	.	.	.	.
.	.	.	.	.

T	度数	パーセント	累積 度数	累積 パーセント
10	2	40.00	2	40.00
20	2	40.00	4	80.00
30	1	20.00	5	100.00

N	度数	パーセント	累積 度数	累積 パーセント
10	1	20.00	1	20.00
10	1	20.00	2	40.00
20	3	60.00	5	100.00

M	度数	パーセント	累積 度数	累積 パーセント
00	3	60.00	3	60.00
10	2	40.00	5	100.00

stage	度数	パーセント	累積 度数	累積 パーセント
10	1	20.00	1	20.00
20	1	20.00	2	40.00
30	2	40.00	4	80.00
40	1	20.00	5	100.00

表：M \* stage

M	stage				合計
度数	10	20	30	40	
00	1	1	1	0	3
10	0	0	1	1	2
合計	1	1	2	1	5

- エクセルの「フィルター」機能だけで行うのは非効率的& 見逃しのリスク
- 2つの変数の関係は分割表を確認しないとわからない

# 分割表分析

	A群	B群
男	XX	XX
女	XX	XX

	A群	B群
40才未満	XX	XX
40才~59才	XX	XX
60才以上	XX	XX

カイ2乗検定 or Fisher検定で

カイ2乗検定 or Fisher検定で

↓

有意じゃない

↓

有意

↓

有意じゃない

↓

有意

↓

群間での男女比に  
違いがあるとは言えない

↓

群間での男女比に  
違いがある

↓

群間での年齢分布  
に違いがあるとは言えない

↓

群間での年齢分布  
に違いがある

↓

A群の方が女性の  
割合が高い

↓

だから何？

# 分割表分析

	合併症なし	合併症あり
40才未満	15	5
40才~59才	10	10
60才以上	5	15

	合併症なし	合併症あり
40才未満	10	10
40才~59才	15	5
60才以上	5	15

Fisher検定なら両者のp値は同じ (p=0.0084)

有意な場合、解釈はいずれも有意に「どこかに差がある」 (対立仮説)

'trend test' (Mantel test etc.) なら

JMPだと  
Cochran-Armitage検定の  
サンプルスクリプトを  
ダウンロードするとできる

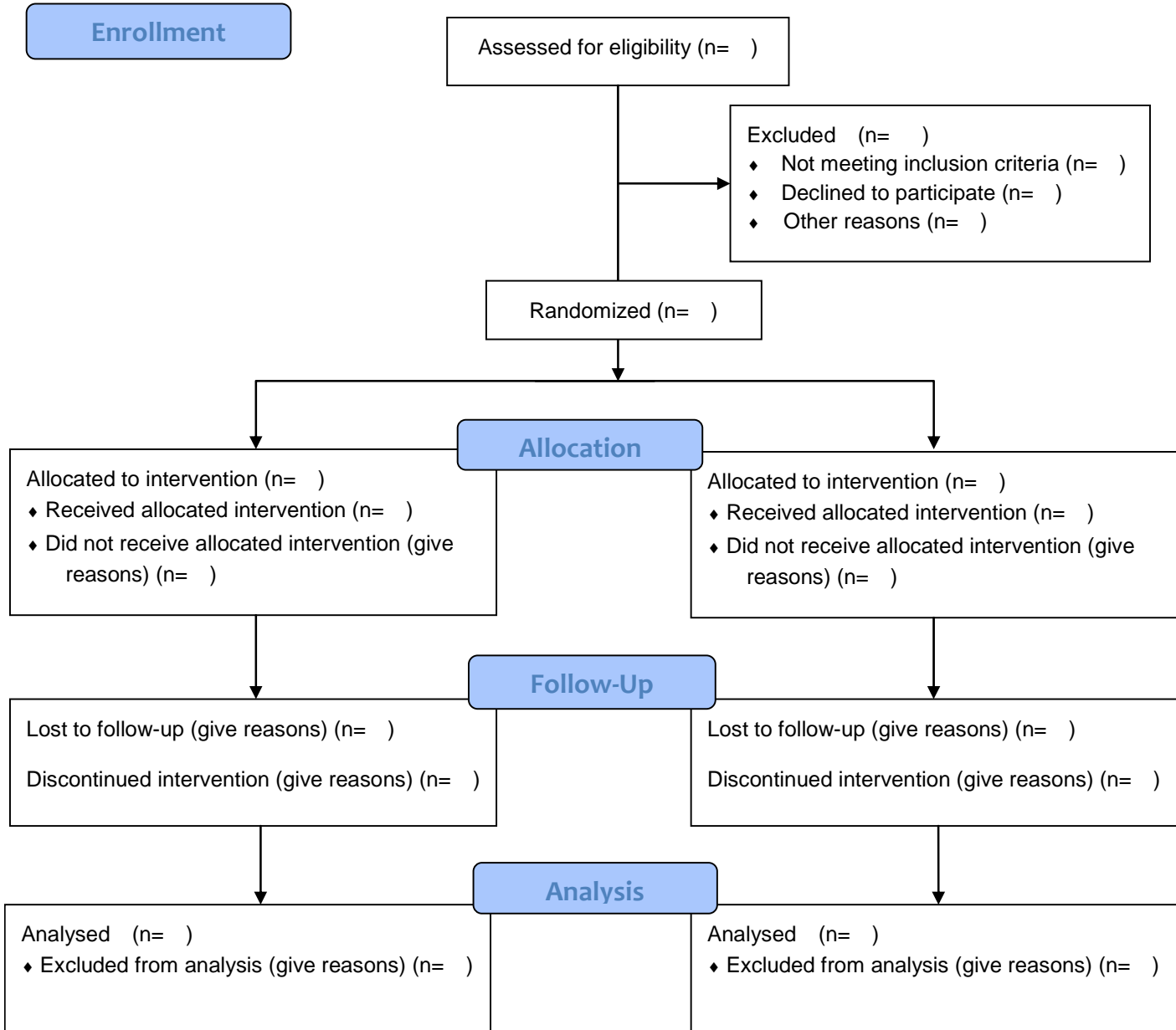
p=0.0017 : 有意  
高齢者で有意に合併症が多い傾向あり

p=0.1169 : 有意じゃない  
高齢者ほど合併症が多いとは言えない

**何を確かめたいのか整理してから検定する！**

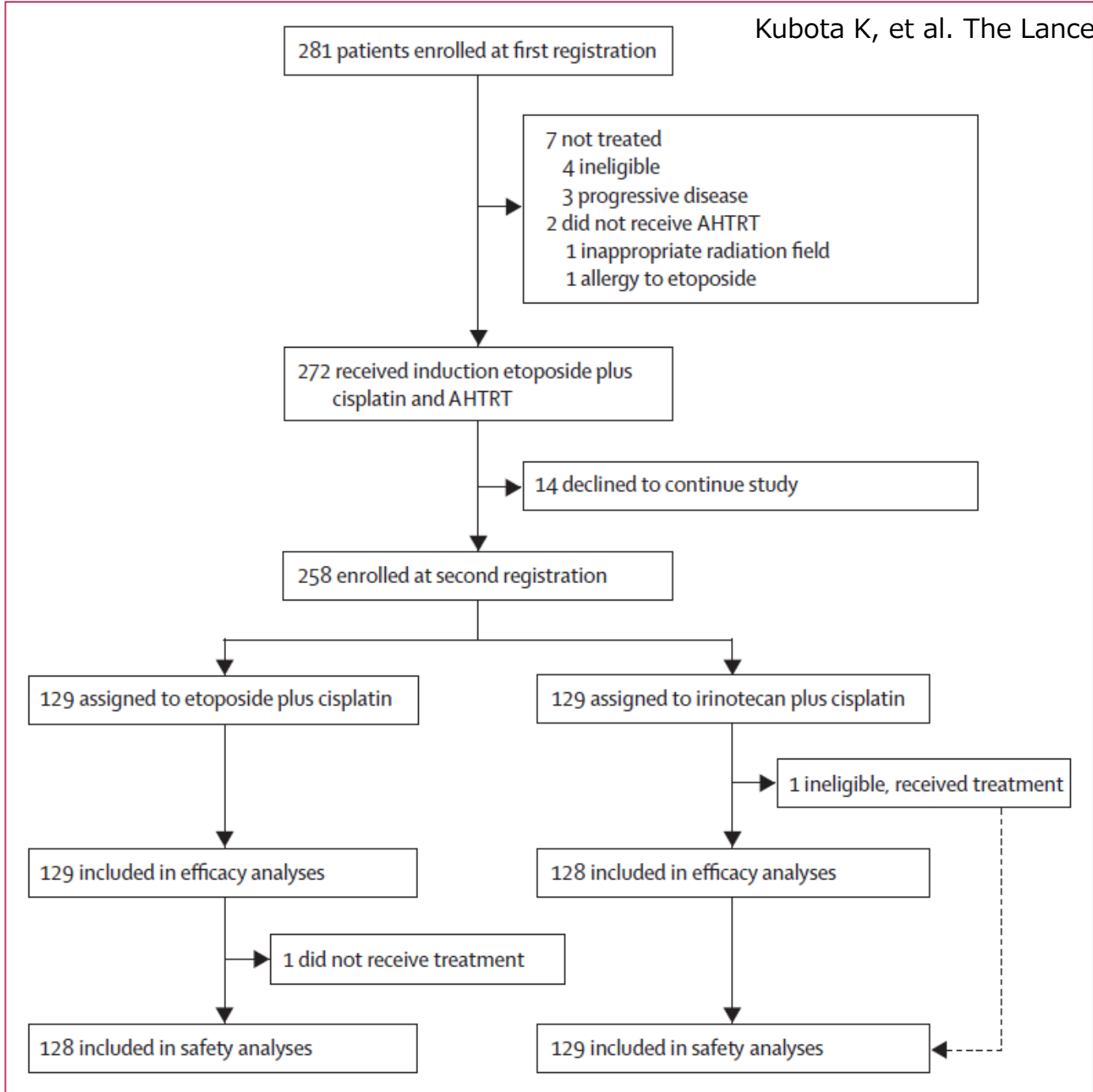
# 解析対象集団の定義と Patients flow diagram

# RCTの場合 : CONSORT flow diagram

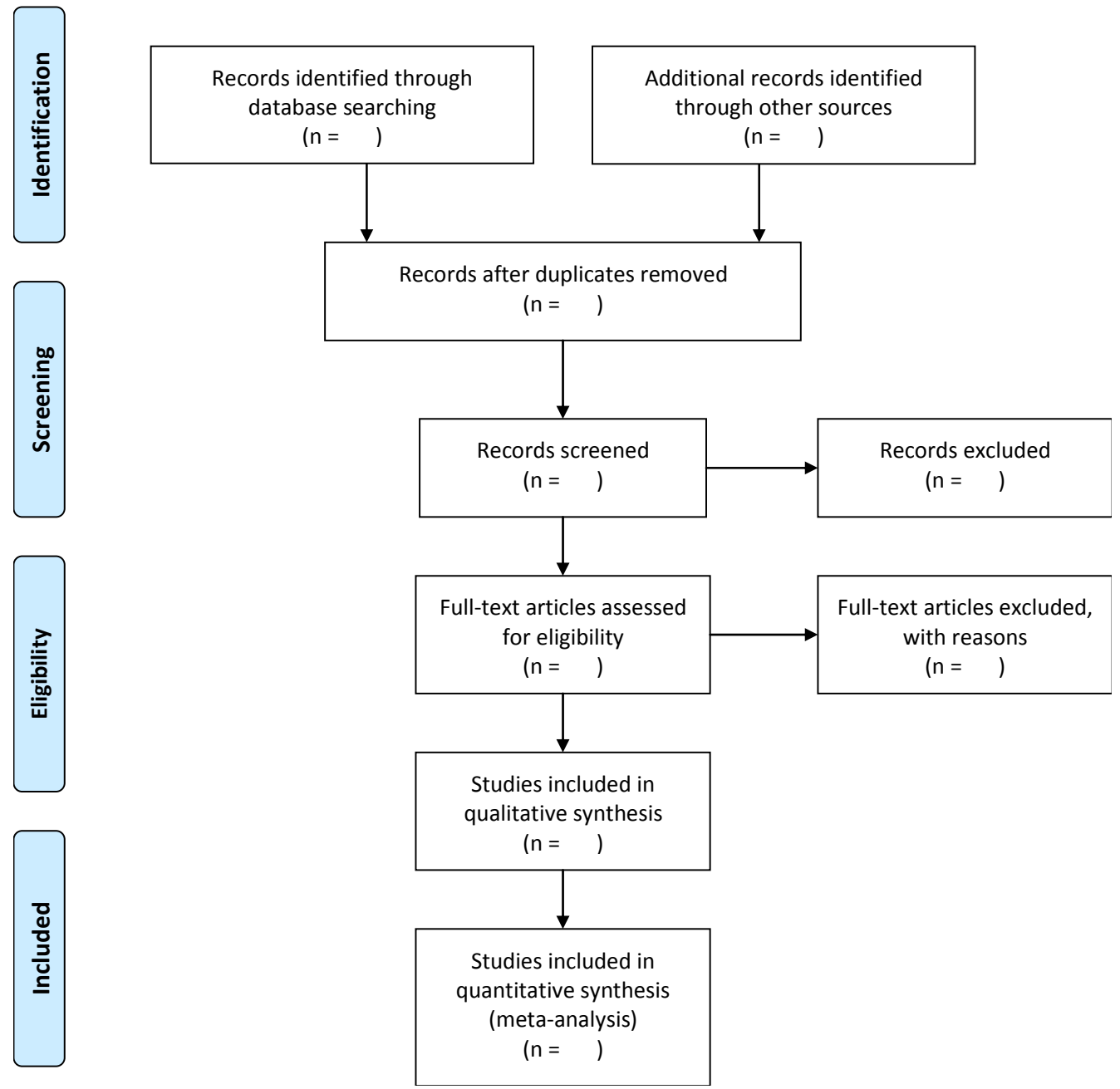


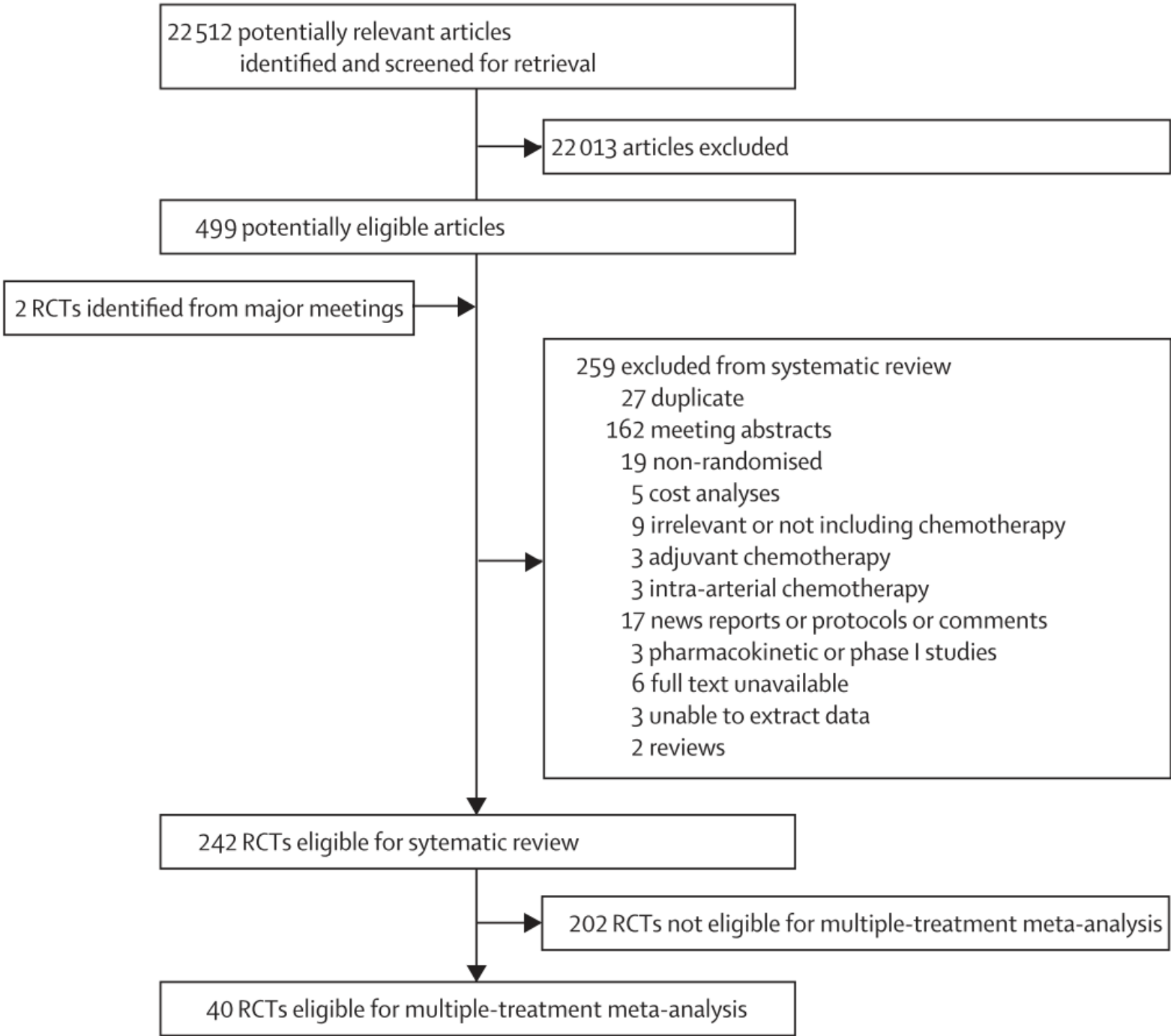


Kubota K, et al. The Lancet Oncology. 2014;15(1):106-13.

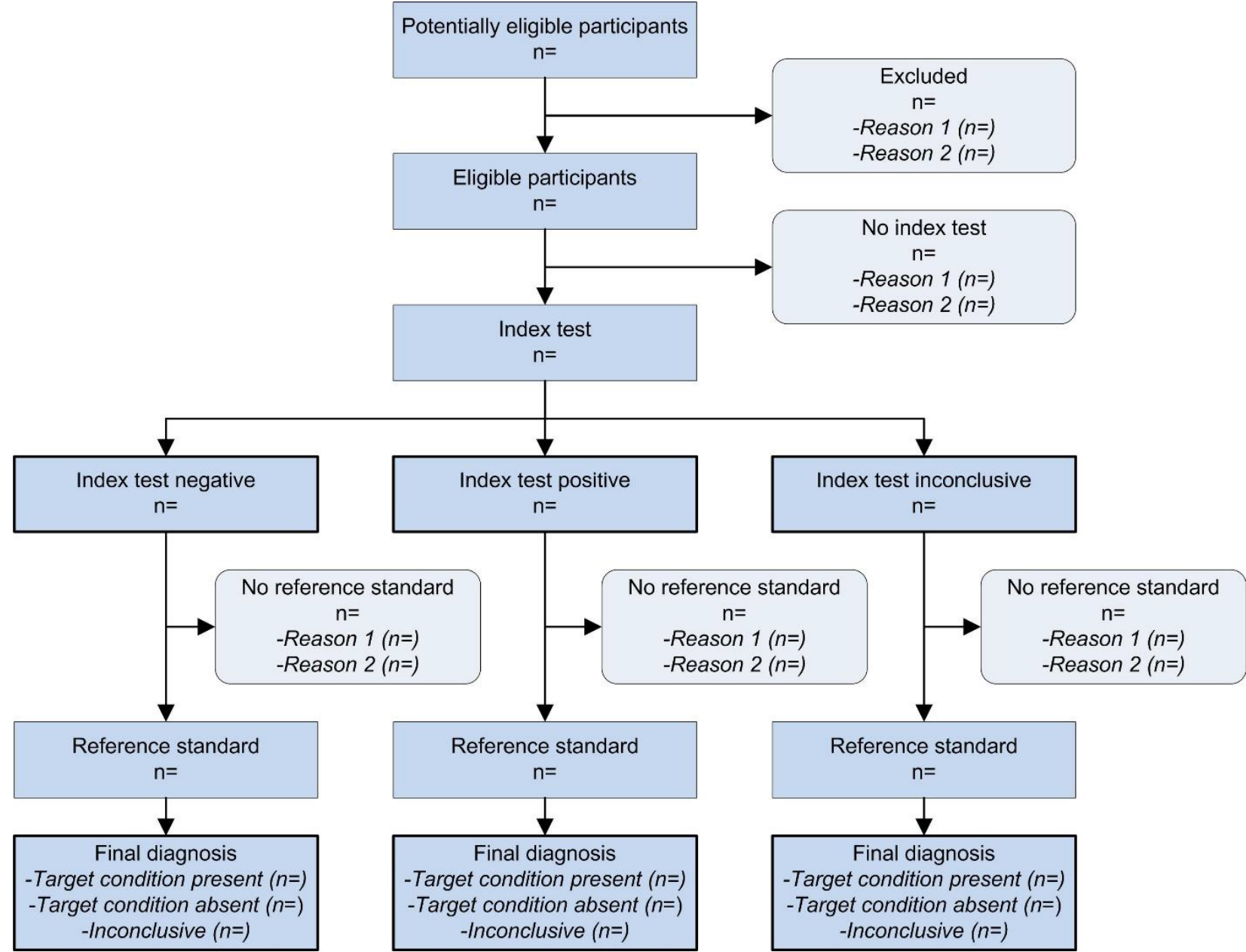


# メタアナリシスの場合 : PRISMA flow diagram





# 診断精度の研究の場合：STARD diagram



# 予後因子解析の場合

## Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)

Lisa M. McShane, Douglas G. Altman, Willi S. Massimo Gion, Gary M. Clark for the Statistic NCI-EORTC Working Group on Cancer Diagn

**Table 1.** Reporting recommendations for tumor marker prognostic studies (REMARK)

### INTRODUCTION

1. State the marker examined, the study objectives, and any prespecified hypotheses.

### MATERIALS AND METHODS

#### Patients

2. Describe the characteristics (e.g., disease stage or comorbidities) of the study patients, including their source and inclusion and exclusion criteria.
3. Describe treatments received and how chosen (e.g., randomized or rule-based).

#### Specimen characteristics

4. Describe type of biological material used (including control samples) and methods of preservation and storage.

#### Assay methods

5. Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study endpoint.

#### Study design

6. State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g., by stage of disease or age) was used. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.
7. Precisely define all clinical endpoints examined.
8. List all candidate variables initially examined or considered for inclusion in models.
9. Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.

#### Statistical analysis methods

10. Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.
11. Clarify how marker values were handled in the analyses; if relevant, describe methods used for cutpoint determination.

### RESULTS

#### Data

12. Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.
13. Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumor marker, including numbers of missing values.

#### Anal

14. Describe the flow of patients through the study,
15. including the number of patients included in each stage
16. of the analysis (a diagram may be helpful) and reasons
17. for dropout.
- 18.

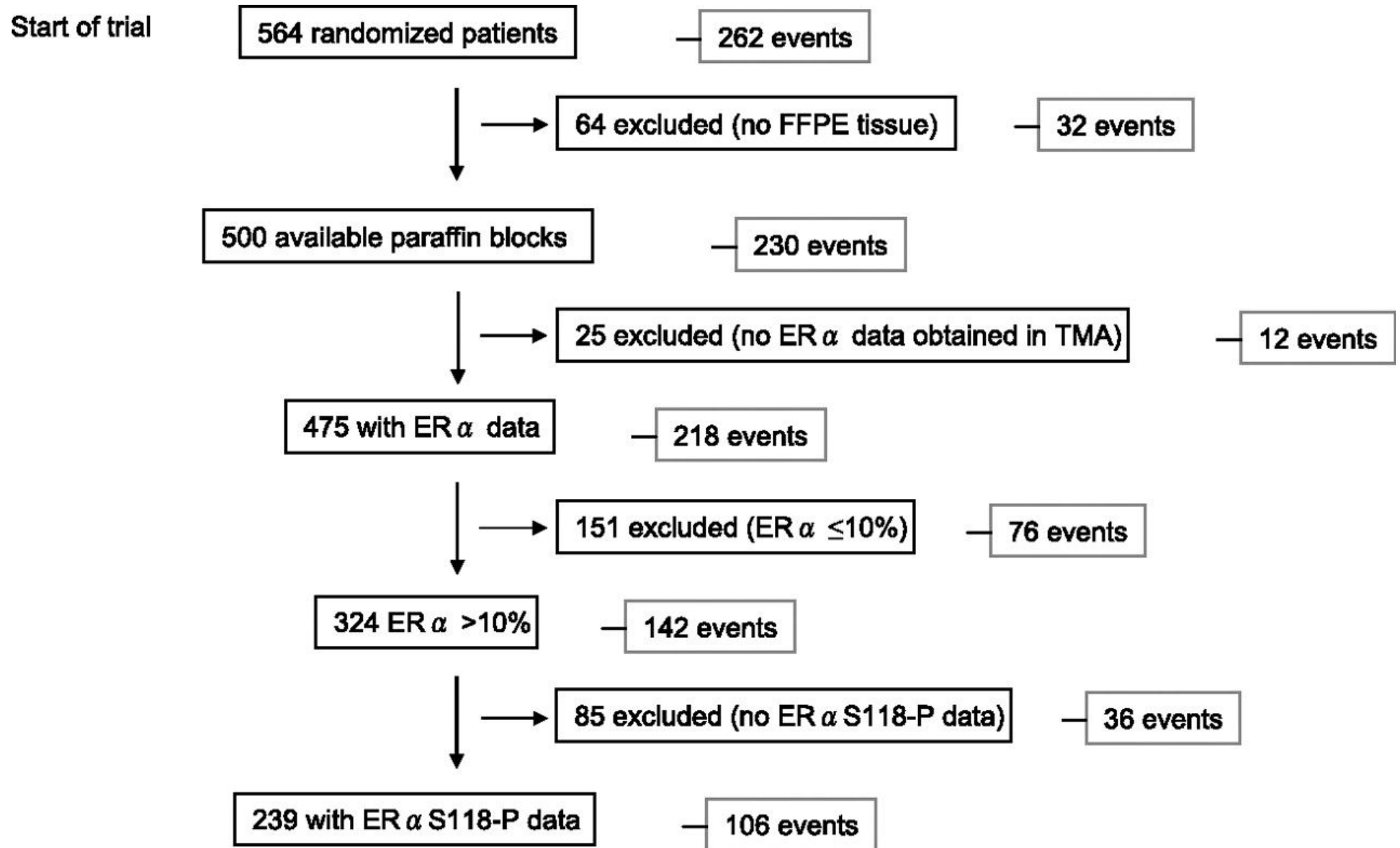
### DISCUSSION

19. Interpret the results in the context of the prespecified hypotheses and other relevant studies; include a discussion of limitations of the study.
20. Discuss implications for future research and clinical value.

Despite years of research and hundreds of reports on markers in oncology, the number of markers that emerged as clinically useful is pitifully small. Often, in reported studies of a marker show great promise, but subsequent studies on the same or related markers yield inconsistent conclusions or stand in direct contradiction to the promising results. It is imperative that we attempt to understand the reasons that multiple studies of the same marker lead to differing conclusions. A variety of methodologic problems have been cited to explain these discrepancies. Unnecessarily, many tumor marker studies have not been reported in a rigorous fashion, and published articles often lack sufficient information to allow adequate assessment of the quality of the study or the generalizability of study results. The development of guidelines for the reporting of tumor marker studies is a major recommendation of the National Cancer Institute-European Organisation for Research and Treatment of Cancer (NCI-EORTC) First International Meeting on Cancer Diagnostics in 2000. As for the successful CONSORT statement for randomized trials and for the STARD statement for diagnostic studies, we suggest guidelines to provide relevant information about the study design, preplanned hypothesis, patient and specimen characteristics, assay methods, and statistical analysis methods. In addition, the guideline suggests helpful presentations of data and important elements to include in discussions. The goal of these guidelines is to encourage transparent and complete reporting so that the relevant information will be available to others to help them to assess the usefulness of the data and understand the context in which the conclusions apply. [*J Natl Cancer Inst* 2005;97:1180-

Despite years of research and hundreds of reports on markers in oncology, the number of markers that have emerged as clinically useful is pitifully small (1-3). Often, initially reported studies of a marker show great promise, but subsequent studies on the same or related markers yield inconsistent conclusions or stand in direct contradiction to the promising results. It is imperative that we attempt to understand the reasons that multiple studies of the same marker lead to differing conclusions. A variety of problems have been cited to explain these discrepancies, such as general methodologic differences, poor study design, assays that are not standardized or lack reproducibility, and inappropriate or misleading statistical analyses that are based on sample sizes too small to draw meaningful conclusions (4-11). For example, in retrospective studies, patient populations are often biased toward patients with available tumor specimens

### Flow diagram of patients.



# 解析対象集団は計画段階で規定するのがベスト

- カルテを調べなければいけない対象・データが明確になる
  - 何と比較したいのか？
  - 比較のために必要なデータは存在するか？
  - 区別するためのフラグ作成
  - 研究のlimitationの明確化
- 解析対象が1例でも変わるとすべての解析をやり直し

## まとめ

- 解析する以前にデータ入力がきちんとされているか確認する
  - ダブルエントリー or 読み合わせ
- 最初から検定するのではなく、データの記述を行う
  - ヒストグラムを確認して適切な要約統計量・検定法を選択
- 散布図や分割表を用いて2変数間の関係を確認する
  - 外れ値や本来は存在するはずのないデータの確認
- 解析対象集団の定義を意識した研究を
  - どの対象にどの解析を行うか意識すること
  - 必ずpatients flow diagramを描いてみる



## 参考文献

- 大橋靖雄.医療とコンピュータ Vol.2, No.1, 1989.
- 丹後俊郎.統計学のセンス.朝倉書店. 2007.
- 永田靖. 入門 統計解析法. 日科技連. 2004.

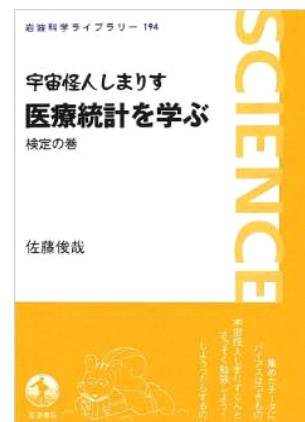
# ※ 生物統計セミナー【入門編】の推薦図書



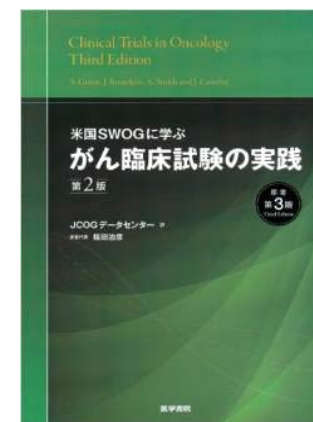
浜田知久馬. 2012  
真興交易医書出版部 ; 新版  
ISBN-13: 978-4880038612



佐藤俊哉. 2005  
岩波書店  
ISBN-13: 978-4000074544



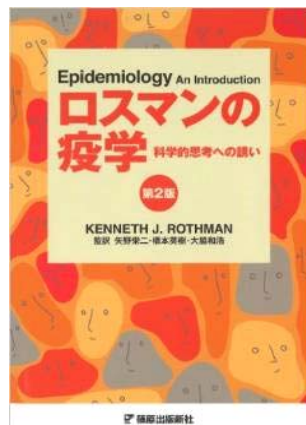
佐藤俊哉. 2012  
岩波書店  
ISBN-13: 978-4000295949



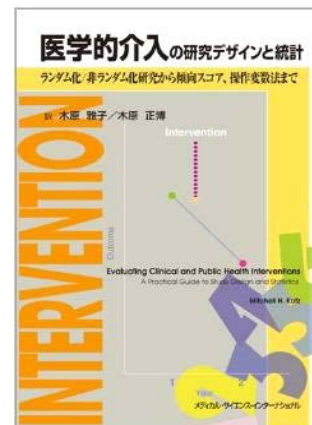
福田治彦 (訳) . 2013  
医学書院; 第2版  
ISBN-13: 978-4260018647



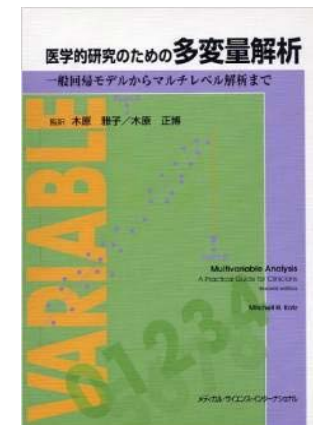
大橋靖雄. 2011  
医歯薬出版  
ASIN: B00I399SY4



矢野 栄二 (訳) . 2013  
篠原出版新社 ; 第2版  
ISBN-13: 978-4884123727



木原雅子 (訳) . 2013  
メディカルサイエンスインターナショナル  
ISBN-13: 978-4895927574



木原雅子 (訳) . 2008  
メディカルサイエンスインターナショナル  
ISBN-13: 978-4895927574