統計学(基礎)

第6回
量的変数の順序化と度数分布表・ヒストグラム

量的変数の順序化とヒストグラム

保健統計におけるデータの種類(再)

- 名義尺度(カテゴリーデータ)
  - 順番に並べることに意味がないもの
- 順序尺度(順序データ)
  - 順番になっているもの
- 量的変数(数量データ)
  - 一定間隔のもの　統計値の計算が可能

量的変数の順序尺度化

- 量的変数を、等間隔の区間に区切って(擬似的に)順序尺度化する
  - 度数分布表が作成できる
  - ヒストグラムが作成できる
- その方がわかりやすい

量的変数の順序尺度化

- 右のデータ
  - 平均値　72.4
  - 標準偏差　12.7
  - 最大値　100
  - 中央値　73.5
  - 最小値　43

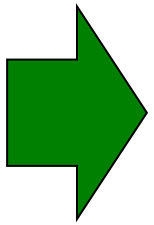
- って言われても何かよくわからない

74	65	75	73	72
63	74	72	56	78
76	91	62	68	85
56	80	56	86	65
71	87	48	79	68
43	100	80	78	90

量的変数の順序尺度化

- 度数分布表にしてみる
  - 区間毎に度数を数えて度数分布表に
  - どれくらいのところにデータがあるかがわかりやすい

74	65	75	73	72
63	74	72	56	78
76	91	62	68	85
56	80	56	86	65
71	87	48	79	68
43	100	80	78	90

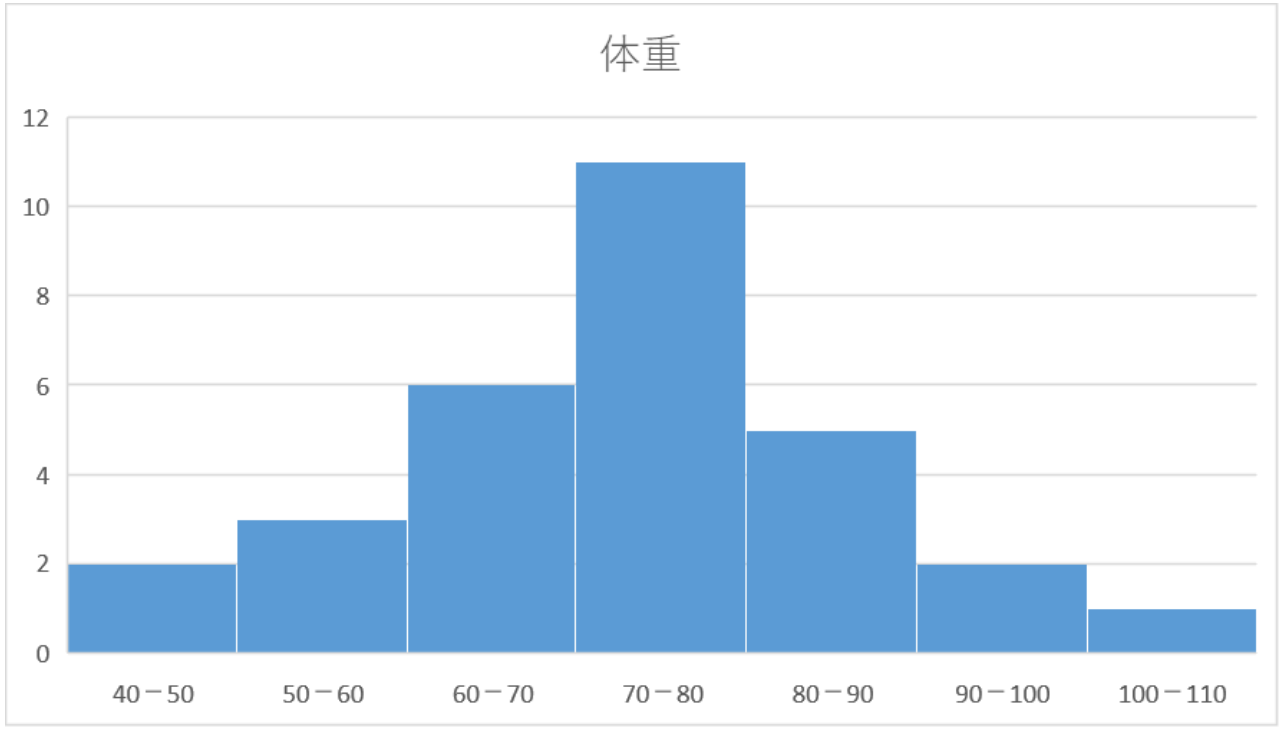


区間 (以上–未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		

ヒストグラム

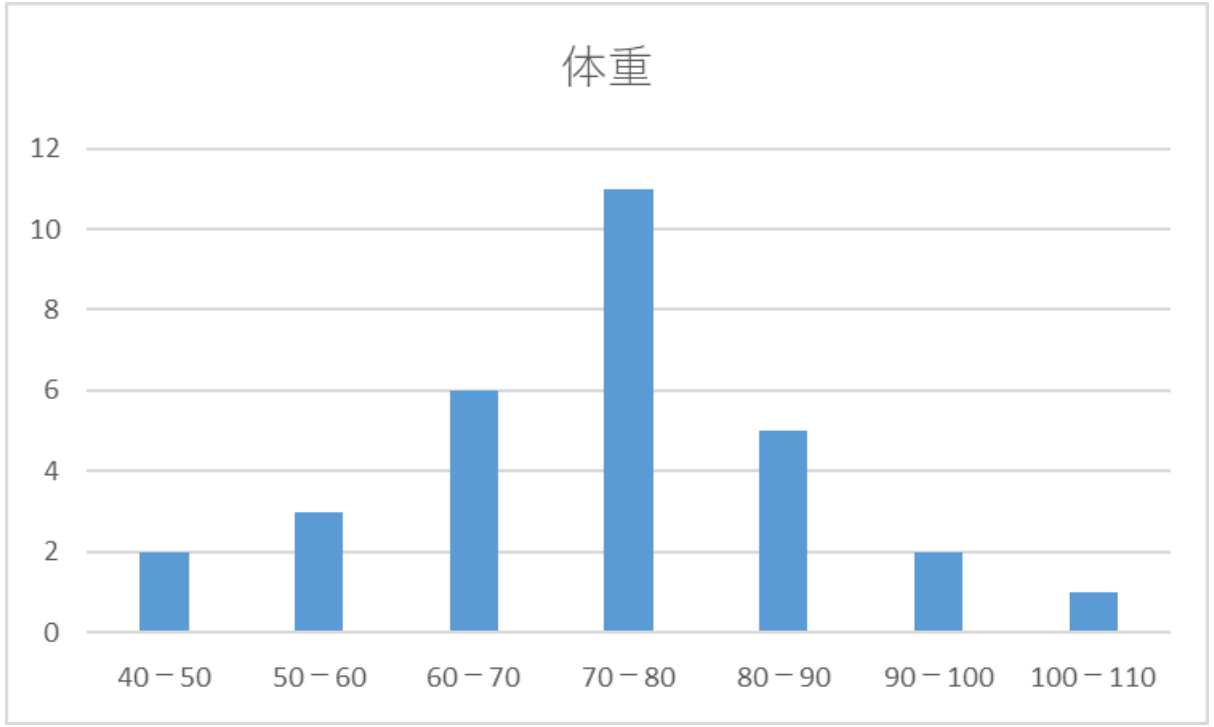
- 度数分布表をグラフにしたのがヒストグラム

区間 (以上−未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		

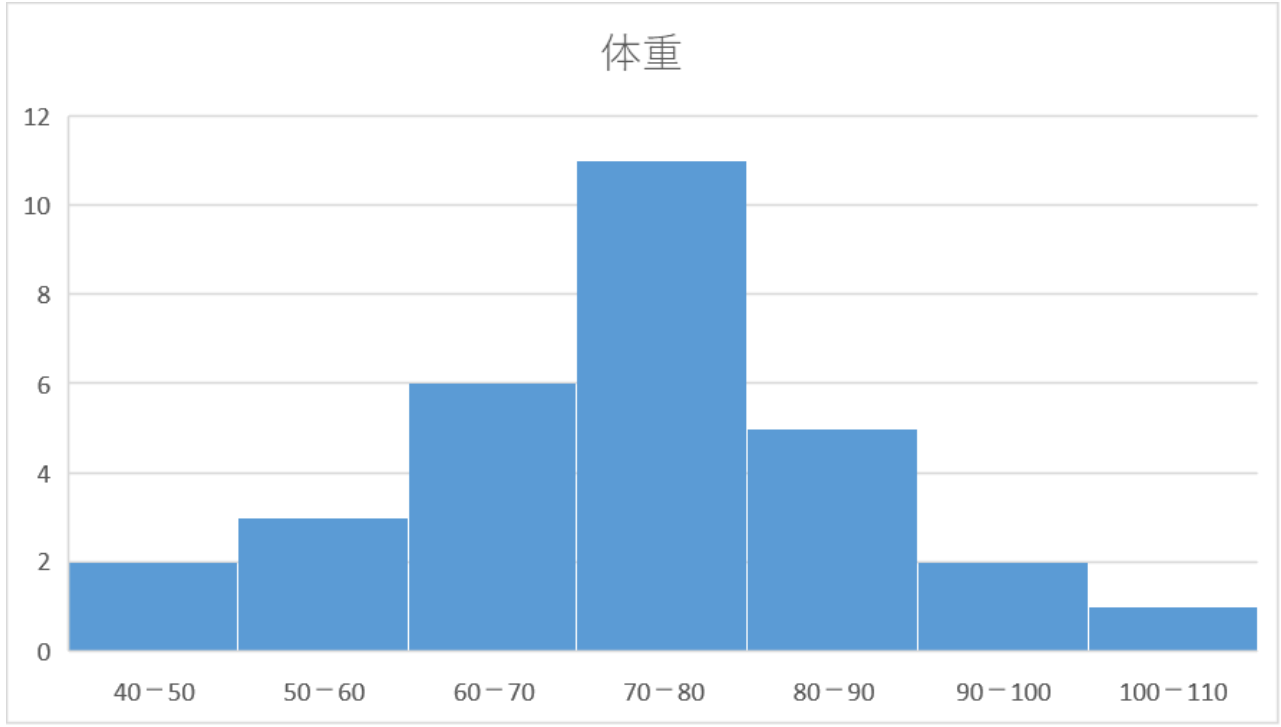


ヒストグラムと棒グラフは違う

棒グラフ

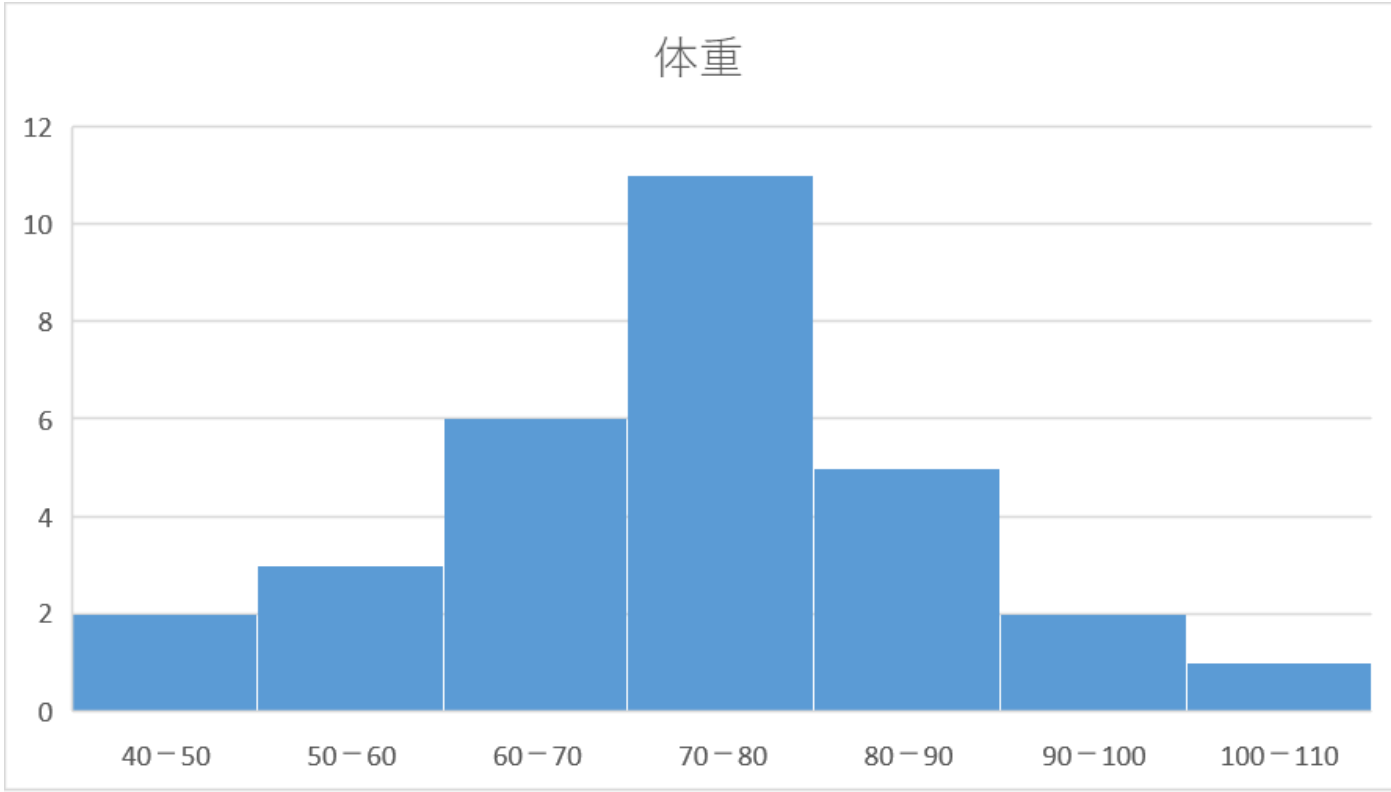


ヒストグラム



ヒストグラム

- 一言で言えば、順序尺度の度数分布表のグラフ
- 量的データを順序尺度化して作成する
- 離散量の場合はそのままで作成することもある
- 多角形の内部の面積を1と考え、累積の割合を表現する
  - 確率密度の近似を表す
  - くっついているのが大事



ヒストグラム作成時の度数分布表

- 区間
- 集計する値の範囲
- 階級値
  - 区間の中央値
  - 今は作らないことが多い
- 度数
- 相対度数
- 累積度数
- 累積相対度数

区間 (以上–未満)	階級値	度数	相対度数	累積度数	累積相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		

ヒストグラム

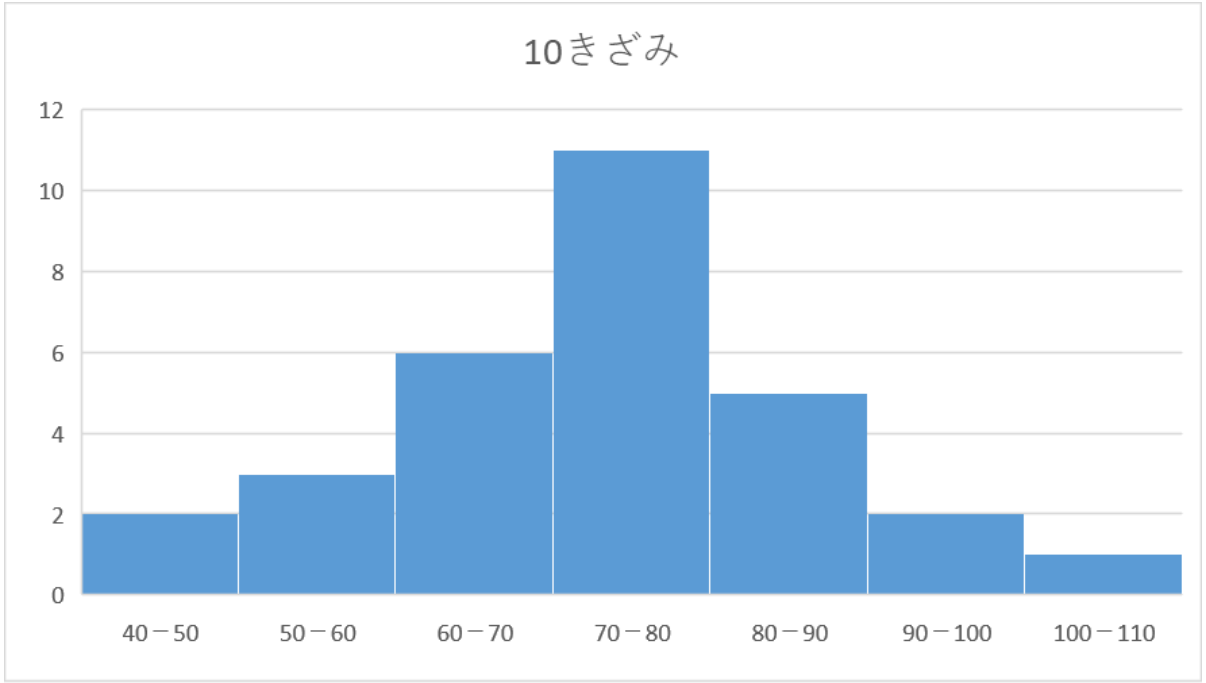
- 区間の細かさで印象が変わる
  - 区間の数は5-7ぐらいがベターと言われているが・・・

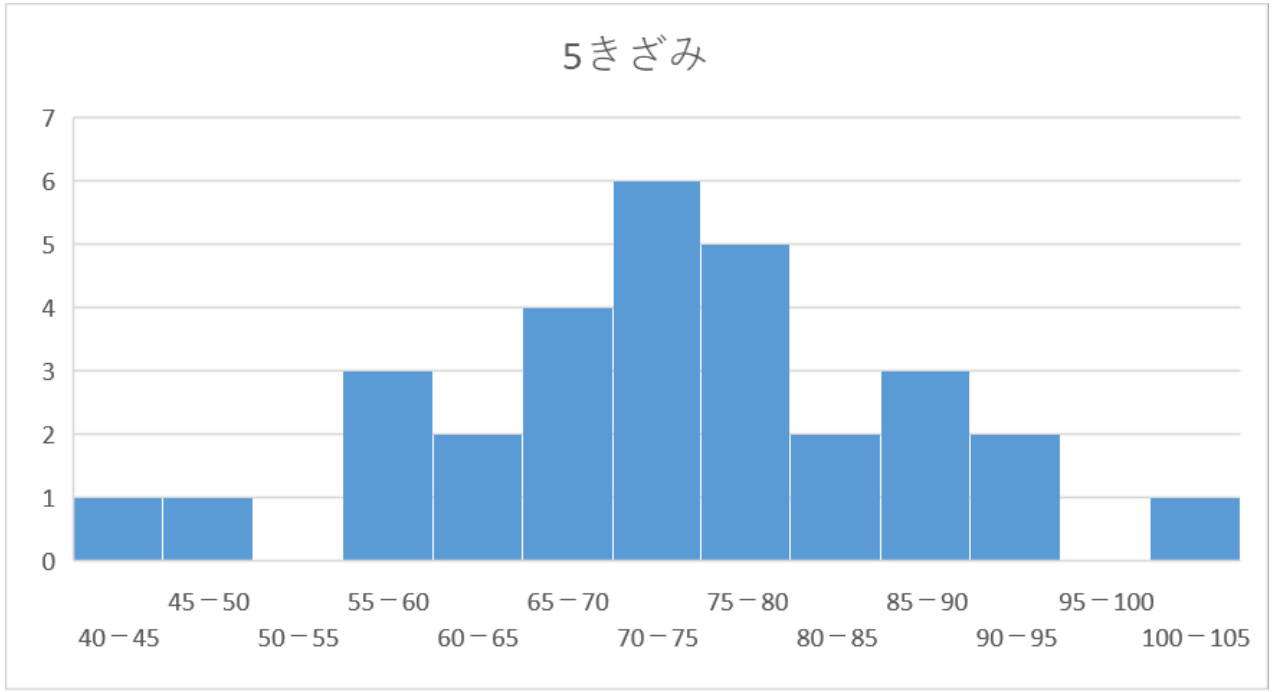
区間 (以上–未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		

区間 (以上–未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-45	42.5	1	0.033	1	0.033
45-50	47.5	1	0.033	2	0.067
50-55	52.5	0	0.000	2	0.067
55-60	57.5	3	0.100	5	0.167
60-65	62.5	2	0.067	7	0.233
65-70	67.5	4	0.133	11	0.367
70-75	72.5	6	0.200	17	0.567
75-80	77.5	5	0.167	22	0.733
80-85	82.5	2	0.067	24	0.800
85-90	87.5	3	0.100	27	0.900
90-95	92.5	2	0.067	29	0.967
95-100	97.5	0	0.000	29	0.967
100-105	102.5	1	0.033	30	1.000
105-110	107.5	0	0.000	30	1.000
合計		30	1.000		

ヒストグラム

- 区間の細かさで印象が変わる





いやもうほんとうにすいません

Execlでやるヒストグラム作成の実際

度数分布表を作ってから ヒストグラムを作ることになってる

- countifs関数を使う
  - 他のアプリやiPadでも可能
  - 絶対参照の利用がほぼ必須
- ピボットテーブルを使う
  - 関数を使わなくていい
  - ピボットテーブル機能がちょいと独特

COUNTIFS関数で度数分布表を作る

- COUNTIFS関数で度数を数える
  =COUNTIFS(データ範囲1,検索条件1,データ範囲2,検索条件2)
  - 検索条件で ”<50”とすると「50未満」という意味になる
    ” ”で囲んでやる必要がある
    ” ”の中は数値の場合は半角

例 =COUNTIFS($A$1:$A$30,”>=40”, $A$1:$A$30,”<50”)
これだと、40以上50未満

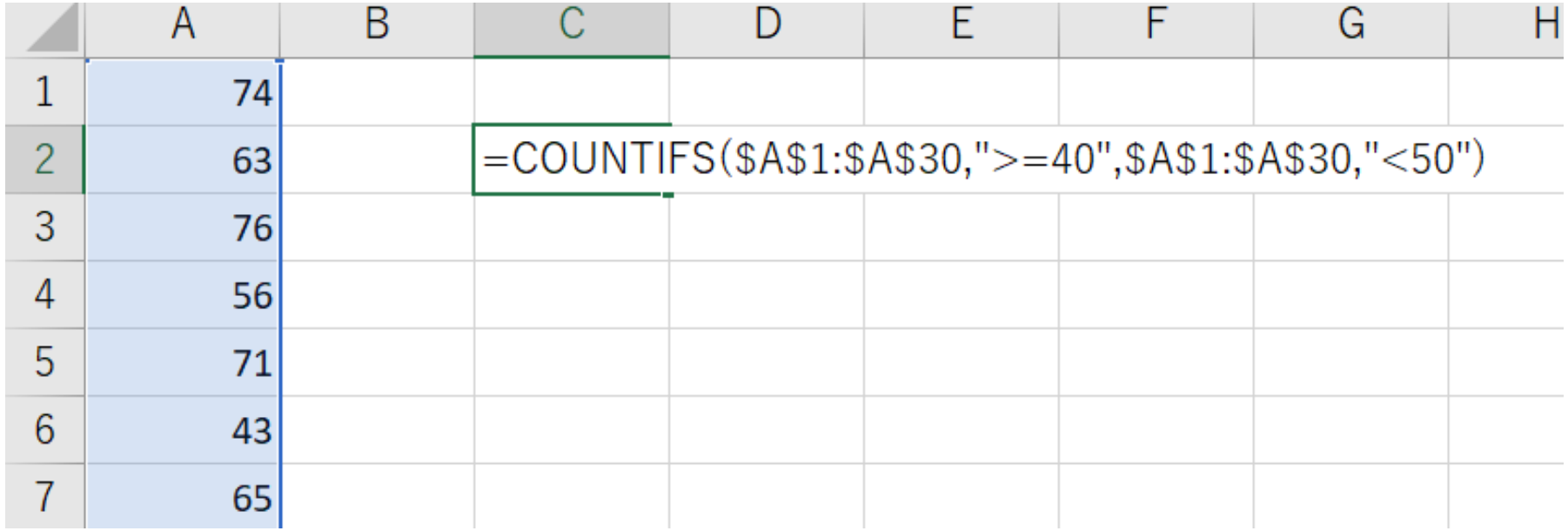
区間 (以上-未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		

検索条件

- 検索条件の書き方
  <=40　40以下
  <40　40未満
  >=40　40以上
  >40　40より上
  ≦や≧はExcelでは使えない
- 例　=COUNTIFS($A$1:$A$30,”>=40”, $A$1:$A$30,”<50”)
- 「以上ー未満」か「より上ー以下」
  - 「より上ー未満」だと境目の値がカウントされない
  - 「以上ー以下」だと境目の値が両方にカウントされる
    - 例:30–40,40–50

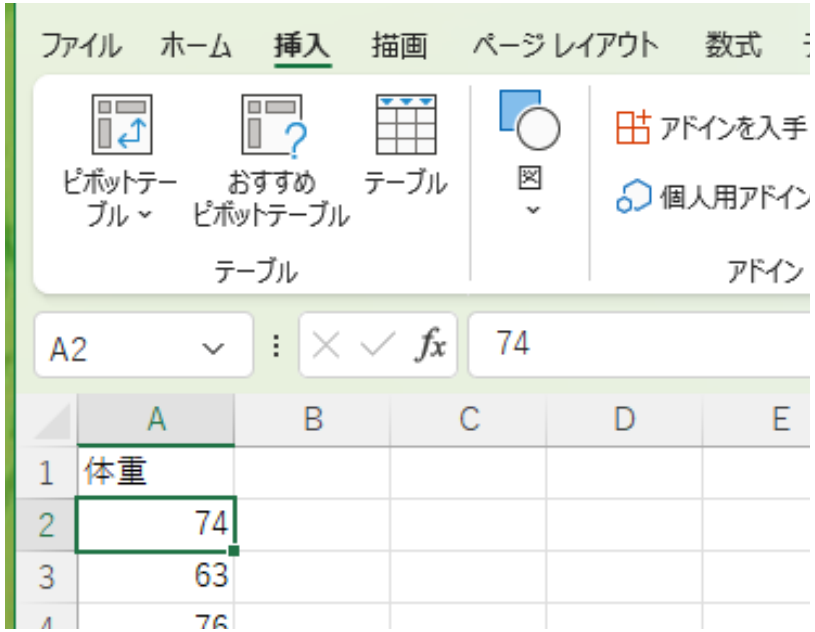
COUNTIFS関数の例

例　=COUNTIFS($A$1:$A$30,”>=40”,$A$1:$A$30,”<50”)



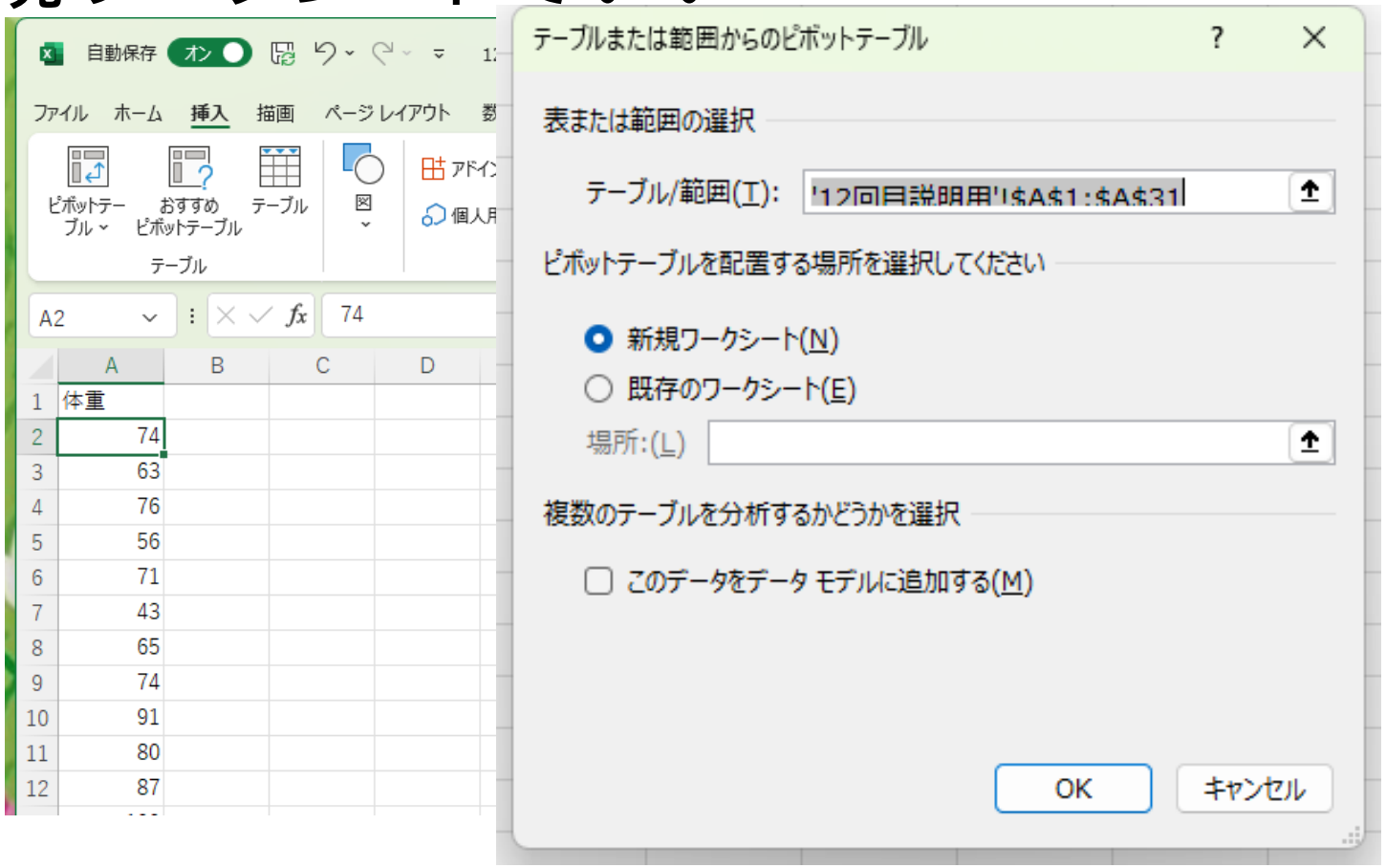
ピボットテーブルで度数分布表を作る

- データの先頭が変数名がないといけない
(ピボットテーブル共通)



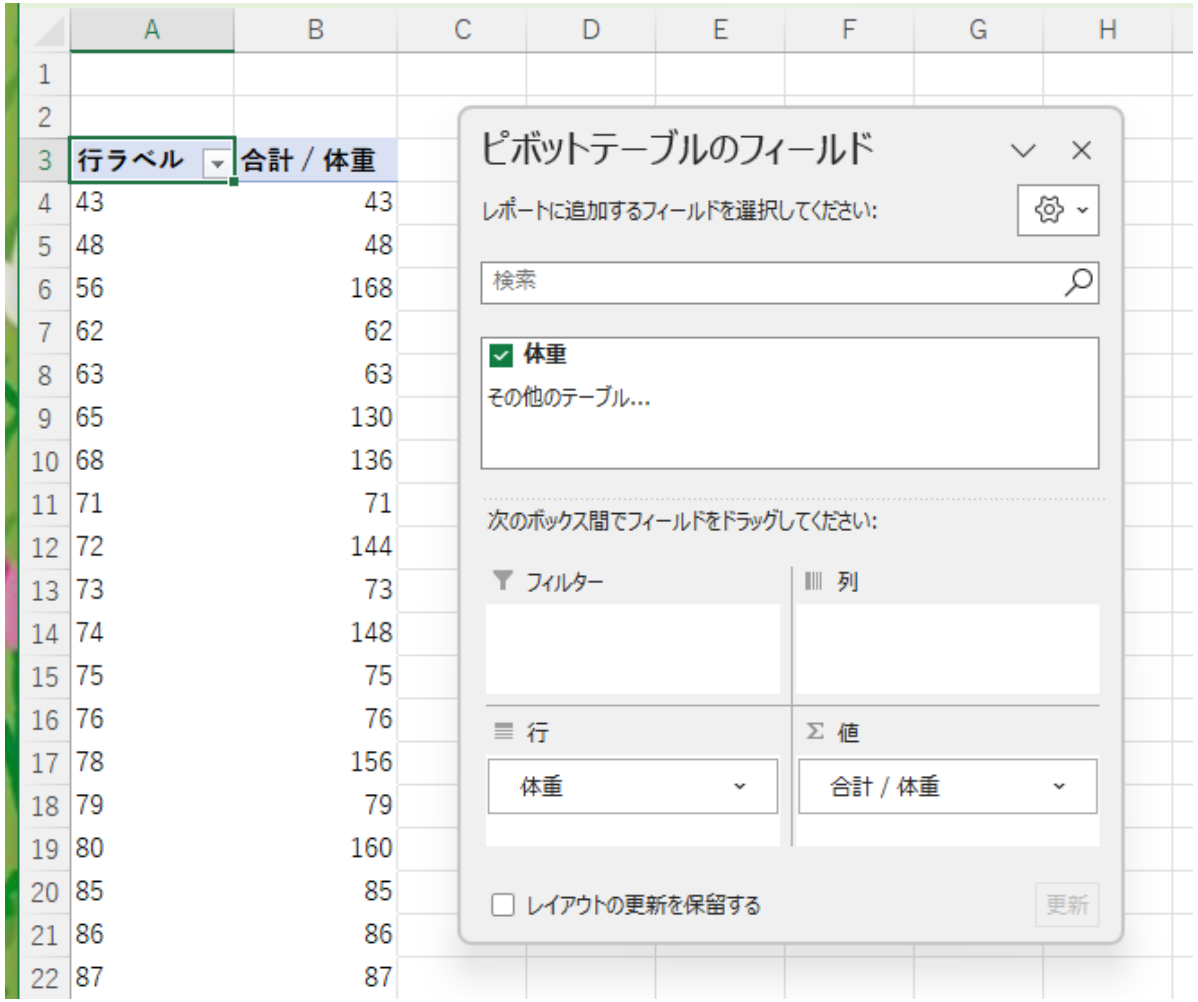
ピボットテーブルで度数分布表を作る

- 新規ワークシートでいい



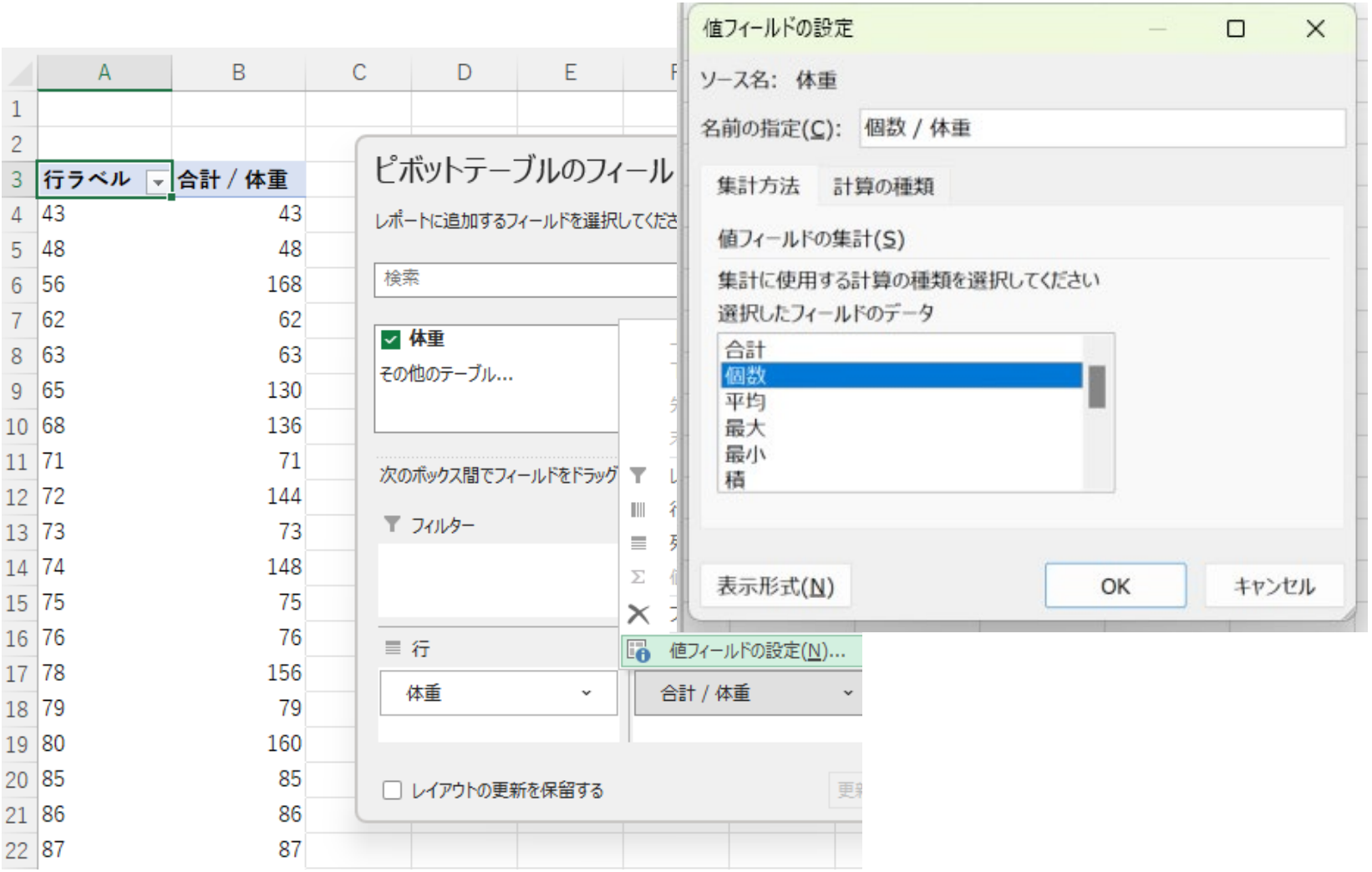
ピボットテーブルで度数分布表を作る

- 「体重」を
「Σ 値」と
「行」に入れる



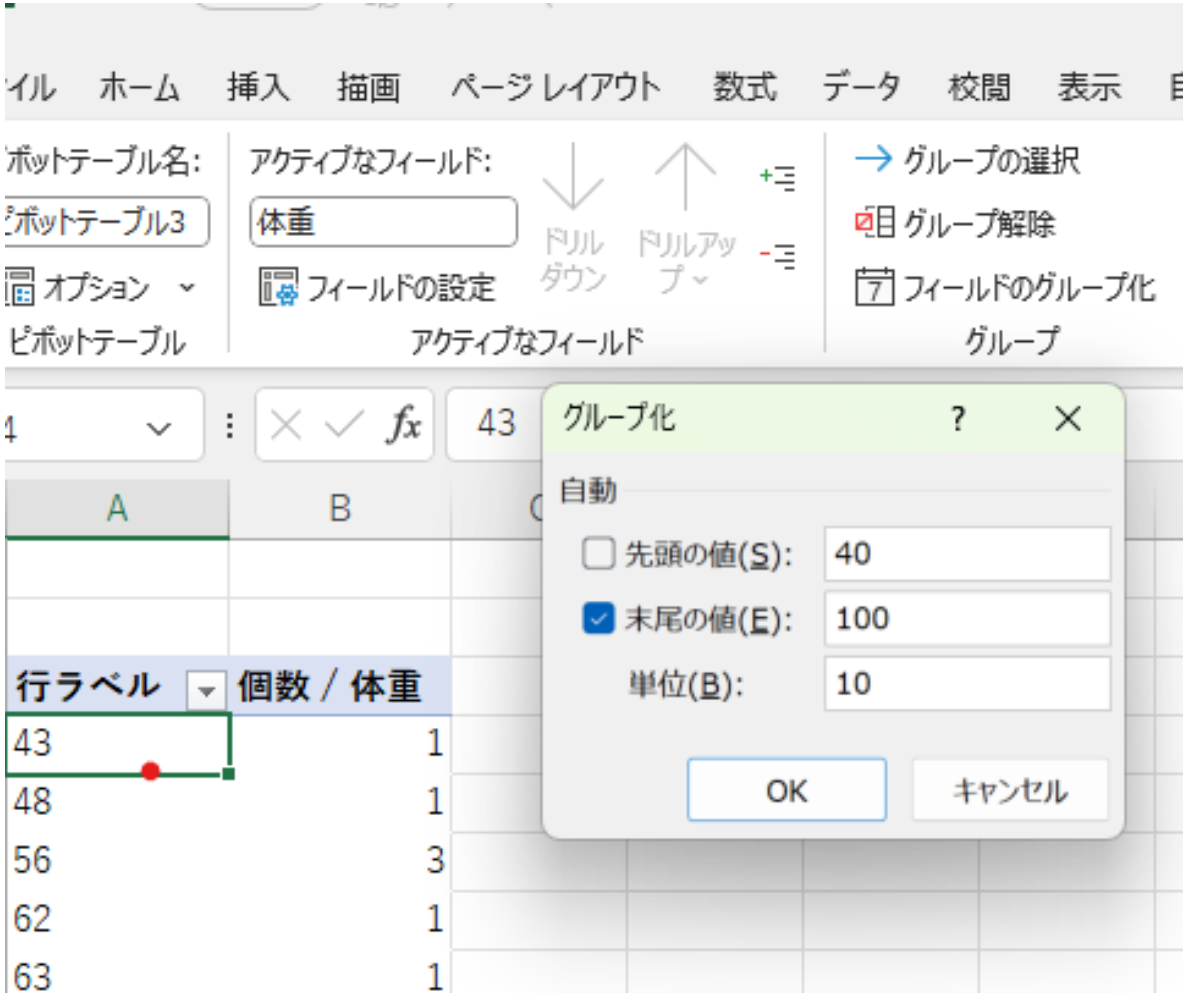
ピボットテーブルで度数分布表を作る

- 「Σ 値」が「合計/体重」になってしまう
- 「合計/体重」の上でマウスを右クリックして「値フィールドの設定」から「選択したフィールドのデータ」を「個数」に変更



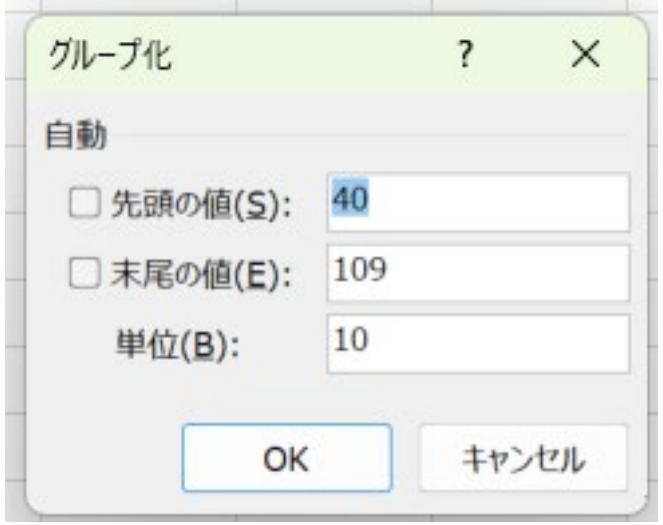
ピボットテーブルで度数分布表を作る

- 個数(度数)になったので
一番上のセル(赤点)をアクティブセルにして、「フィールドのグループ化」でグループ指定



ピボットテーブルのグループ化の問題点

- 最大値が区間の次の値と同じ時、最後の区間の端がその値まで含まれてしまう。
- 防止策
  - 最小値、最大値を算出してグループ化の端の値はそれを超えるように指定する。



ピボットテーブルからの度数分布表

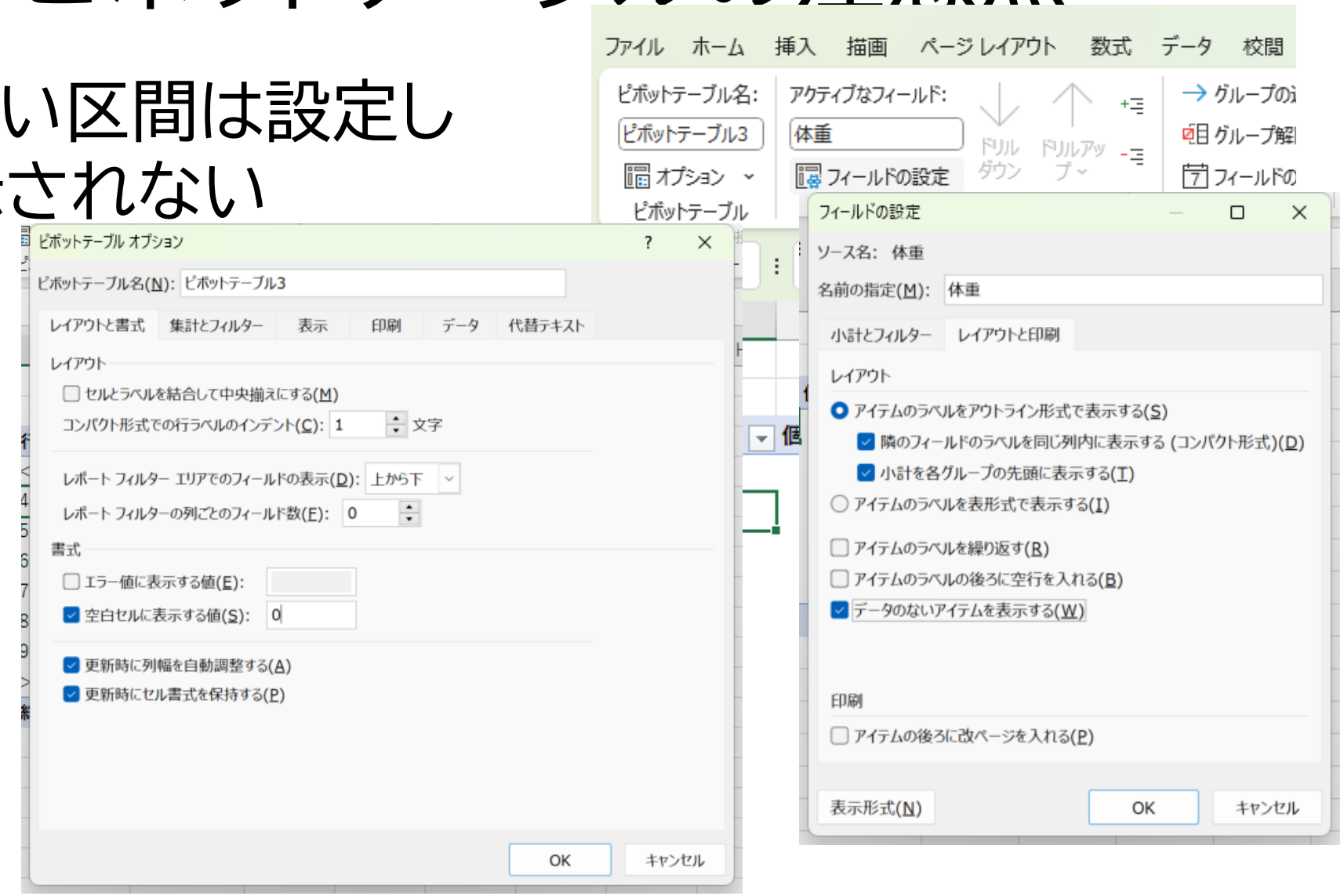
- 階級での集計になっているので、部分をコピーして残りの統計量を計算して表作成

行ラベル	個数 / 体重
40-49	2
50-59	3
60-69	6
70-79	11
80-89	5
90-99	2
100-109	1
総計	30

40-49	2
50-59	3
60-69	6
70-79	11
80-89	5
90-99	2
100-110	1
総計	30

ピボットテーブルの注意点

- データのない区間は設定しないと表示されない



Excelと集計について

- メニューから行う機能は、あまり信頼しすぎない
- 任せると結構ダメ
  - ヒストグラムの自動作成
  - ピボットテーブルのグループ化の端判定

  ※ 将来的にAI(Copilot)エンジンが導入されたときに、結果がちゃんとなるのか、手順のみ自動化なのか、気をつけないといけない
- 関数(countifs)ならほぼ間違いない

Excelでのヒストグラムの作り方

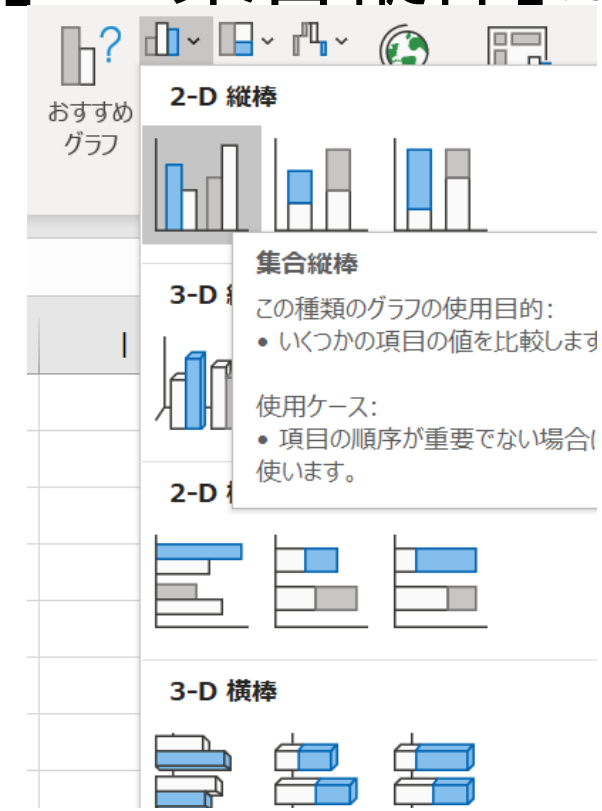
- 縦棒グラフを改造する
- グラフメニューにあるヒストグラムは超微妙
  - 階級幅が自動計算されるため、意図しない端数区間になることが多い

Excelでのヒストグラムの作り方

- 縦棒グラフを普通に作る
- グラフの棒の上で右クリックして「データ系列の書式設定」をクリック
- 「要素の間隔」を「0%」にする
  - Excel2016以降はグラフに「ヒストグラム」があるが、自動作成なので大抵うまくいかない。
    - 最大、最小値から自動で区間を求めるので、ちょうどよくならないことが多い

Excelで作るヒストグラム

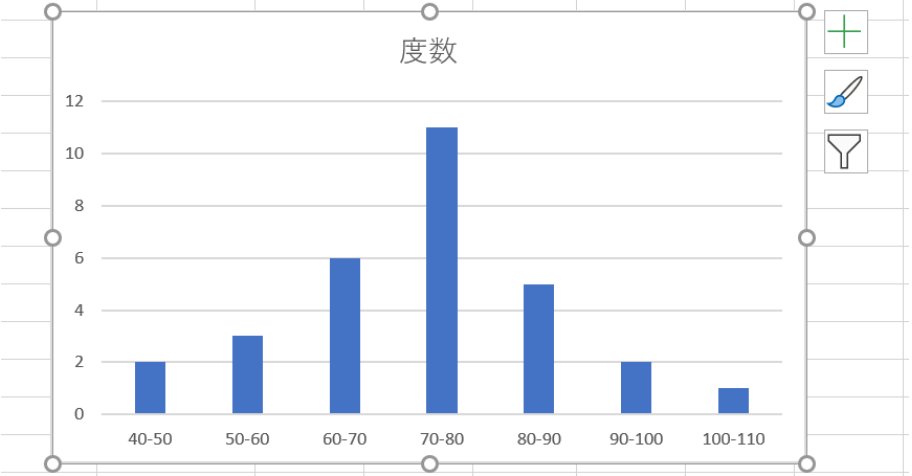
- まず「2-D縦棒」の「集合縦棒」から棒グラフをつくる



Excelで作るヒストグラム

- 先に度数分布表で集計しておく
- 離れたところを選択したいときは、一カ所目を選択した後で「ctrl」キーを押しながら次を選択する
- うまくいかないときは、グラフ用に離れていない表を作ってもいい。

区間 (以上-未満)	階級値	度数	相対度数	累積度数	累積 相対度数
40-50	45	2	0.067	2	0.067
50-60	55	3	0.100	5	0.167
60-70	65	6	0.200	11	0.367
70-80	75	11	0.367	22	0.733
80-90	85	5	0.167	27	0.900
90-100	95	2	0.067	29	0.967
100-110	105	1	0.033	30	1.000
合計		30	1.000		



Excelで作るヒストグラム

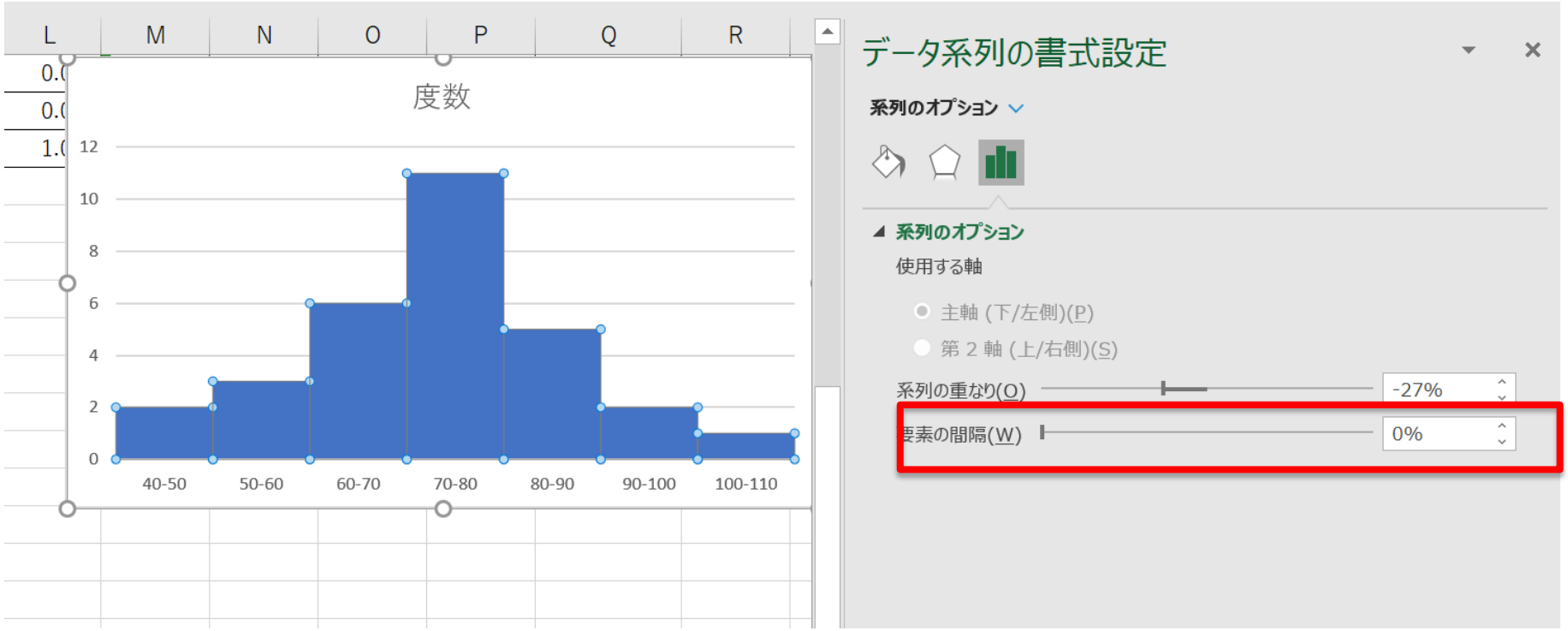
- 棒をくっつけるには、グラフをクリックして、棒の上で右クリックして出るメニューから、「データ系列の書式設定」を選ぶ

ヒストグラムの作り方

- データ系列の書式設定で「要素の間隔」を0%にすると
くっつく

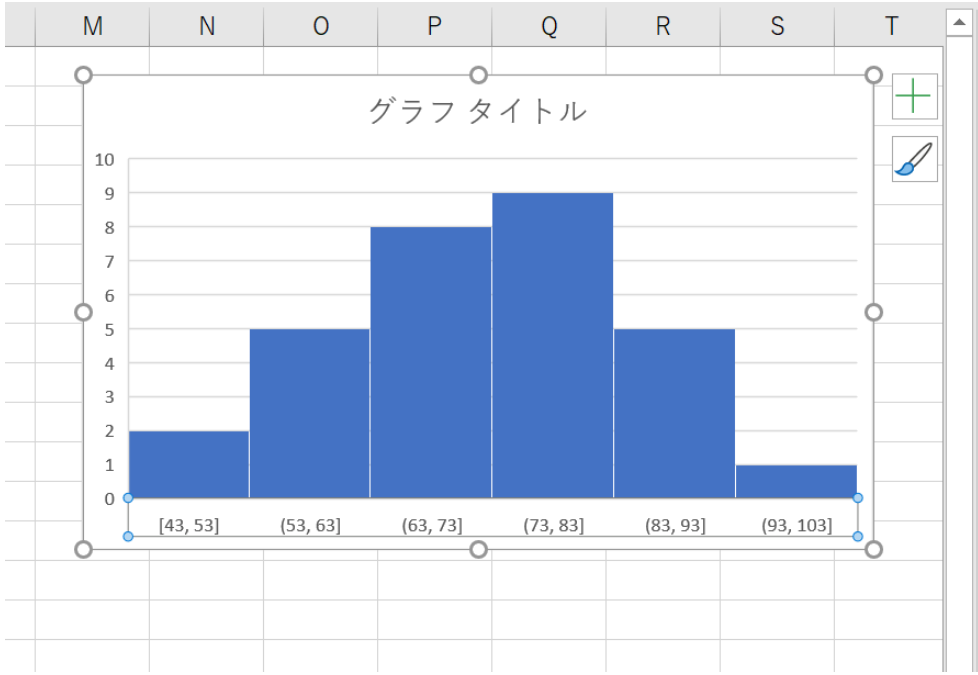
ヒストグラムの作り方

- データ系列の書式設定で「要素の間隔」を0%にするとくっつく



Excelでヒストグラムメニューを使わない理由

- 任意で値を設定しても反映されず、必ず最小値から間隔を作り出すので、区切りがおかしくなるから

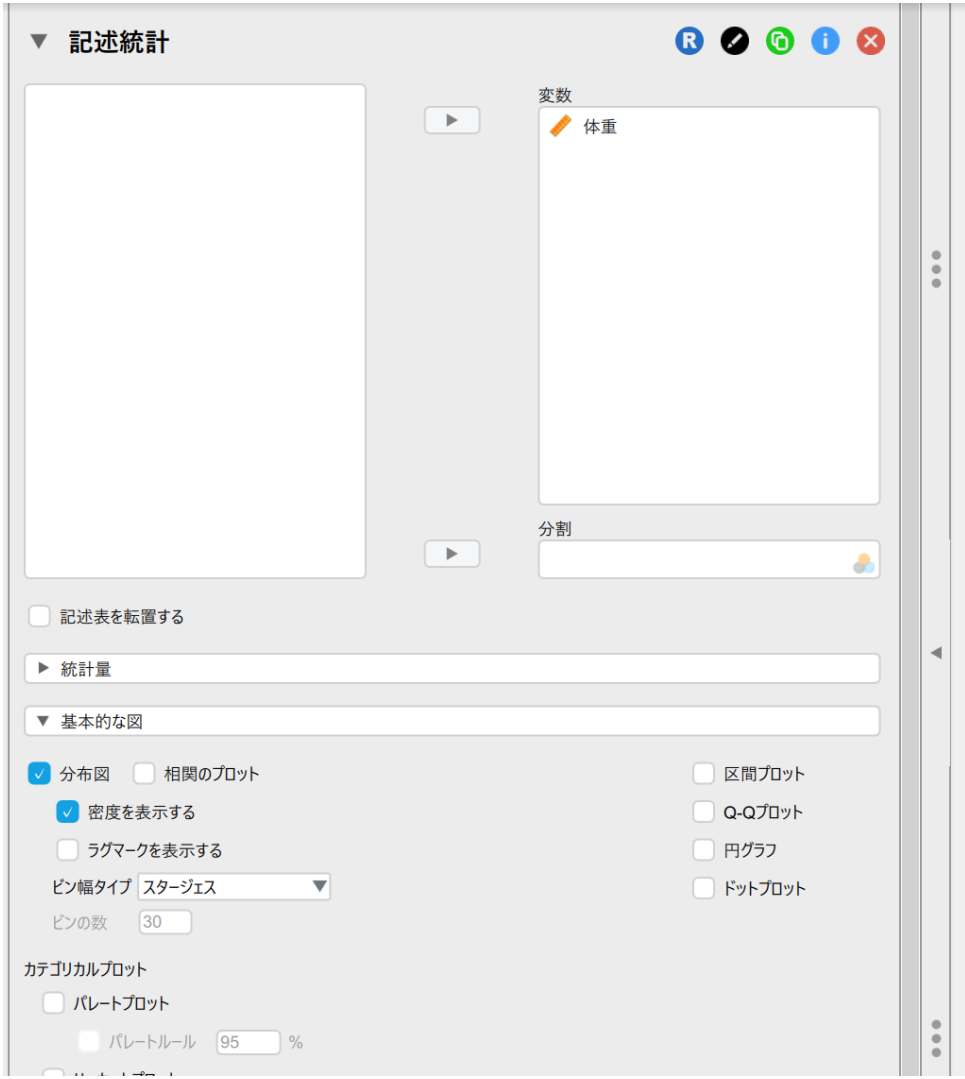


統計アプリでやる場合

あんまりうまいこといかない

- JASP
  - とりあえずヒストグラムを作ってはくれる
    - あまりうまくは行かない
    - 区間が　より上～以下　になる(以上～未満ではない)
  - 度数表は作ってくれない
    - 自分で変数を作る必要がある
    - 作成した変数からは棒グラフしか作れない

JASPの場合



結果

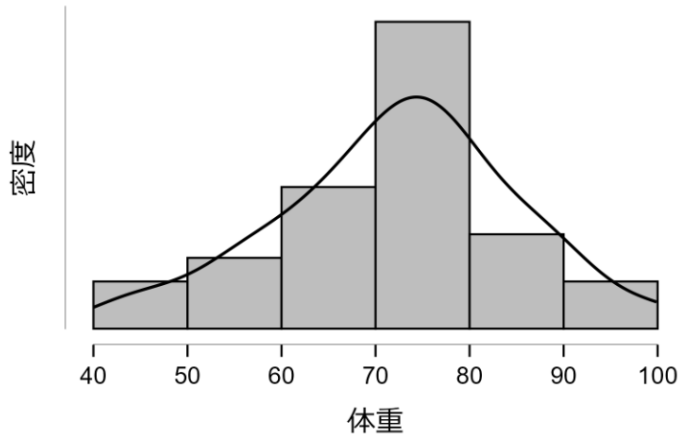
記述統計

記述統計

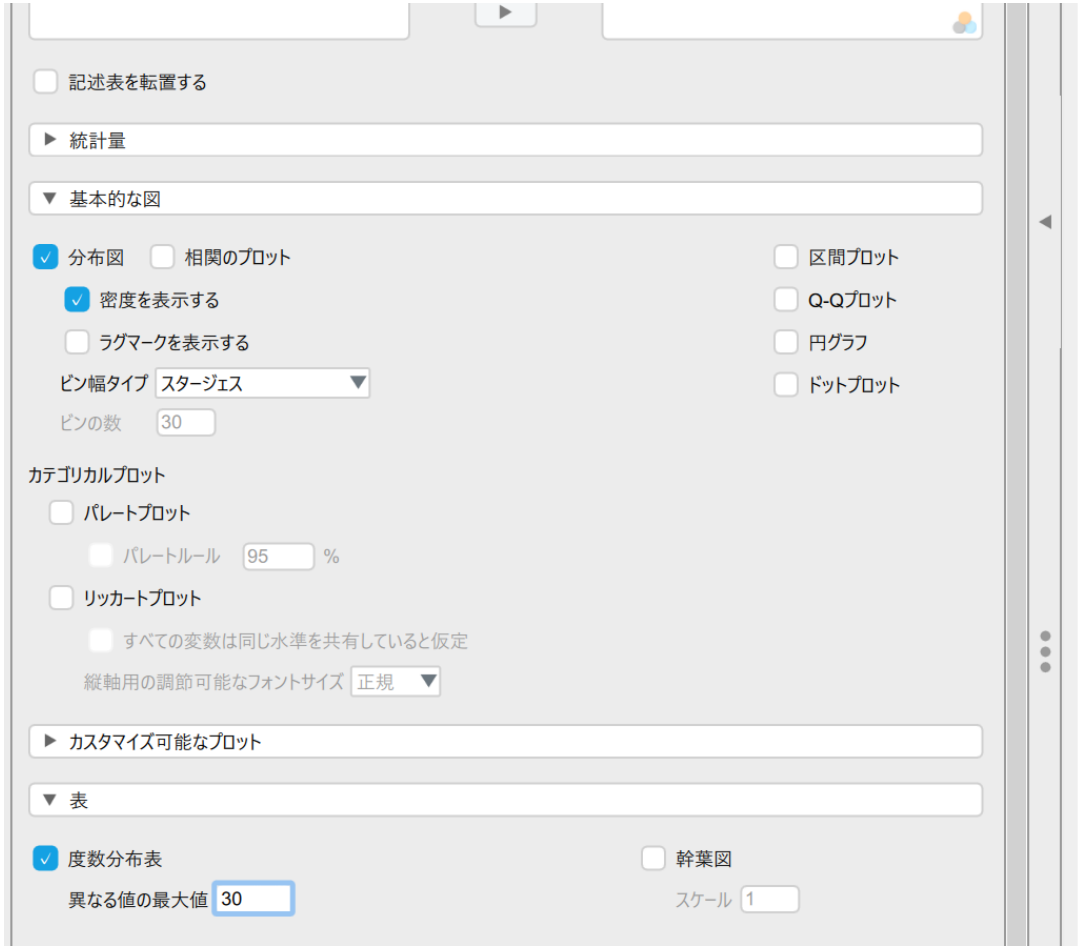
	体重
妥当	30
欠損値	0
平均値	72.37
標準偏差	12.91
最小値	43.00
最大値	100.0

分布のプロット

体重



JASPの場合



度数分布表

体重 の頻度

体重	頻度	パーセント	有効パーセント	累積パーセント
43	1	3.3	3.3	3.3
48	1	3.3	3.3	6.7
56	3	10.0	10.0	16.7
62	1	3.3	3.3	20.0
63	1	3.3	3.3	23.3
65	2	6.7	6.7	30.0
68	2	6.7	6.7	36.7
71	1	3.3	3.3	40.0
72	2	6.7	6.7	46.7
73	1	3.3	3.3	50.0
74	2	6.7	6.7	56.7
75	1	3.3	3.3	60.0
76	1	3.3	3.3	63.3
78	2	6.7	6.7	70.0
79	1	3.3	3.3	73.3
80	2	6.7	6.7	80.0
85	1	3.3	3.3	83.3
86	1	3.3	3.3	86.7
87	1	3.3	3.3	90.0
90	1	3.3	3.3	93.3
91	1	3.3	3.3	96.7
100	1	3.3	3.3	100.0
欠損値	0	0.0		
合計	30	100.0		

分布のプロット

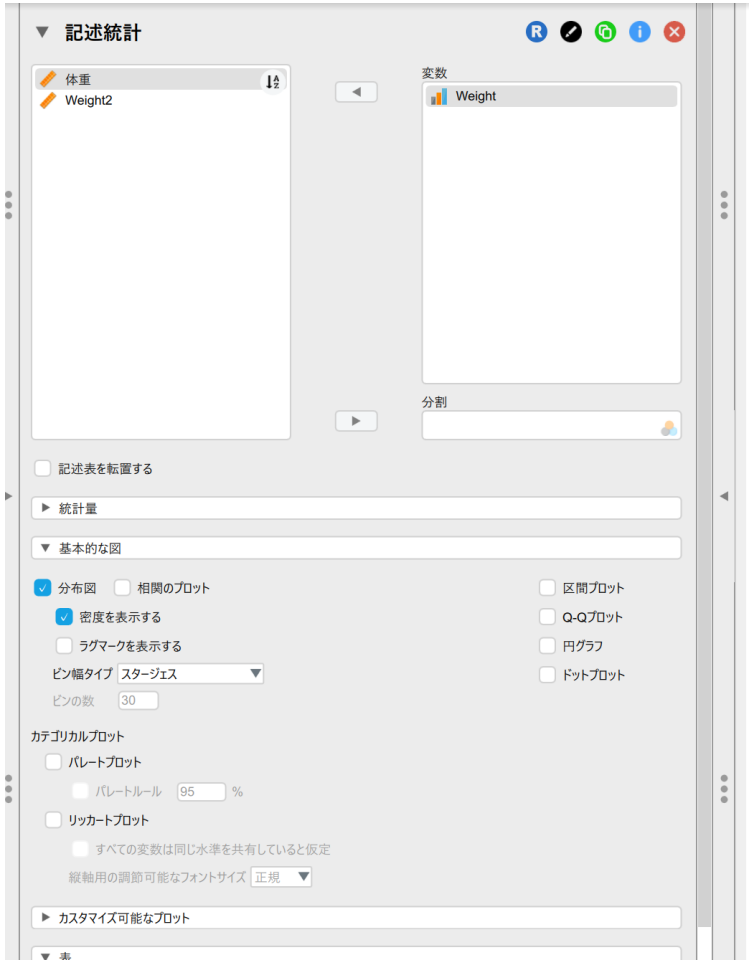
JASPで区間の度数分布表

	体重	Weight
1	74	70-79
2	63	60-69
3	76	70-79
4	56	50-59
5	71	70-79
6	43	40-49
7	65	60-69
8	74	70-79
9	91	90-99

区間分割の式

- cut(体重, breaks = c(0, 40, 50, 60, 70, 80, 90, 100,110), labels = c("～40", "40-49", "50-59", "60-69", "70-79","80-89","90-99", "100以上"), right = FALSE)

JASPで区間の度数分布表



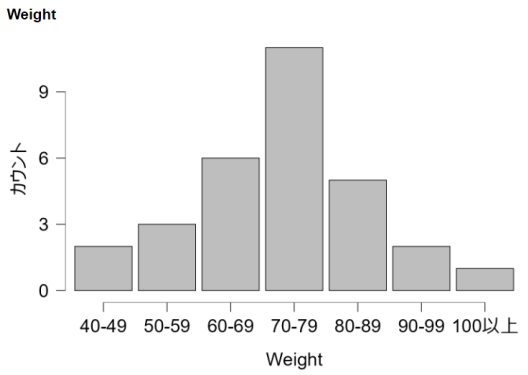
	Weight
妥当	30
欠損値	0
平均値	
標準偏差	
最小値	40
最大値	100

度数分布表

Weight の頻度

Weight	頻度	パーセント	有効パーセント	累積パーセント
40-49	2	6.7	6.7	6.7
50-59	3	10.0	10.0	16.7
60-69	6	20.0	20.0	36.7
70-79	11	36.7	36.7	73.3
80-89	5	16.7	16.7	90.0
90-99	2	6.7	6.7	96.7
100以上	1	3.3	3.3	100.0
欠損値	0	0.0		
合計	30	100.0		

分布のプロット



JASPで区間の度数分布表

名前: Weight2　長い名前: Weight2

列の種類: スケール　説明:

計算された種類: Rコードで計算

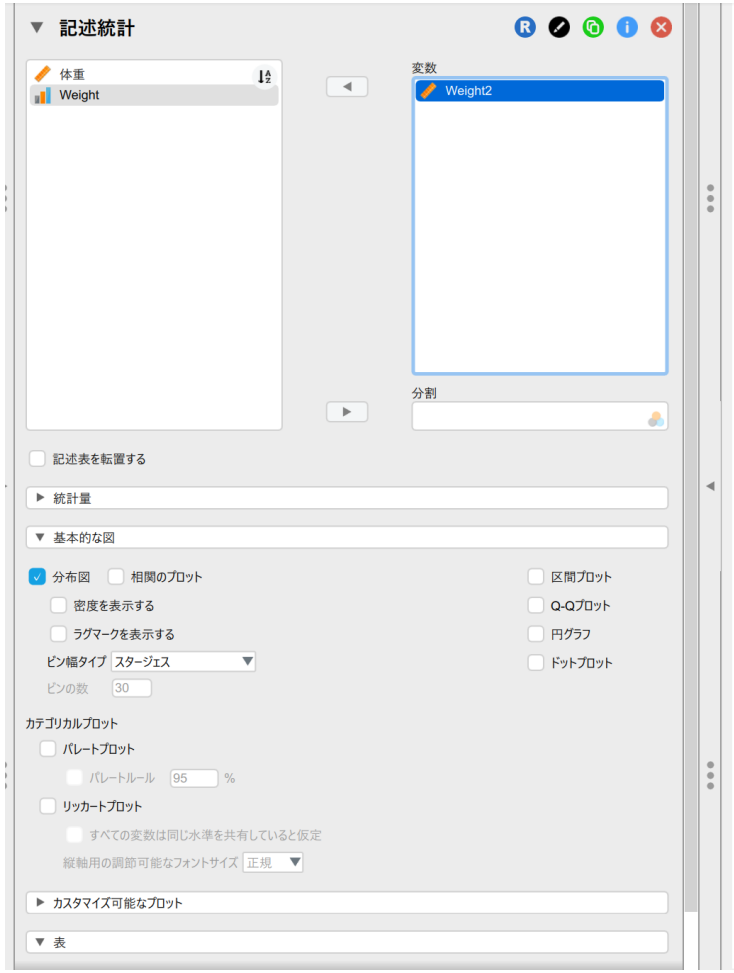
計算された列の定義 | ラベルエディタ | 欠測値

```
as.numeric(cut(体重, breaks = c(40,50,60,70,80,90,100,110), right = FALSE))
```

計算された列:

	体重	Weight	Weight2
1	74	70-79	4
2	63	60-69	3
3	76	70-79	4
4	56	50-59	2
5	71	70-79	4
6	43	40-49	1
7	65	60-69	3

JASPで区間の度数分布表



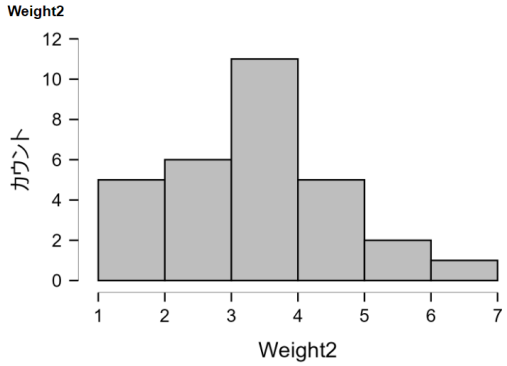
	Weight2
妥当	30
欠損値	0
平均値	3.800
標準偏差	1.400
最小値	1.000
最大値	7.000

度数分布表

Weight2 の頻度

Weight2	頻度	パーセント	有効パーセント	累積パーセント
1	2	6.7	6.7	6.7
2	3	10.0	10.0	16.7
3	6	20.0	20.0	36.7
4	11	36.7	36.7	73.3
5	5	16.7	16.7	90.0
6	2	6.7	6.7	96.7
7	1	3.3	3.3	100.0
欠損値	0	0.0		
合計	30	100.0		

分布のプロット



区間分割の式

- as.numeric(cut(体重, breaks = c(40,50,60,70,80,90,100,110), right = FALSE))

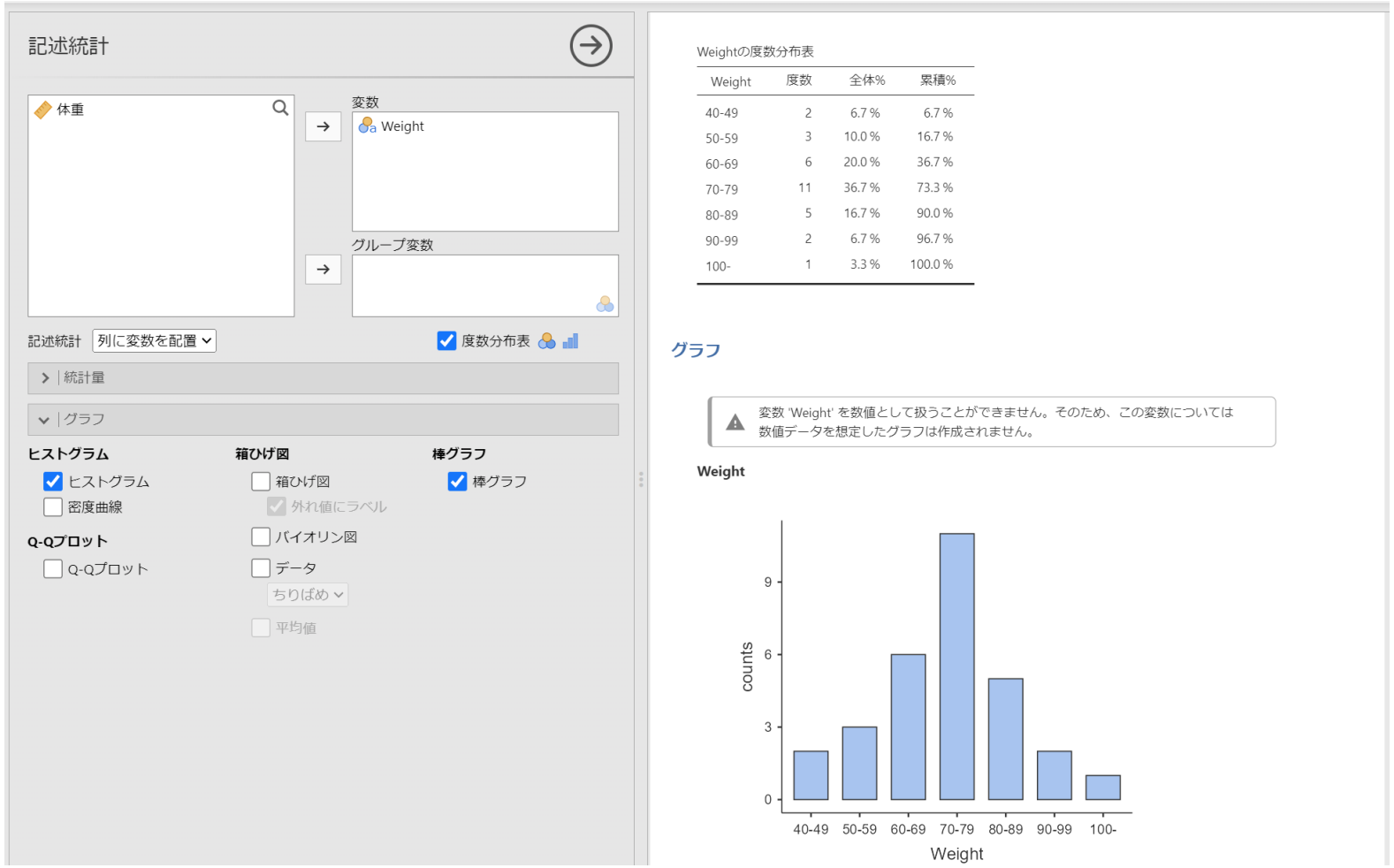
あんまりうまいこといかない

- jamovi
  - とりあえずヒストグラムを作ってはくれる
    - あまりうまくは行かない
    - 区間にならない
  - 度数表は作ってくれない
    - 自分で変数を作る必要がある
    - 作成した変数からは棒グラフしか作れない

jamoviで区間の度数分布表

	体重	Weight
1	74	70-79
2	63	60-69
3	76	70-79
4	56	50-59
5	71	70-79
6	43	40-49
7	65	60-69
8	74	70-79
9	91	90-99
10	80	80-89
11	87	80-89
12	100	100-

jamoviで区間の度数分布表



区間分割の式(jamovi)

- IF(体重 < 40, ”0-39”,IF(体重 < 50, ”40-49”,IF(体重 < 60, ”50-59”,IF(体重 < 70, ”60-69”,IF(体重 < 80, ”70-79”,IF(体重 < 90, ”80-89”,IF(体重 < 90, ”80-89”,IF(体重 < 100, ”90-99”,”100-”)))))))

式が同じで無いのは

- JASPとjamoviには関数制限がかかっている
- データ構造を壊す可能性があると判断した関数は使えないようになっている
- だったら、R使えよっていうこと

対策

- 編集した変数を作る
  - Excelでやる
  - JASPやjamoviで関数を使って作る
- ヒストグラム
  - Excelで棒グラフを改造
  - 元となるデータをExcelで作るか、JASPやjamoviで集計してExcelに渡すか

ほんとにねぇ

なんでこんな面倒なことをするのか

データの正規性の判断

- 歪度(Skewness)
  - 0:左右対称(正規分布)
  - 正の値(＋):右に裾が長い(右歪み)
  - 負の値(－):左に裾が長い(左歪み)
  - ±1以内ならおおむね対称☆
- 尖度(Kurtosis)※Fisherの定義(統計ソフト標準)
  - 0:正規分布判定
  - 正の値(＋):尖っている、外れ値多い傾向
  - 負の値(－):平らである、外れ値は少ないが中央が多くない
  - ±1以内ならおおむね正規に近い☆

☆教科書やソフトによって基準は異なる(±1が一般的)

データの正規性の判断

記述統計

記述統計

	体重
妥当	30
欠損値	0
平均値	72.37
標準偏差	12.91
歪度	−0.232
歪度の標準誤差	0.427
尖度	0.103
尖度の標準誤差	0.833
最小値	43.00
最大値	100.0

データの正規性の判断

データの正規性の判断

- 「正規分布らしいか」は結局「形」で判断する
  - 「正規性検定」
    - Shapiro–Wilk 検定
    - Kolmogorov–Smirnov 検定
  - サンプル数にものすごく敏感
    - nが多い→ ほんのわずかな歪みでも帰無仮説棄却レベル
    - nが少ない→検出力が低く、歪みがあっても棄却されない

正規性があるとみなしていい場合

- 各群のサンプルサイズが目安として30程度以上
  - 中央極限定理が働く
    - サンプルサイズを大きくすると、元のデータの分布に関係なく、標本平均の分布は正規分布に近づく

    → だから「正規分布を仮定する統計手法」が広く使える

正規性があるとみなしていい場合

- 外れ値が極端に多くない
  - 片側だけに外れ値が集中していない
- 群間の分散が大きく異ならない
  - 分散の等質性がある程度保たれている
- 中央がピークでだいたい対称
- 単峰性である

正規分布かどうか

- 数字よりも形と常識で判断
- 完璧な正規分布なんて実データにはほとんど存在しない

t検定や分散分析の頑健性

- 「母集団が完全に正規分布でなくても、ある程度のサンプルサイズがあれば結果はほぼ保たれる」

→分析の頑健性(robustness)

  - 多少の条件の違いやデータの乱れに影響されにくい
  - 仮定(正規性、等分散など)が完全に満たされていなくても、結果が大きく崩れない

これを判断するのが

- 歪度や尖度、正規性の検定よりも、ヒストグラムで形を見た方がわかりやすい
  - 科学的に値がいくつだと正規分布という決まりがない
- 形で見てみて、「単峰」、「だいたい対称」、「中央にピーク」、「外れ値があんまりない」ということであれば、t検定や分散分析はロバストだからあまり問題ない

面倒だけれども

- 分析の前に、各データの区間での度数分布とヒストグラムを作成して、データがどういう傾向を持っているか把握するべき
- 外れ値があったり、対称でない、単峰でない場合
  - データが本当にそう
  - 測定やデータ作成の時に何か間違った