データ分析の基礎-1

2021年7月14日

本日の内容

1 データの代表値
　平均・バラツキ、中央値

2 データの視覚化

3 正規分布、t 分布

4　t 検定
　検定のしくみ

中央値（MEDIAN）

230	287	252	313	825	331	292	851	229

平均値＝401.1

平均値は極端な値の影響を受け易い

1	2	3	4	5	6	7	8	9
229	230	252	287	292	313	331	825	851

大きさの順番に並び替えて、中央に位置する値を代表値

四分位数

・データを小さい順に並べる。

範囲＝最大値－最小値

・4分の1ずつの場所にある値

Q1：第1四分位数（25%点）
Q2：第2四分位数（50%点・中央値）
Q3：第3四分位数（75%点）

四分位範囲（IQR）　：　Q3－Q1

230	287	252	313	825	331	292	851	229

1	2	3	4	5	6	7	8	9
229	230	252	287	292	313	331	825	851

平均　　：401.1
Q1 四分位点（25%）　：252
Q2 中央値　　：292
Q3 四分位点（75%）　：331

四分位範囲　Q3－Q1＝331－252＝79

外れ値

◇四分位範囲（IQR）を基準

Q1 第1四分位数（25%）　：　252
Q2 中央値　　：　292
Q3 第3四分位数（75%）　：　331
IQR　Q3－Q1＝331－252＝79

Q1－1.5×IQR以下　　Q3＋1.5×IQR以上

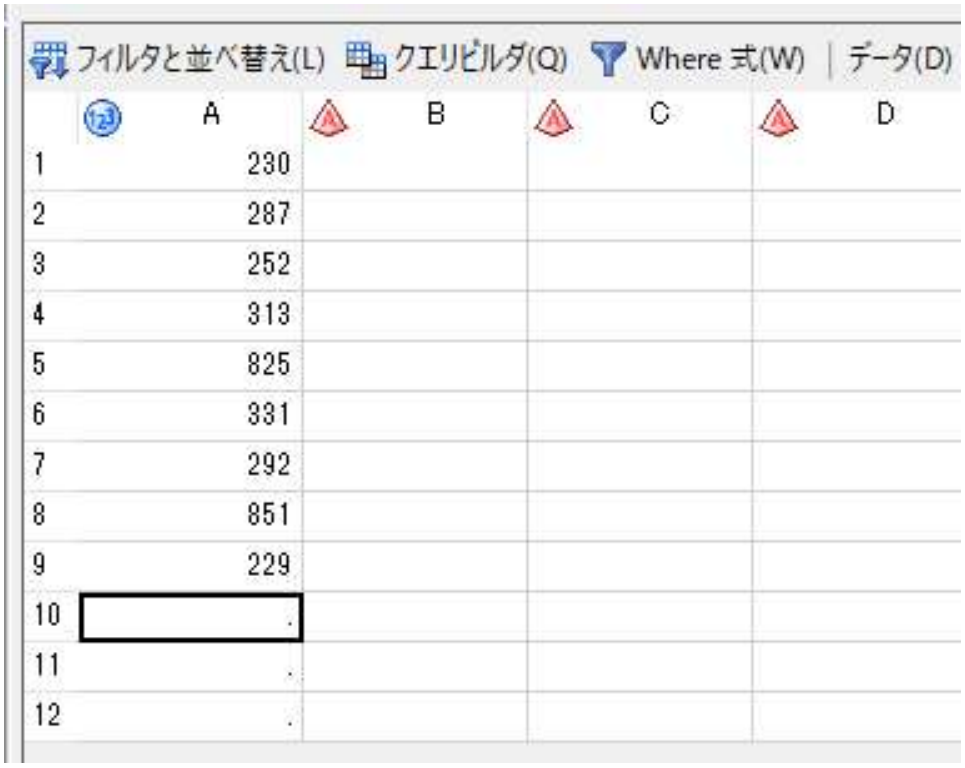
外れ値　：　　133.5以下　　449.5以上

注）252－1.5×79＝133.5　331+1.5×79＝449.5

EG

中央値、四分位数を求める。（SAS EG)

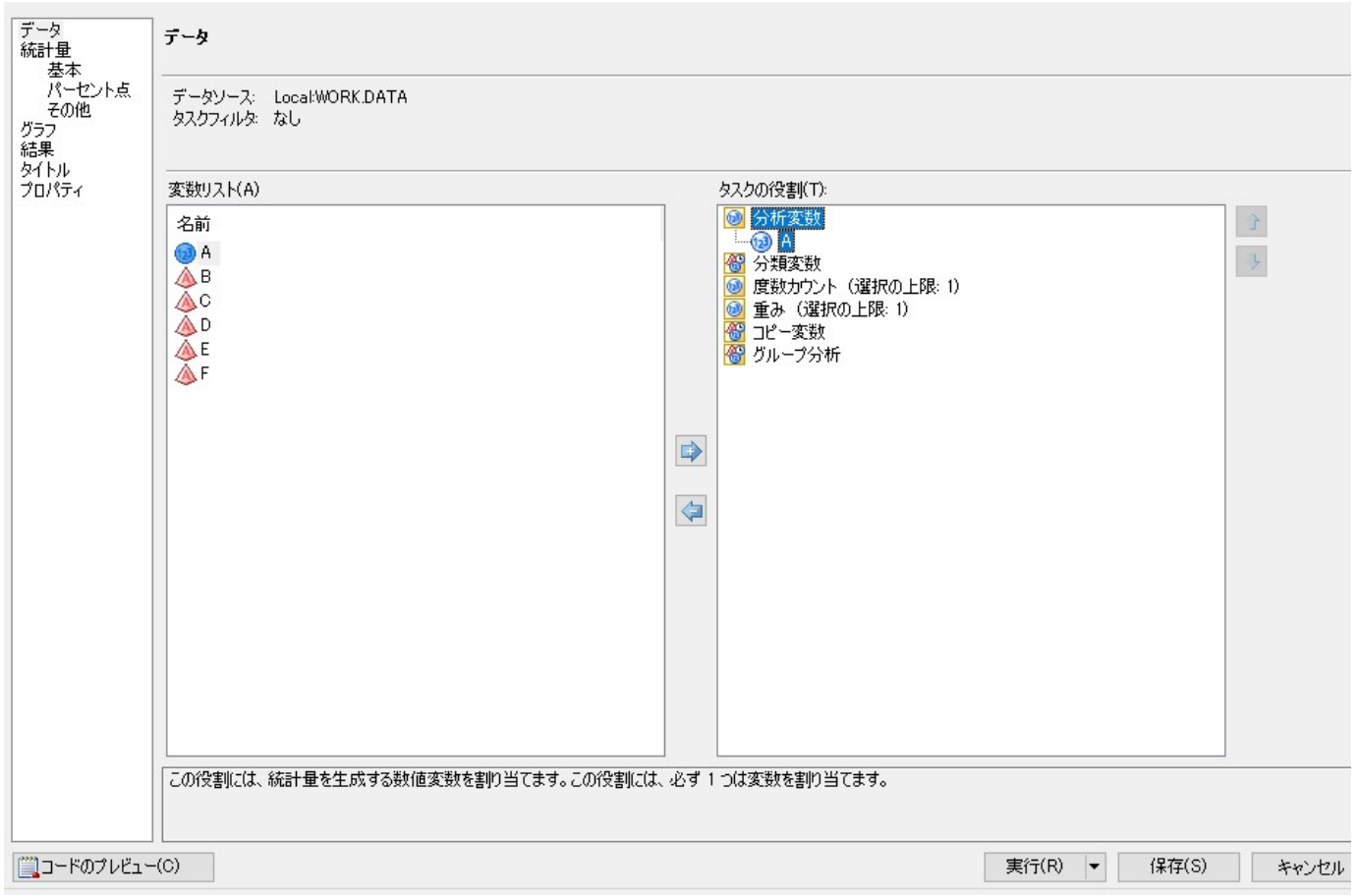
（1）データを入力する。



	A	B	C	D
1	230			
2	287			
3	252			
4	313			
5	825			
6	331			
7	292			
8	851			
9	229			
10	.			
11	.			
12	.			

EG

（2）「記述統計」–「要約統計量」をクリックし、変数リストから分析変数を指定する。



EG

（3）「統計量」－「パーセント点」をクリックし、出力項目を設定する。

EG

（4）出力結果が表示される。

分析変数：A							
平均	標準偏差	最小値	最大値	N	下側四分位点	中央値	上側四分位点
401.1111111	250.2056376	229.0000000	851.0000000	9	252.0000000	292.0000000	331.0000000

平均　401.1
Q1 四分位点（25%）　252
Q2 中央値　292
Q3 四分位点（75%）　331

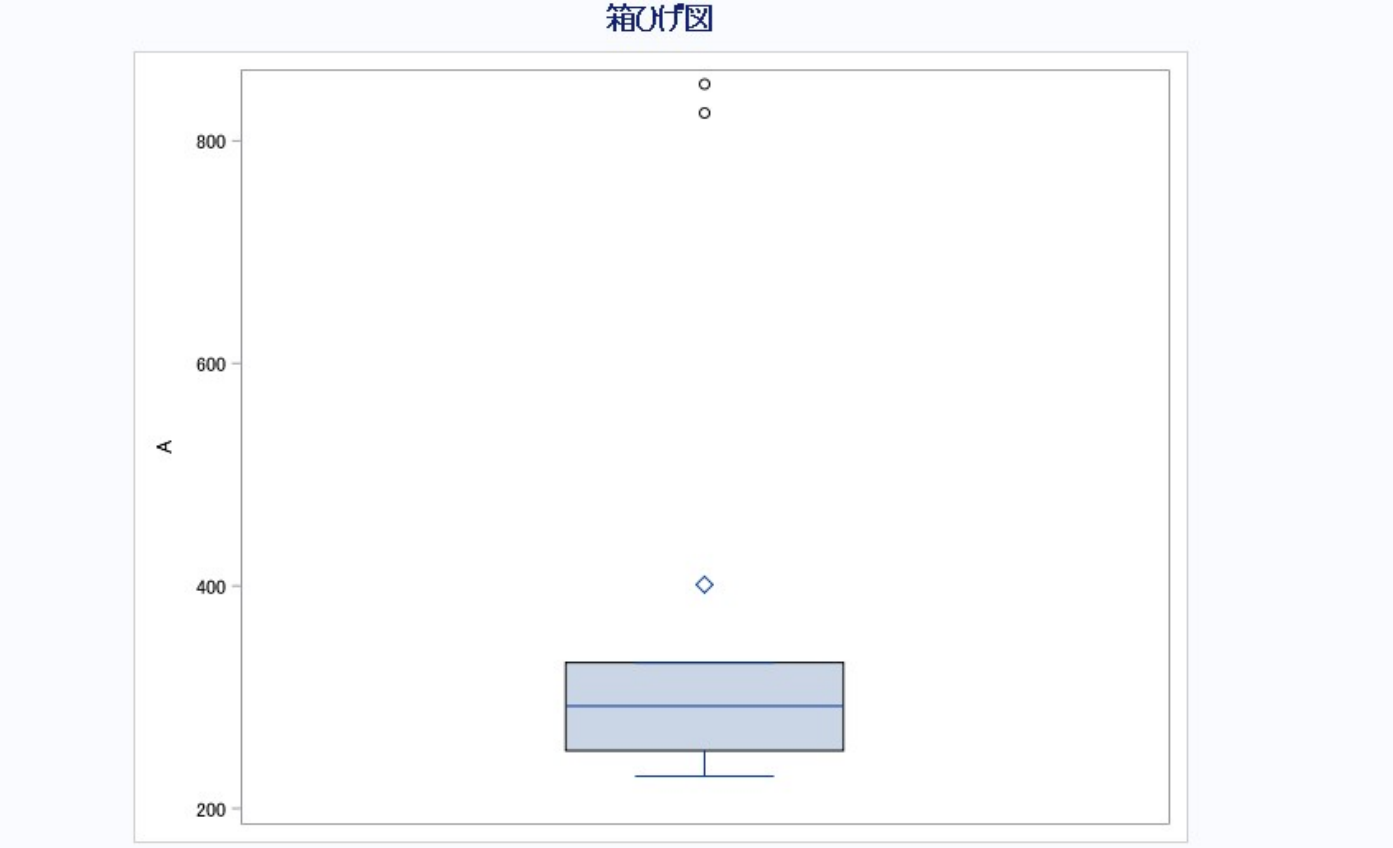
EG

箱ひげ図の作成（SAS EG）

「グラフ」をクリックし、「箱ひげ図」をチェックする。

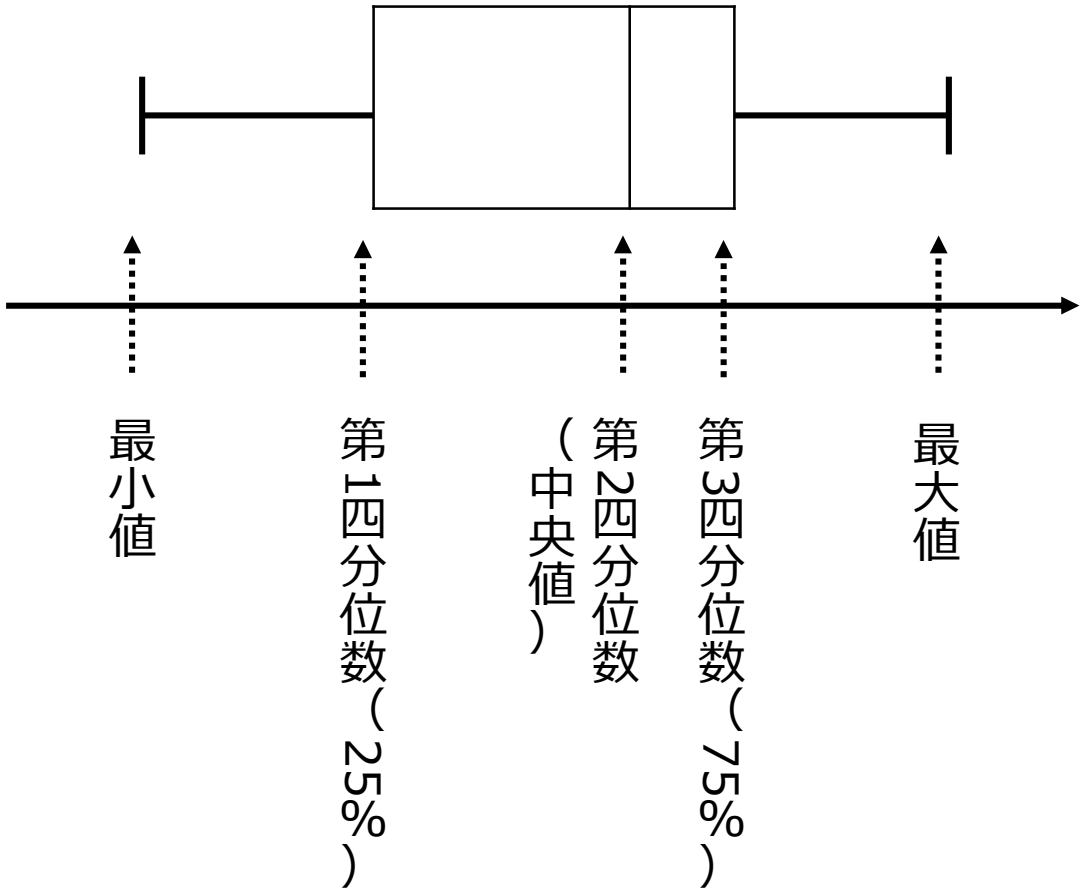
EG

分析変数：A							
平均	標準偏差	最小値	最大値	N	下側四分位点	中央値	上側四分位点
401.1111111	250.2056376	229.0000000	851.0000000	9	252.0000000	292.0000000	331.0000000

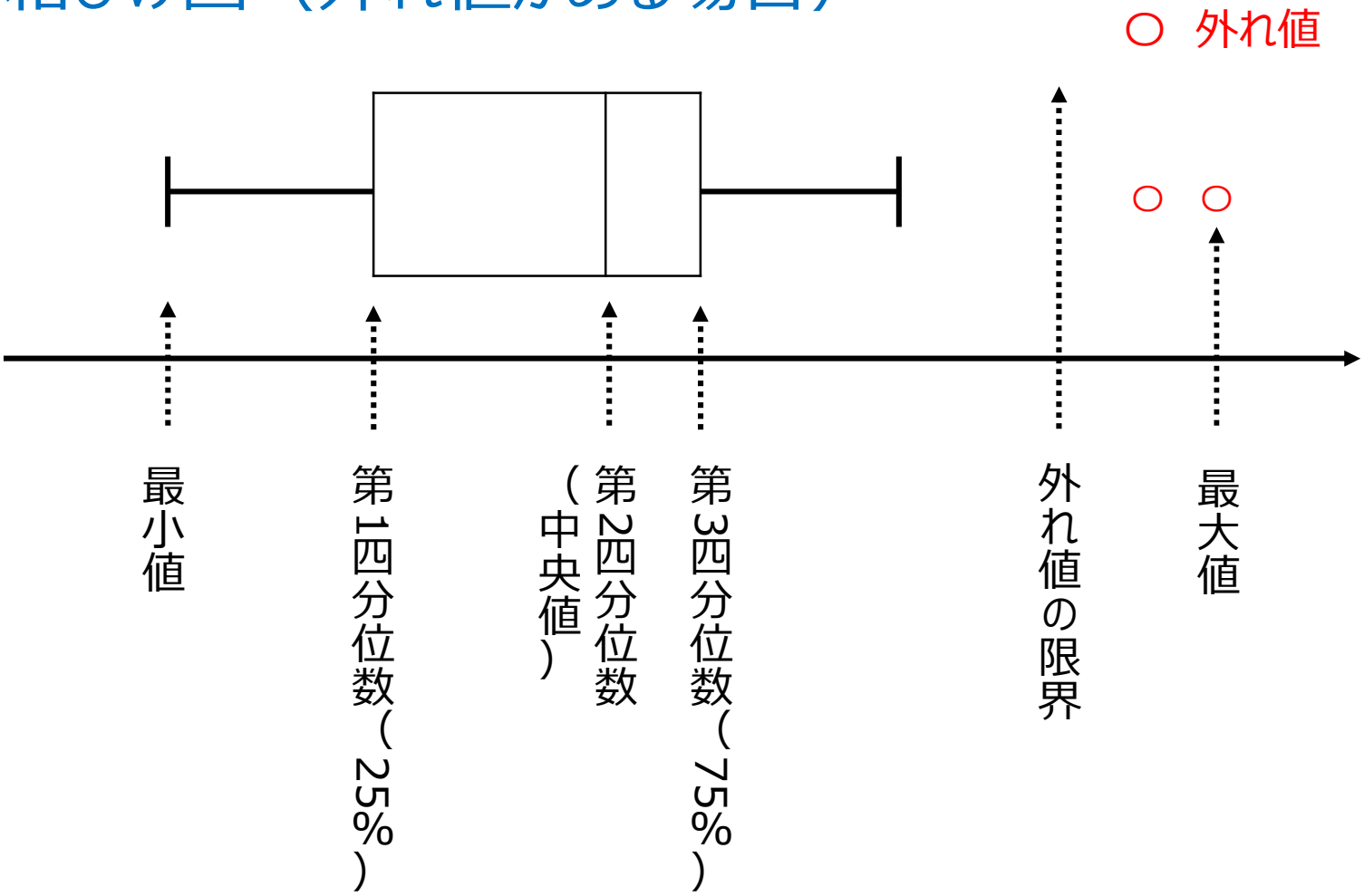


箱ひげ図

（外れ値がない場合）



箱ひげ図（外れ値がある場合）

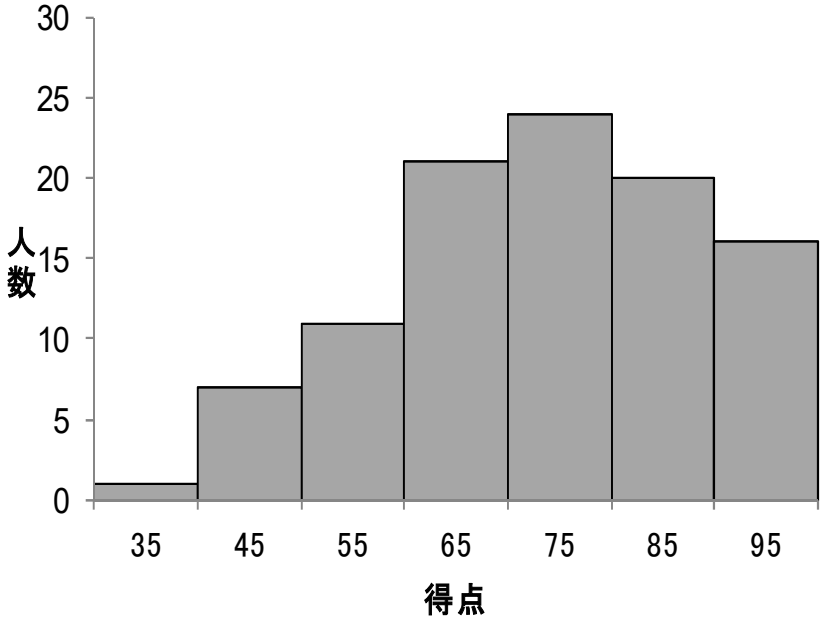


EG

度数分布表とヒストグラム

データ

下限		上限	階級値	度数
30	～	39	35	1
40	～	49	45	7
50	～	59	55	11
60	～	69	65	21
70	～	79	75	24
80	～	89	85	20
90	～	99	95	16



成績データA

88	88	60	86	86
48	86	60	46	40
64	84	70	34	62
78	36	77	72	76
65	68	90	90	56
72	54	54	52	66
68	68	60	74	72
78	56	62	96	90
36	86	48	48	90
48	68	84	66	46

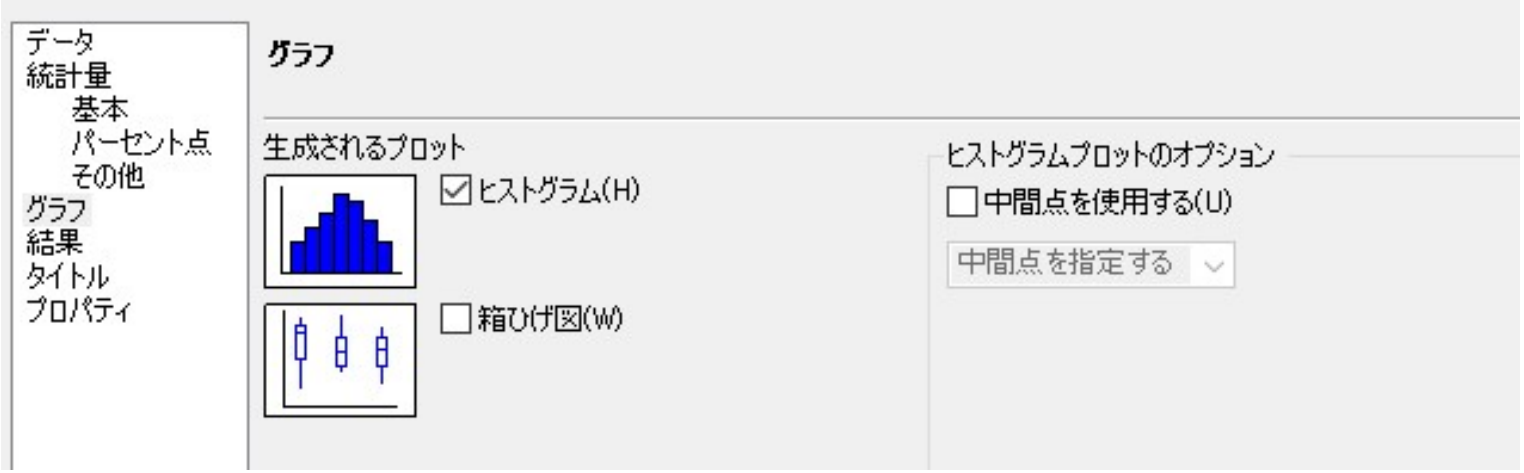
EG

ヒストグラムの作成（SAS EG）

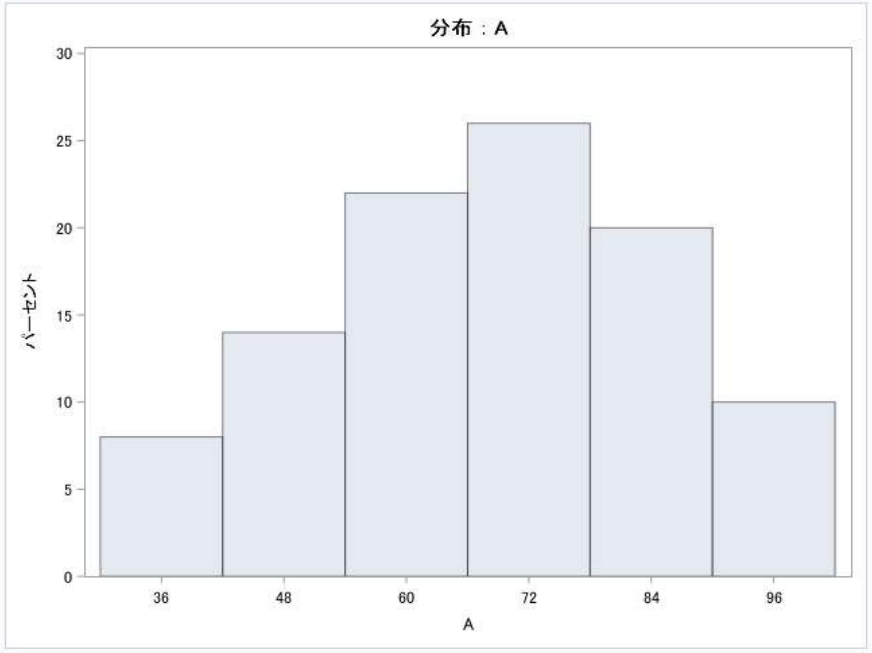
（1）データを入力する。

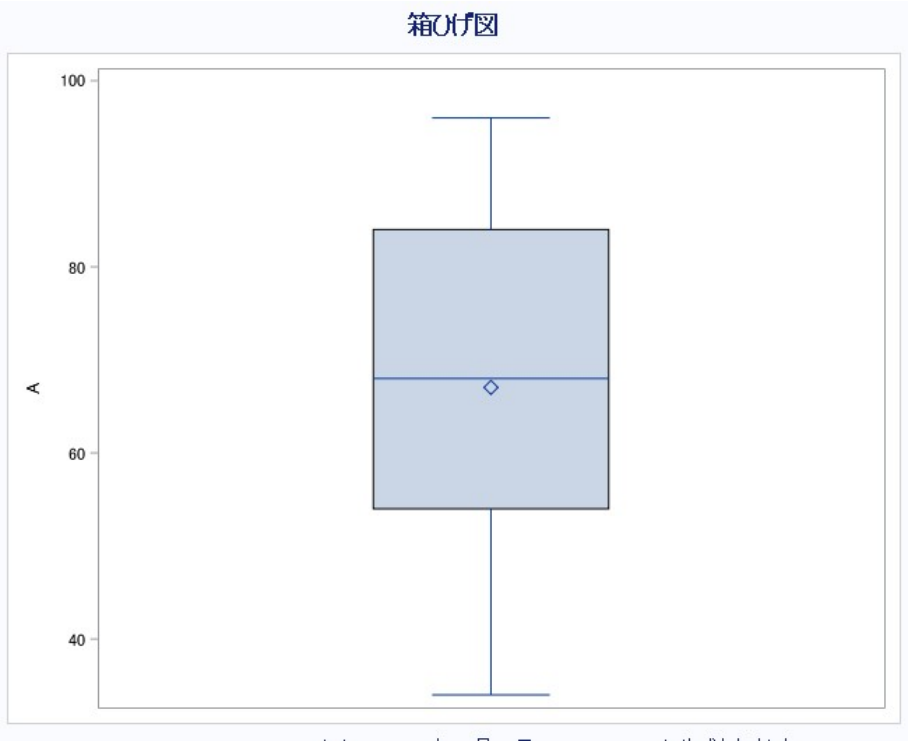
（2）「記述統計」－「要約統計量」をクリックし、変数リストから分析変数を指定する。

（3）「グラフ」をクリックし、「ヒストグラム」にチェックする。



EG





分析の基本的な考え方

英語の研修前後の成績（20人）

	A君の得点	全体の平均点	得点-平均点
研修前	70	58.3	+11.7
研修後	72	58.3	+13.7

A君の成績の評価は？

平均を評価基準とする合理性？

平均の信頼性

（例）暑い日に暑い場所で待ち合わせをした。
いつも遅れてくる人が何分後に来るのか予測。

	＜過去10回の遅刻データ（分）＞										平均
①	21	46	8	28	19	34	13	33	19	31	25.2分
②	25	26	23	24	26	27	26	25	24	26	25.2分

①、②それぞれにおける待ち時間の行動は？

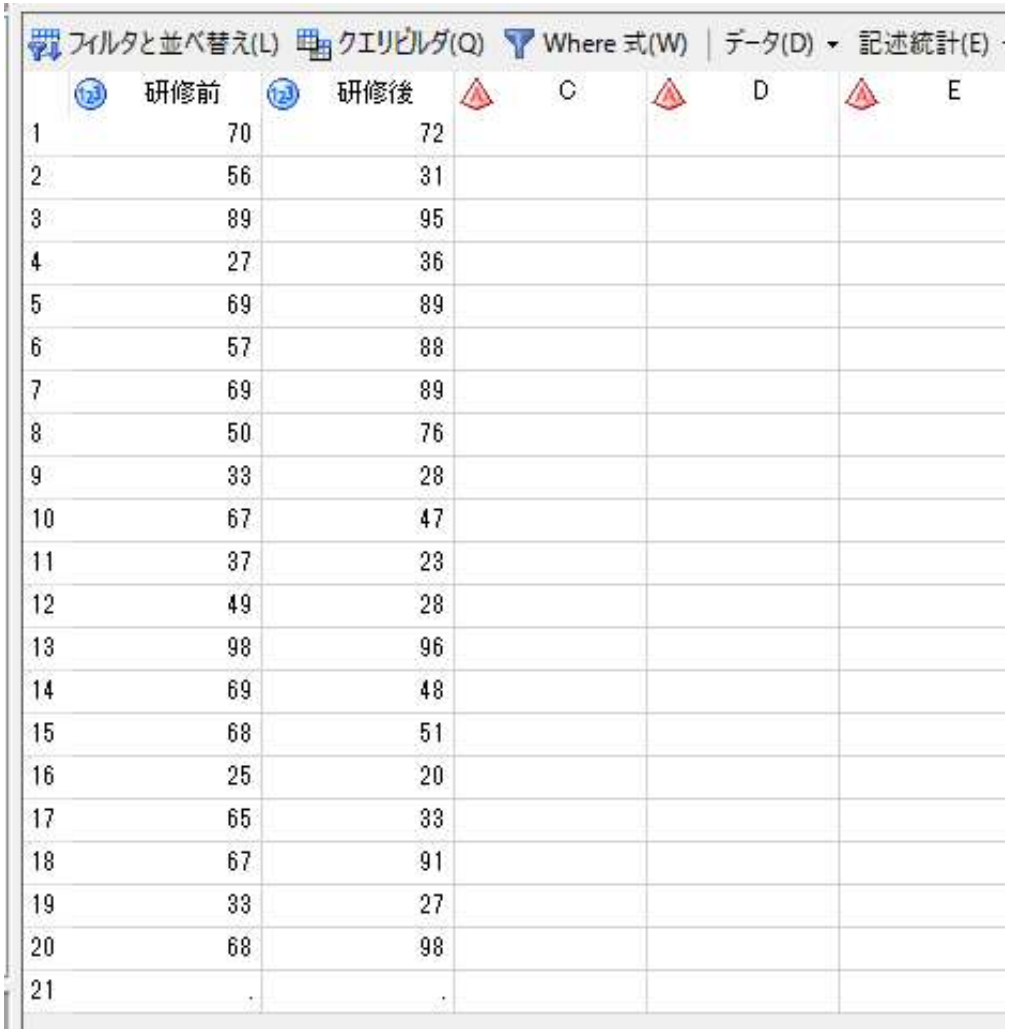
①のデータは「バラツキ」が大きい。

「バラツキ」の大きいデータの平均は信頼できない。

バラツキ　⇒　分散、標準偏差

EG

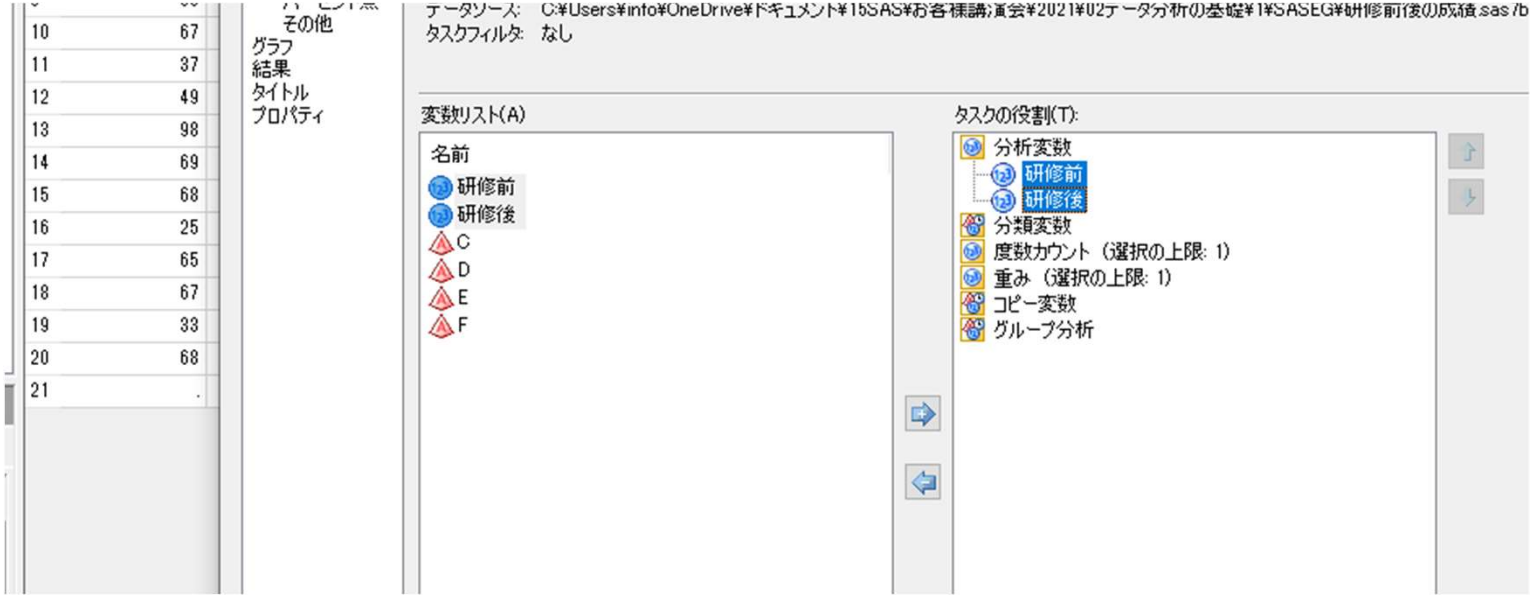
平均、分散、標準偏差を求める。（SAS EG）



	研修前	研修後	C	D	E
1	70	72			
2	56	31			
3	89	95			
4	27	36			
5	69	89			
6	57	88			
7	69	89			
8	50	76			
9	33	28			
10	67	47			
11	37	23			
12	49	28			
13	98	96			
14	69	48			
15	68	51			
16	25	20			
17	65	33			
18	67	91			
19	33	27			
20	68	98			
21	.	.			

EG

（2）「記述統計」–「要約統計量」をクリックし、変数リストから分析変数を指定する。



EG

（3）「統計量」－「基本」をクリックし、出力項目を設定する。

「平均」、「標準偏差」、「分散」、「オブザベーションの数」をチェックし、
「標準偏差と分散の除数」において「オブザベーションの数」を選択する。

要約統計量

結果

MEANS プロシジャ

変数	平均	標準偏差	分散	N	変動係数
研修前	58.3000000	19.1940095	368.4100000	20	32.9228293
研修後	58.3000000	28.6183508	819.0100000	20	49.0880802

A君の成績の評価

	成績	平均	成績－平均	標準偏差
研修前	70	58.3	11.7	19.2
研修後	72	58.3	13.7	28.6

研修前　　研修後

$$\frac{70-58.3}{19.2} = 0.609 > \frac{72-58.3}{28.6} = 0.479$$

$$\frac{\text{成績と平均値の差}}{\text{標準偏差}} \Rightarrow \text{Z値}$$

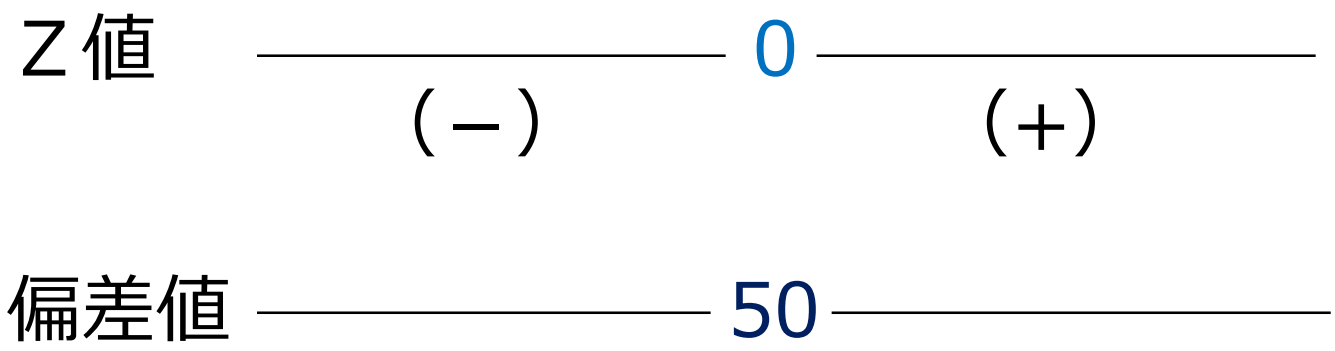
Z 値の比較によりデータの評価・比較が可能

Z 値×10＋50

偏差値

研修前の偏差値＝0.609×10＋50＝56.09

研修後の偏差値＝0.479×10＋50＝54.79



Z 値は、平均値＝0
平均値より大のときプラスの値、小のときマイナスの値

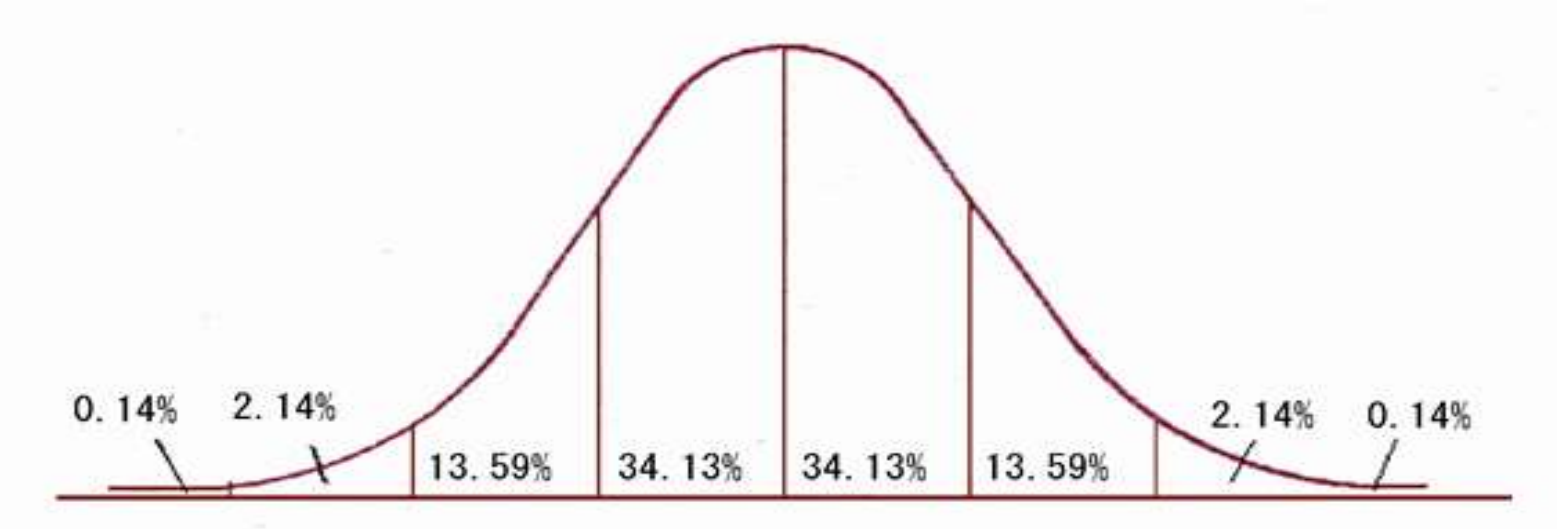
偏差値は、平均値＝50
平均値より大のとき50より大きな値、平均値より小のとき50より小さな値

Z値の特性

	データ	Z値	データ	Z値
	1	-1.4142	3	-1.1844
	2	-0.7071	8	0.55739
	3	0	6	-0.1393
	4	0.70711	4	-0.8361
	5	1.41421	11	1.60249
平均値	3	0	6.4	0
分散	2	1	8.24	1
標準偏差	1.41421	1	2.87054	1

Z値の平均値＝0　標準偏差＝1

標準正規分布（standard normal distribution）



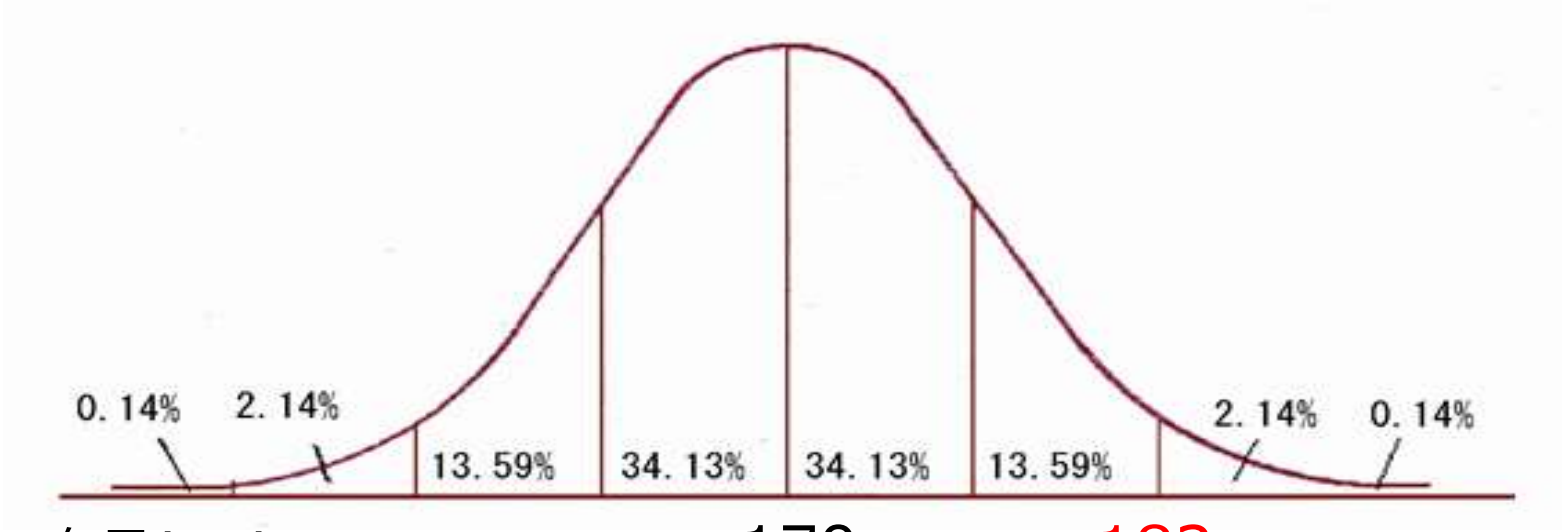
Z値	-3	-2	-1	0	1	2	3
偏差値	20	30	40	50	60	70	80

Ｚ値＝1（偏差値60） 以上の割合は約15.87%

Ｚ値＝2（偏差値70） 以上の割合は全体の約2.28%

標準正規分布の活用

■ 日本の成年男子の身長が平均170cm 標準偏差6cm、
182cm以上の人は、全体の何%を占めるか？



身長(cm)	170	182
Z値	0	2

$$Z = \frac{182\text{-}170\ (\text{cm})}{6\ (\text{cm})} = 2$$

182cm以上の人は約2.28%

■通信販売会社における顧客売上高の平均5,568円、
標準偏差2,216円、10,000円以上購入した顧客は、
全体の何%を占めるか？

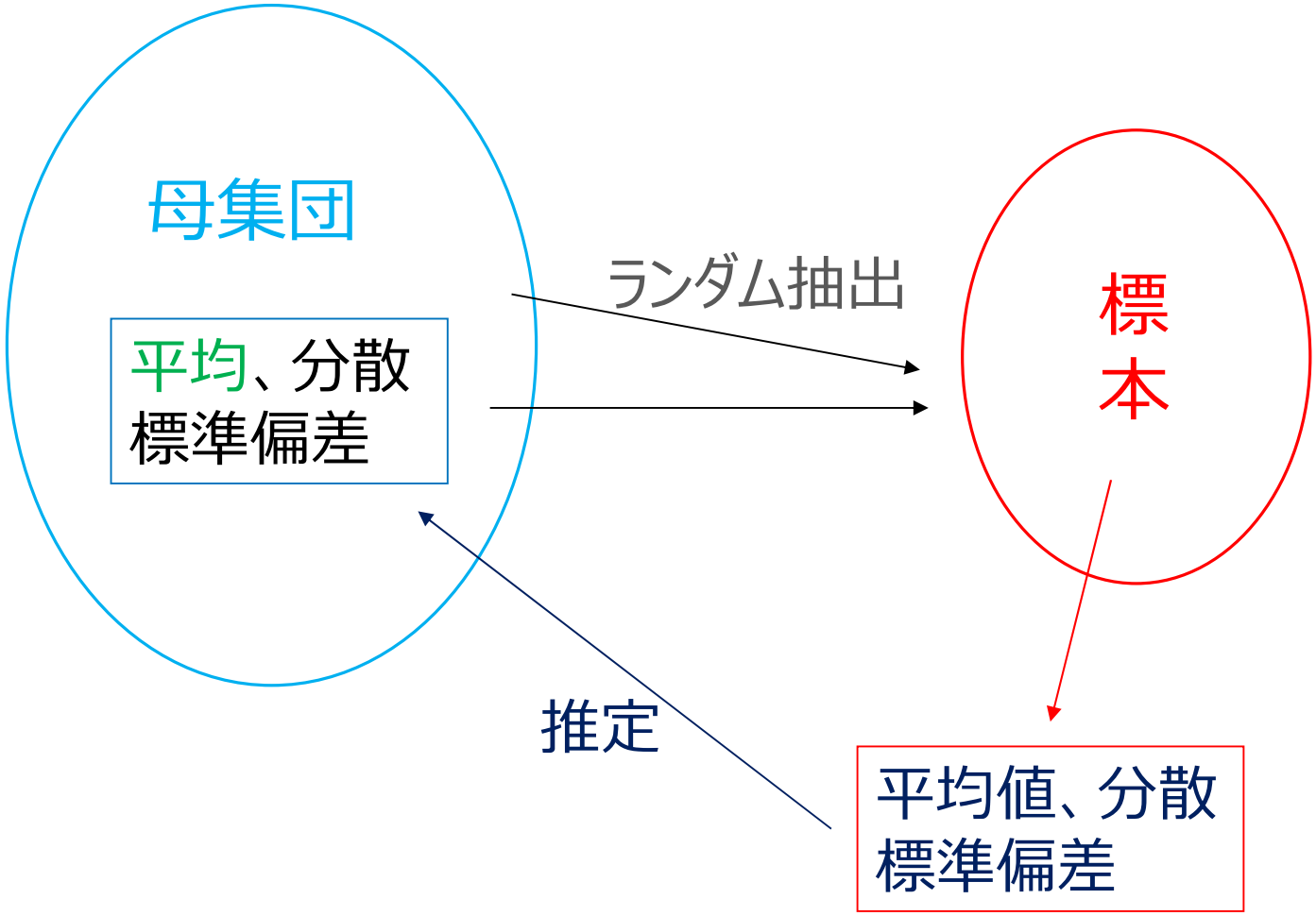
$$Z = \frac{10{,}000\text{-}5{,}568\ (円)}{2{,}216\ (円)} = 2$$

10,000円以上：約2.28%

Z値は単位とは無関係。

$$Z値 = \frac{求める値 - 平均値}{標準偏差}$$

母集団と標本

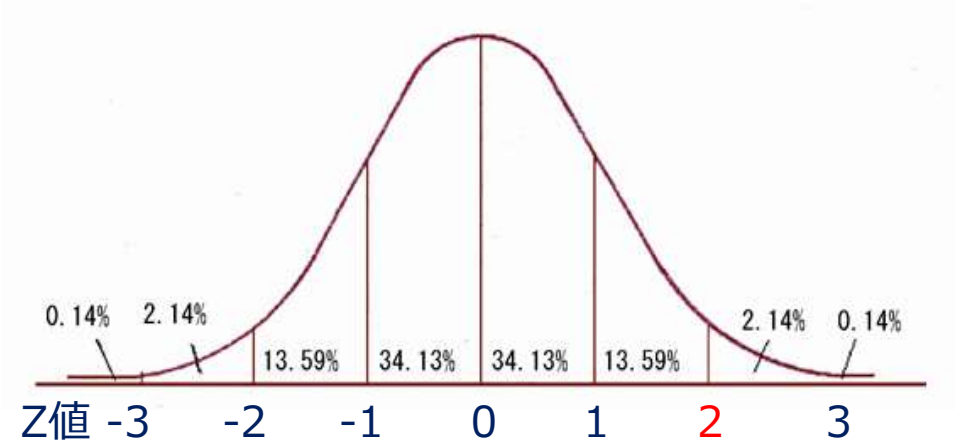


◇競合メーカー　有効成分含有量
平均＝60mg、 標準偏差＝5mg

市場から4つ購入した平均：65mg
60mgより多いのか？

P値（有意確率）を求める！

違うと判定したときの危険率



$$Z = \frac{65-60}{5} \times \sqrt{4} = 2$$

P値（有意確率）＝ 2.28%

サンプルサイズが大きくなると、Z 値も大きくなる

注）中心極限定理を参照

Z 値による検定 ⇔ 正規分布

Z 値による検定は標本数が少ないときに精度が良くない。

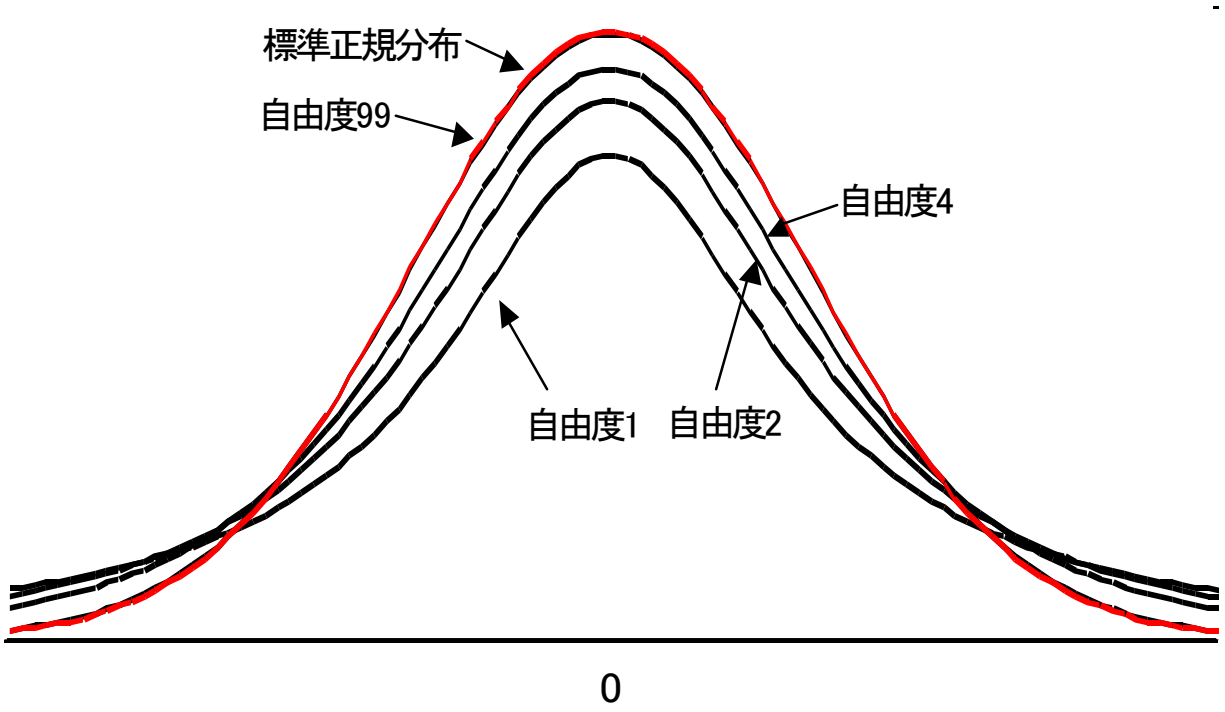
t 値による検定 ⇔ t 分布

t 分布とは

1908年、W.S.ゴセット（ギネスビールの研究員）が考案。
データが少ないときに精度を上げるために t 値を使う。

■ 全て（データの大小に関係なく） t 分布を使う。
■ この検定方法を t 検定という。

t 分布



自由度（ｎ-1）によって形が変わる。
自由度が大きくなると正規分布と同じ形。

$$\text{Z値、t 値} = \frac{\text{①差 (65-60)}}{\text{②バラツキ (5)}} \times \text{③}\sqrt{\text{n (4)}}$$

① 差が大 ⇒ t 値は大
② バラツキが小 ⇒ t 値は大
③ n（サンプルサイズ）が大 ⇒ t 値は大

t 値が大 ＝ P値（有意確率）は小
t 値が小 ＝ P値（有意確率）は大

1標本 t 検定

（A）ランダムに選んだ製品10個の有効成分含有量
60mg（母平均）と違いはみられるか？

64	62	58	61	60	59	63	66	62	63

10個の平均値を求めると
（64+62+・・・+60+63）÷10 ＝ 61.8

標本平均（61.8）と母平均（60.0）の比較

EG

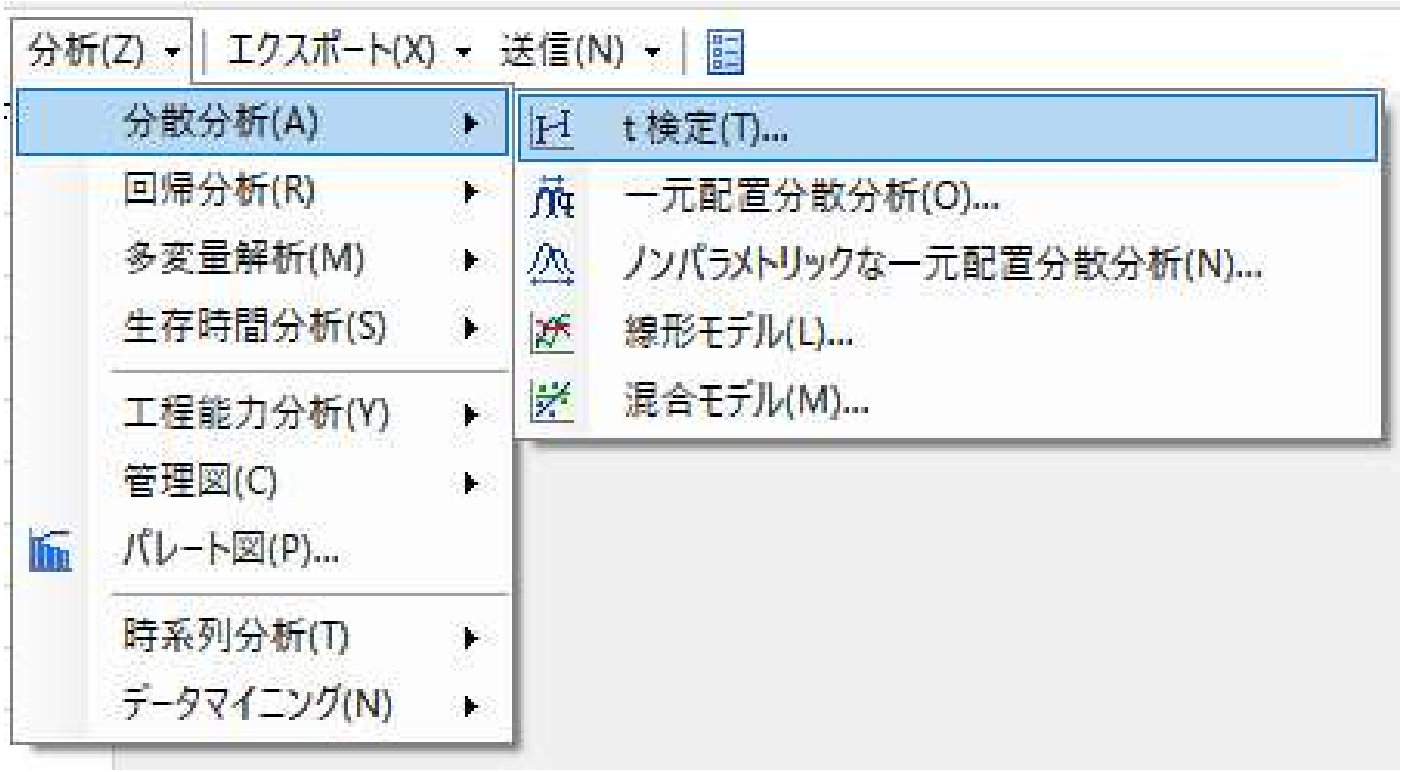
1標本 t 検定（SAS EG）

（1）データを入力する。

	含有量	B	C	D
1	64			
2	62			
3	58			
4	61			
5	60			
6	59			
7	63			
8	66			
9	62			
10	63			
11	.			
12	.			

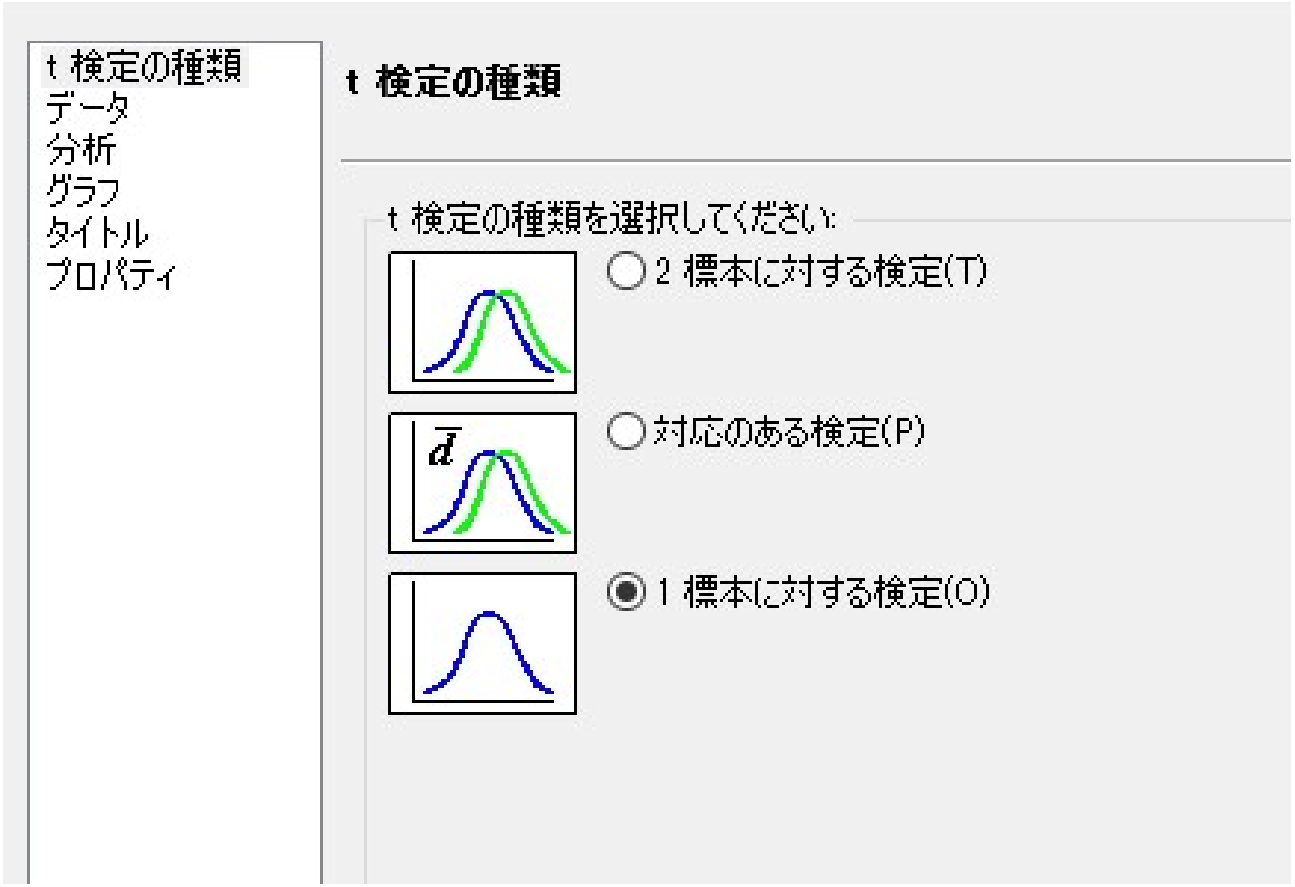
EG

（2）「分析」-「分散分析」-「ｔ検定」をクリックする。



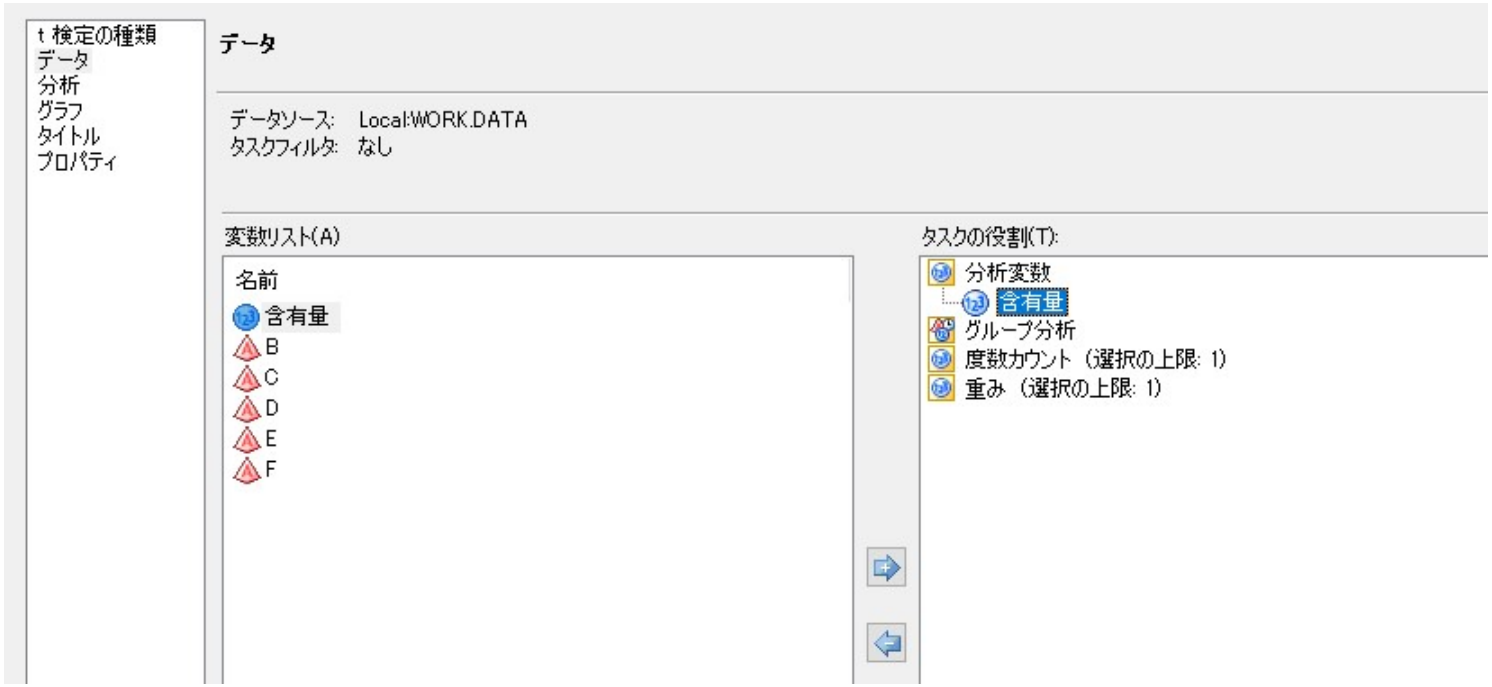
EG

（3）「1標本に対する検定」を選択する。



EG

（4）「分析変数」を指定する。



EG

（5）「分析」をクリックし、「母集団の平均」（帰無仮説）を入力する。

EG

t 検定

TTEST プロシジャ

変数 : 含有量

N	平均	標準偏差	標準誤差	最小値	最大値
10	61.8000	2.3944	0.7572	58.0000	66.0000

平均	平均の 95% 信頼限界		標準偏差	標準偏差の 95% 信頼限界	
61.8000	60.0871	63.5129	2.3944	1.6470	4.3713

自由度	t 値	Pr > |t|
9	2.38	0.0414

t 値 = 2.38

有意確率 = 0.0414

サンプルサイズが拡大した場合

（B）ランダムに選んだ製品20個の有効成分含有量
母平均：60ｍｇ　違いはみられるか？

64	62	58	61	60	59	63	66	62	63
64	62	58	61	60	59	63	66	62	63

20個の平均値を求めると
（64+62+・・・+60+63）÷20 ＝ 61.8

標本平均（61.8）と母平均（60.0）の比較

EG

t 検定

TTEST プロシジャ

変数： 含有量3

N	平均	標準偏差	標準誤差	最小値	最大値
20	61.8000	2.3306	0.5211	58.0000	66.0000

平均	平均の 95% 信頼限界		標準偏差	標準偏差の 95% 信頼限界	
61.8000	60.7093	62.8907	2.3306	1.7724	3.4040

自由度	t 値	Pr > |t|
19	3.45	0.0027

t 値 ＝ 3.45 ＞ 2.38

P値（有意確率）＝ 0.0027 ＜ 0.0414

効果量（Cohen’s d）

t値（P値）はサンプルサイズの影響を受ける

$$t\text{ 値} = \frac{\text{差}}{\text{バラツキ}} \times \sqrt{n}$$

$$= \text{効果量} \times \sqrt{n}$$

1標本 t 検定　効果量の比較

	P値	t 値	n	効果量
(A)	0.0413	2.38	10	0.75
(B)	0.0073	3.45	20	0.77

◇効果量（Cohen’s d）の大きさの評価

0.2	0.5	0.8
小	中	大

2標本 t 検定

（A）新製品の好感度について、男女別各10人に
10点満点にて調査した。
男女間の評価に違いは見られるか？

2020年

										平均
男性	7	3	8	5	9	2	5	7	6	9	6.1
女性	6	4	7	3	6	3	6	6	4	8	5.3

◇ 1 標本 t 検定

$$t\text{ 値} = \frac{\text{差}}{\text{バラツキ}} \times \sqrt{n}$$

◇2標本 t 検定

$$t\text{ 値} = \frac{\text{差（男女の平均値の差）}}{\text{男性のバラツキ+女性のバラツキ}} \times \sqrt{n}$$

EG

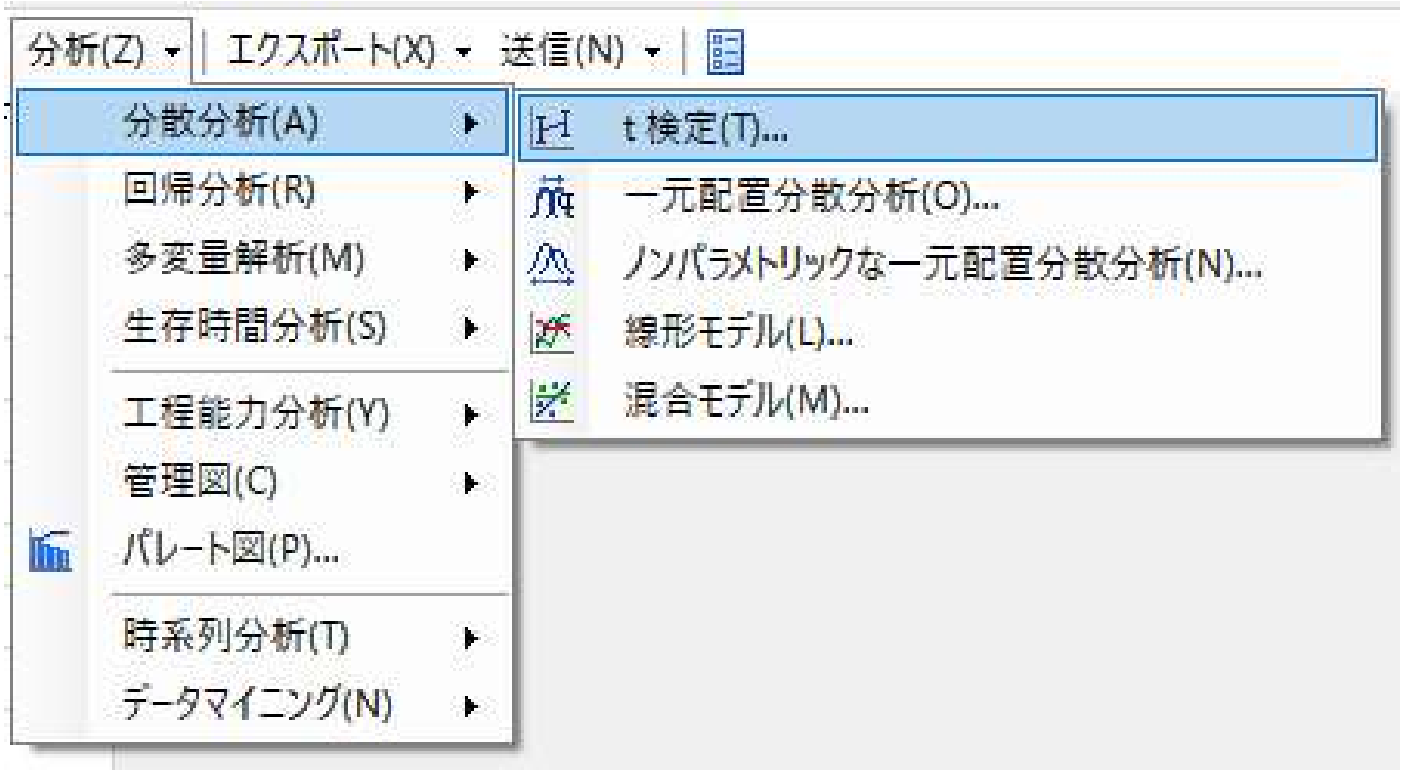
2標本 t 検定（SAS EG）

（1）データを入力する。

	性別	評価	C	D
1	m	7		
2	m	3		
3	m	8		
4	m	5		
5	m	9		
6	m	2		
7	m	5		
8	m	7		
9	m	6		
10	m	9		
11	f	6		
12	f	4		
13	f	7		
14	f	3		
15	f	6		
16	f	3		
17	f	6		
18	f	6		
19	f	4		
20	f	8		
21		.		

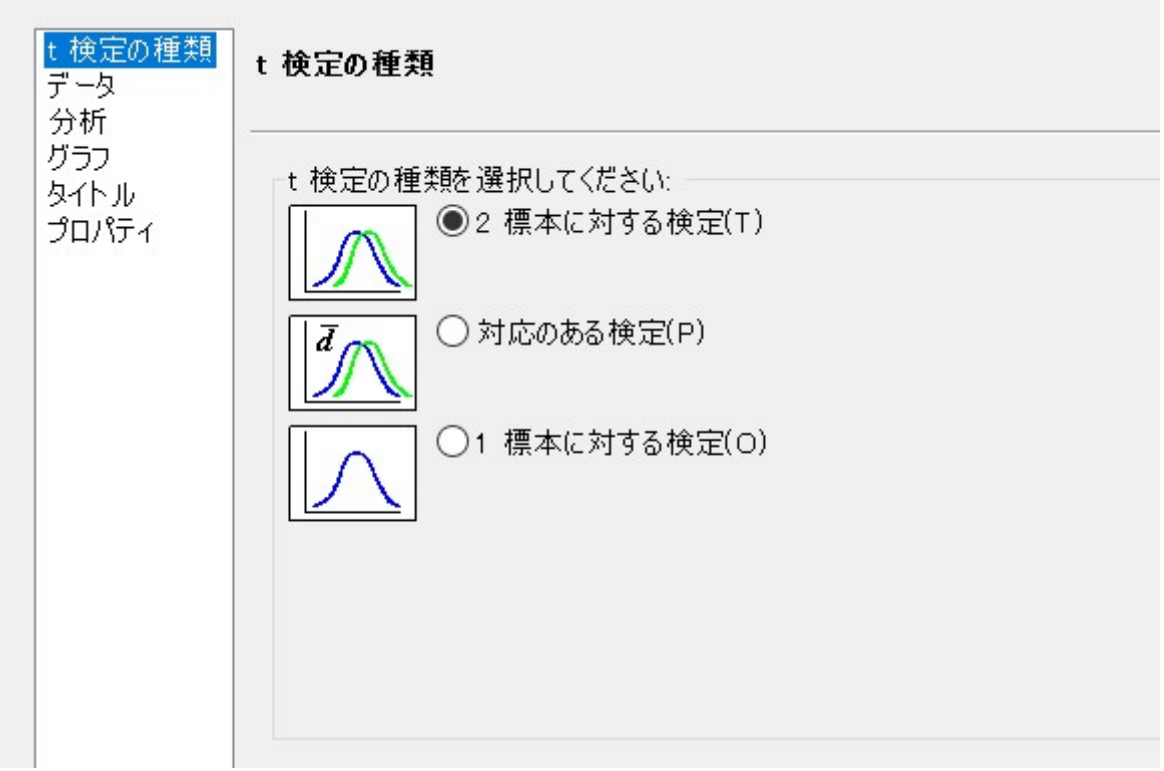
EG

（2）「分析」-「分散分析」-「ｔ検定」をクリックする。



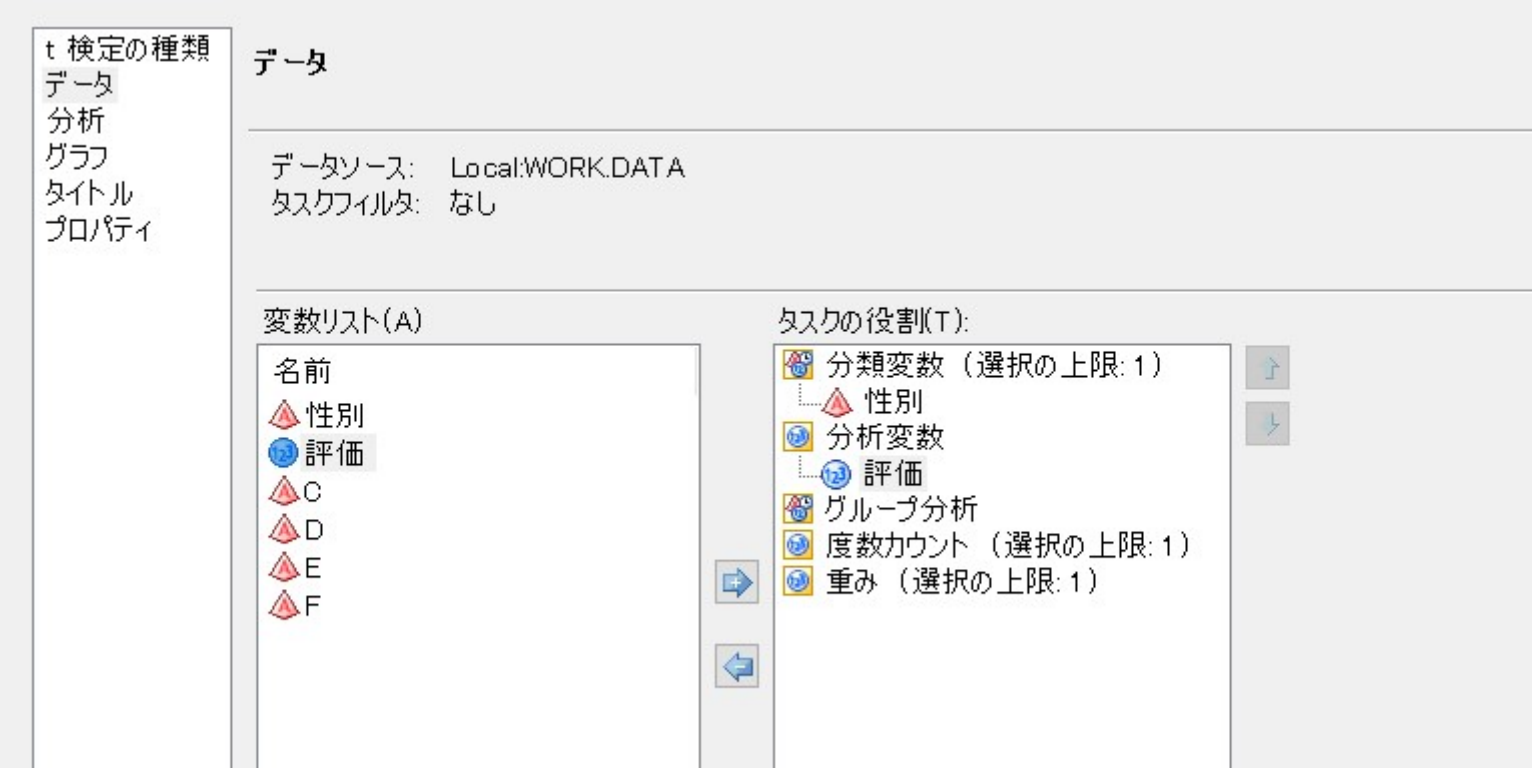
EG

（3）「2標本に対する検定」を選択する。



EG

（4）「分類変数」と「分析変数」を指定する。



EG

手法	分散	自由度	t 値	Pr > |t|
Pooled	Equal	18	-0.86	0.3985
Satterthwaite	Unequal	16.308	-0.86	0.3996

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	9	9	1.95	0.3341

◇分散が違うとは言えないとき
P値（有意確率） ＝ 0.3985

◇分散が違うとき
P値（有意確率） ＝ 0.3996

＜男女別好感度調査　2標本 t 検定の結果＞

P値（有意確率）＝ 0.3996 ＞ 0.05（有意水準）

有意水準5%において、男女間の好感度に違いが見られるとは言えない。
（有意水準5%において、有意ではない。）

サンプルサイズが拡大した場合

（B）新製品の好感度について、男女別各30人に
10点満点にて調査した。
男女間の評価に違いは見られるか？

男性	7	3	8	5	9	2	5	7	6	9	5.3
	7	3	8	5	9	2	5	7	6	9	
	7	3	8	5	9	2	5	7	6	9	
女性	6	4	7	3	6	3	6	6	4	8	6.1
	6	4	7	3	6	3	6	6	4	8	
	6	4	7	3	6	3	6	6	4	8	

男女各10人から男女各30人に増加

EG

手法	分散	自由度	t 値	Pr > |t|
Pooled	Equal	58	-1.55	0.1260
Satterthwaite	Unequal	52.549	-1.55	0.1265

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	29	29	1.95	0.0772

◇分散が違うとは言えないとき
P値（有意確率）＝ 0.1260

◇分散が違うとき
P値（有意確率）＝ 0.1265 ＜ 0.3996

効果量（Cohen’s d）

◇2標本 t 検定

$$t\text{ 値} = \frac{\text{差（男女の平均値の差）}}{\text{男性のバラツキ+女性のバラツキ}} \times \sqrt{n}$$

$$= \frac{\text{差（男女の平均値の差）}}{\text{男性と女性の平均的なバラツキ}} \times \sqrt{n/2}$$

$$= \text{効果量} \times \sqrt{n/2}$$

2標本 t 検定　効果量の比較

	P値	t 値	n	効果量
(A)	0.3996	0.86	10	0.38
(B)	0.1265	1.55	30	0.40

◇効果量（Cohen’s d）の大きさの評価

0.2	0.5	0.8
小	中	大

「違い」と「効果」

「違いの大きさ」はサンプルサイズの影響を受ける。

ビッグデータにおいては
「違いがある！」という分析結果が出やすい。

違いがある ≠ 効果が大きい

サンプルサイズと効果量を検討することが重要！

まとめ

１ データの代表値
　平均・バラツキ（分散・標準偏差）
　中央値
２ データの視覚化
　箱ひげ図、ヒストグラム
３ 正規分布、ｔ 分布
４ ｔ 検定
　検定のしくみ
　P値（有意確率）、有意水準
　効果量

アンケートのお願い・ご質問

7月14日 データ分析の基礎-1

今後の参考にさせていただくため、ぜひともアンケートにご協力をお願いします。

- 無記名
- 所要時間目安: １～３分

アンケートURL

https://sas.qualtrics.com/jfe/form/SV_3vMgvkw3ivBQKiy

・お客様講演会のアーカイブは、2021年7月19日～2022年3月31日迄視聴できます。

・本日の内容に関するご質問は、以下宛にご連絡ください。

que@datascience.co.jp

ご視聴ありがとうございました。