基礎セミナー
じっくり勉強すれば身につく統計入門

「じっくり勉強すれば身につく統計解析」を副題としたシリーズ全 3 巻がサイエンティスト社から刊行されました．タイトルは「医薬品開発のための統計解析，第 1 部 基礎，第 2 部 実験計画，第 3 部 非線形モデル」です．「じっくり勉強すれば身につく統計 ～ Excel，JMP による基礎から応用統計解析実務者コース」（SAS（株）JMP ジャパン事業部主催，年 12 回）のテキストとして使用されています．定例会に先立って，この本をベースに「基本に戻ろうー基本統計量とデータの比較ー」を開催いたします．統計をじっくりと勉強して身に付けたいと思われる方々の参加をお待ちしています．

基本に戻ろう 基本統計量とデータの比較

橘田 久美子（スギ生物科学研究所）
杉本 典子 （バイオスタティスティカルリサーチ）

東京，横浜，静岡，名古屋，京都，大阪，神戸，広島に住む 8 人の同級生が同窓会を開きたい．同級生の移動距離が最小になる集合場所はどこでしょうか？ どのような統計量となるのでしょうか．どのように定式化すればいいのでしょうか．最小 2 乗法で平均値を求めたい．どのように計算するのでしょうか．算術平均値に等しくなるのでしょうか？ 基本に戻って考えて見ましょう．Excel 上にあるデータから直接ヒストグラム書きたいのだけれども，どうしたらいいのだろうか？ データを箱ひげ図で作りたいのだが，Excel ではできそうもないが，どうしたらよいのだろうか．基本に戻ってといっても，どのように学習したらいいのだろうか．

じっくり勉強すれば身につく 統計学入門

基本に戻ろう 基本統計量とデータの比較

スギ生物科学研究所株式会社
橘田久美子

1. 基本統計量とは?

基本統計量 (Fundamental Statistics)

データ (n個の観測値) の特長を表す数値

：代表値：
[データ集団を代表する]

- 平均値
- 中央値

：ばらつき：
[データの変化の大きさを表す]

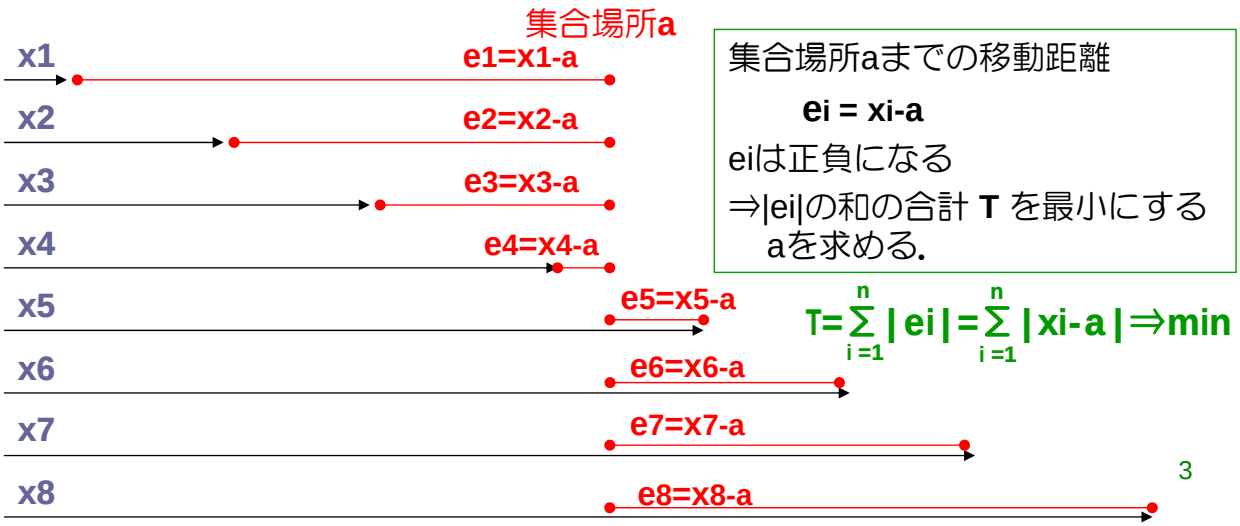
- 平方和
- 平均平方
- 標準偏差
- 変動係数
- 四分位値
- 四分位範囲
- ひずみ・とがり

(1)中央値

東京, 横浜, 静岡, 名古屋, 京都, 大阪, 神戸, 広島に住む8人の同級生が同窓会を開きたい.

⇒同級生の移動距離が最小になる集合場所はどこ？

東京　横浜　静岡　名古屋　京都　大阪　神戸　広島



集合場所aまでの移動距離

$e_i = x_i - a$

eiは正負になる

⇒|ei|の和の合計 **T** を最小にする aを求める.

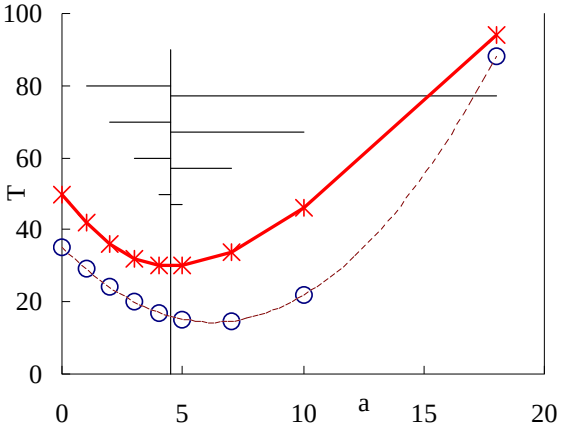
$$T=\sum_{i=1}^{n}|e_i|=\sum_{i=1}^{n}|x_i-a|\Rightarrow \min$$

(1)中央値

東京・横浜…という場所の代わりに i =1, 2, …8 とし, aを0～18まで変化させたとき, T の変化はグラフの赤折れ線のようになる.

a が 4～5の間のとき T は最も小さくなる.

i	x
1	1
2	2
3	3
4	4
5	5
6	7
7	10
8	18
中央値	4.5
n	8



T が最小になるのは

nが偶数：中央の2つの値の間

nが奇数：中央の値

⇒**中央値(Median)** $\tilde{x}$

中央値は観測値が変化した場合でも影響を受けにくい,という性質がある.

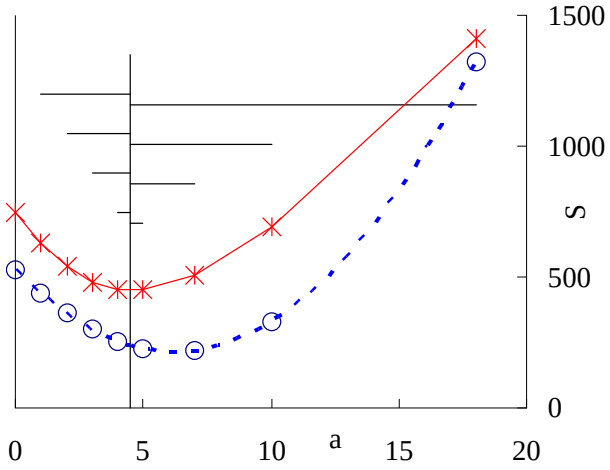
⇒頑健性がある

EXCEL関数

中央値 = MEDIAN(データ範囲)

(2) 平均値

中央値の場合は観測値 xi と中央値 a の差 ei を ei = | xi – a |と絶対値で考えた.
ei は2乗すると符号が + になる.
⇒ ei^2 の合計が最小になる a の位置を S で表す. $S=\sum_{i=1}^{n}(ei)^2=\sum_{i=1}^{n}(xi-a)^2 \Rightarrow \min$
S の変化はグラフの青点線のようになる

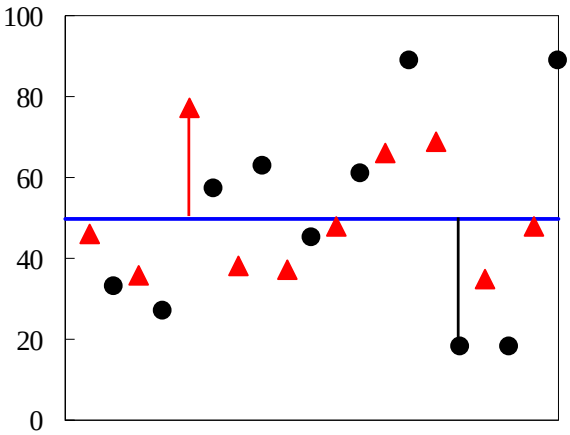


Sが最小になるときの a が
⇒ **平均値 (Mean)**
また, この算出方法が
最小2乗法 (Least Squares Methods)
である.

EXCEL関数
平均値 = AVERAGE(データ範囲)

誤差が**正規分布に従う**ときは最小2乗法による推定値である**平均値**を代表値として使用するのが最適であると数理統計学で証明されている.

(3) 平方和



平均値 (—)が同じ2つのデータ集団 (●▲)がある.
この2つの集団の違いをどのように表現したらよいだろうか？
グラフ化して観察すると, ●の方が▲に比べ各観測値が平均値から離れているように見える.
⇒個々の観測値xiが平均値からどれだけ離れているか, ばらつきの違いで2つの集団の違いを表せそう.

平均値を求めたときの式 $S=\sum_{i=1}^{n}(ei)^2=\sum_{i=1}^{n}(xi-a)^2 \Rightarrow \min$

の a に平均値 $\bar{x}$. を代入すると $S=\sum_{i=1}^{n}(xi-\bar{x}.)^2=\sum(ei)^2$

ei = (xi - $\bar{x}$.)：偏差 (Deviation)
S ：偏差平方和，又は平方和(Sum of Squares)と呼び
ばらつきの総量をあらわす.

(4) 平均平方・標準偏差

平方和はデータ数が多くなる (n数が増える)と大きくなるのでばらつきの大きさの比較には使えない.
→データ1個当たりのばらつきで比較すればよい.

平方和 S は ei^2 を n 個加えたもの
→データ1個当たりのばらつきにするには,
平方和を **(n-1)**(=自由度, Degree of Freedom, ν) で割る.

$V=S/(n-1)= S/\nu$ ←平均平方 (Mean Square), 分散 (Variance)

S (平方和) は ei^2 (偏差)^2 だった → 平均平方の単位も観測値の2乗
観測値と単位を揃える (単位を戻す) には**平方根**を取ればよい

$SD = \sqrt{V}$ ←平均平方に平方根を取った値が
標準偏差 (Standard Deviation, sd, SD)

7

(5) 変動係数

標準偏差は n 数に影響されずに平均的なばらつきの大きさを定量的に表すことができる.
しかし, 単位が異なる量の標準偏差を比較するのは不適当な場合がある.

<例1>
あるグループの身長の標準偏差が7 cm, 体重の標準偏差が5 kgであった.
どちらのばらつきが大きいか？

<例2>
20歳の成人の身長の標準偏差は12 cm, 5歳児の身長の標準偏差は3 cmであった.
どちらのばらつきが大きいか？

このような場合には, **標準偏差と平均値の比**を用いるとよい.

$CV = S / \bar{x}$ ←変動係数(Coefficient of Variation)
100を掛けて%表示することも多い

!! 温度のように, 摂氏(℃)と華氏(°F)で0点の取り方に任意性がある量では変動係数は意味を持たない.
<例> [37±10℃] vs. [100±18°F]を比較しても意味がない
また, データに負の値が含まれるときは変動係数を用いることは好ましくない.

8

(6) 四分位値と四分位範囲

n個のデータを大きさ順に並べたときのデータを2分する値
=**中央値** (Median)
中央値で分割された下半分/上半分を2分してデータ集団を4分割する値
= **四分位値(Quartile)**

nが奇数の場合 (n=5)

EXCEL

中央値

x(1) < x(2) < x(3) < x(4) < x(5)

下側四分位値　上側四分位値

JMP

中央値

x(1) < x(2) < x(3) < x(4) < x(5)

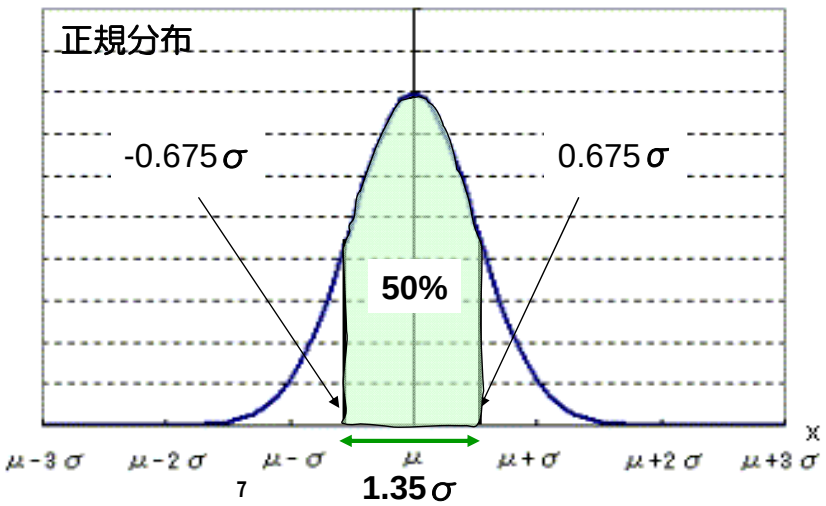
下側四分位値 x(1)とx(2)の間　上側四分位値 x(4)とx(5)の間

nが偶数の場合
プログラムによって異なるので注意!!

上側四分位値と下側四分位値の間に**ほぼ半分のデータが含まれる**
= 四分位範囲 = 上側四分位値－下側四分位値

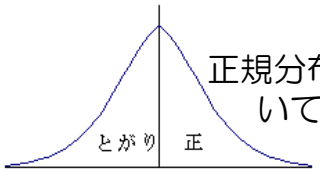
(6) 四分位値と四分位範囲

データが正規分布に従うとき，データの50%が含まれる範囲は
$\mu \pm 0.675\sigma$
→ 四分位範囲≈ $0.675\sigma \times 2 = 1.35\sigma$となるはず
⇒ **四分位範囲÷1.35 = 標準偏差(σ)の推定値(= s)**
この推定値は外れ値の影響を受けにくい.
(頑健性がある)



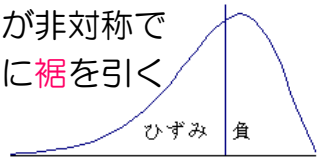
(7) ひずみ・とがり

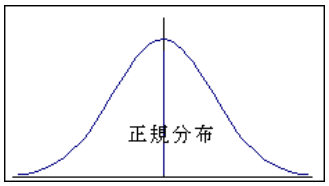
データが正規分布に従うとき，分布は**平均**と**標準偏差**だけで決まる.
しかし，現実のデータでは正規分布が当てはまるとは限らない.
正規分布からの外れを
ひずみ(歪度, Skewness)
とがり (尖度, Kurtosis)
で表すことができる.



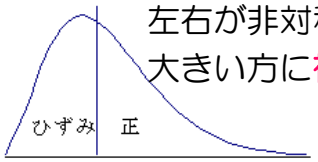
正規分布よりも頭が尖っていて**裾が長い**

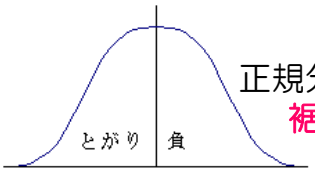
左右が非対称で小さい方に裾を引く





左右が非対称で大きい方に裾を引く





正規分布よりも頭が丸く**裾が短い**

EXCEL関数
ひずみ = SKEW (データ範囲)
とがり = KURT (データ範囲)

11

(7) ひずみ・とがり

母集団が正規分布であっても, サンプルのひずみ・とがりは 0 からかなり振れた値をとることがある.
nが数十程度以下のとき，正規分布から外れているかどうかは
ひずみ・とがりの値が**±1.5以内**かどうかを目安とし,_この範囲を超えたときは，以下の点について検討する.

ひずみ，とがりの絶対値が大きくなるのは,

a) 分布が全体として**正規分布から外れる**場合
b) 少数個の**外れ値が含まれる**場合

⇒ ヒストグラムや箱ひげ図で確認する.

12

(8) 度数表とヒストグラムの作成

EXCELファイルの『ヒストグラム』シートの ALTのデータを使用して
度数表・ヒストグラムの作成します.

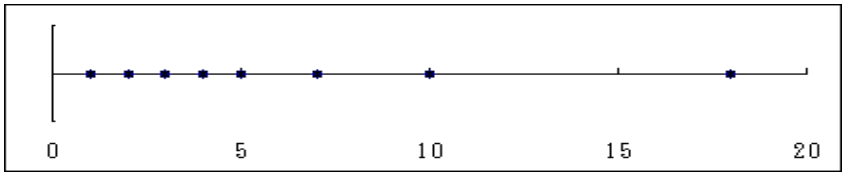
ALTのデータ　　⇒10前後に分割しよう！
　最大値：242.4　　221.0÷10 = 22.1
　最小値：21.4　　→20 間隔で分けることにする.
　差　　：221.0　　境界値は『〇以上△未満』とする

〇と△の間に当てはまる数が度数
度数の和が累積度数　　→　最後の区間の累積度数 = データ数

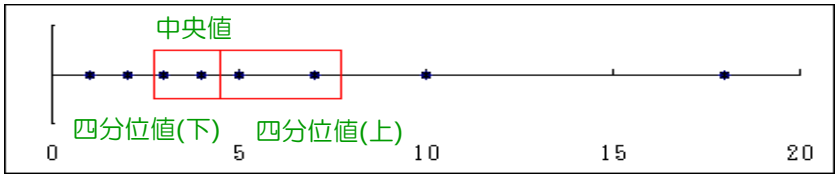
EXCEL関数
　累積度数　　= FREQUENCY(データ範囲, 上限値)
　度数　　= 累積度数ー前の累積度数

(9) 箱ひげ図の作成

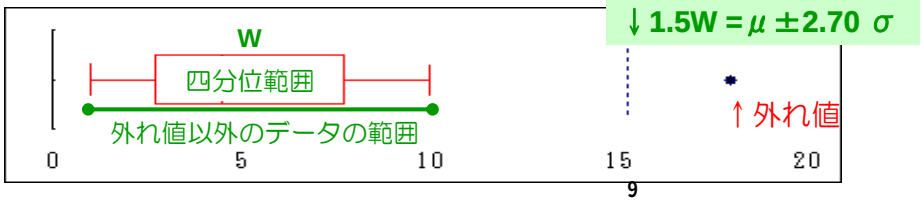
1. 観測値を一列にプロットする.



2. 中央値と四分位値を追加して箱を書く



3. 箱の中の点を消し, 外れ値以外のデータ範囲をひげで表す.
四分位範囲の幅から1.5倍外れたところに点線を引く(外れ値の目安).



(9) 箱ひげ図の作成

箱の幅Wは四分位範囲に相当
データが正規分布に従うとき，四分位範囲は1.35σの推定値となる.
箱の幅Wから1.5倍離れた値は
$\mu \pm (1.35/2+1.5\times1.35)\ \sigma = \mu \pm \mathbf{2.70}\ \sigma$
となる.
正規分布がこの範囲を外れる確率は
{1-NORMSDIST(2.70)}×2 = 0.00693…
より約1%となる.
よって，箱の両端から1.5W離れた位置に点線を引き，
その外側の値を外れ値(Outlier)とする.

EXCELファイルの『箱ひげ図』シートで箱ひげ図を作成します.
(使用データは ALT)

15

(10) 外れ値の取扱い

外れ値が生じたとき, Smirnovの棄却検定などで検定を行なう方法もある.
しかし, 外れ値とは**全体の分布から離れている**というだけであって**異常値と断定したことではない**ことを意味している.

異常かどうかは統計理論では判断できない.

得られたデータの外れ値は, 特異体質や併用薬の影響等の原因による場合がある.

外れ値を機械的に削除して解析することは厳に慎むべきである.

外れ値であるかどうかについては, **観測値が正規分布に従う**という仮定の元で導かれたものである.
したがって, 分布が歪んでいるときは異常がなくても外れ値になりやすい.
分布にゆがみがある場合は, **正規分布になるように変換してから適用**すべきである.

16

(11) 分布がゆがんでいるとき

ひずみが正の値を取り，とがりが大きく分布が全体として上の方に裾を引いている場合 (例えば，臨床検査項目のALT，ASTなど) は**測定値 (観測値) を対数変換**することで正規分布に近づくことが多い．

このような分布を対数正規分布という．

対数変換前

	ALT	AST
平均	55.73	35.84
標準偏差	48.08	9.82
変動係数	0.86	0.27
ひずみ	2.74	1.16
とがり	7.56	1.37
最小値	17.80	23.10
第1四分位値	30.55	28.35
中央値	41.10	33.80
第3四分位値	59.40	41.25
最大値	242.40	65.70

対数変換によりひずみ・とがりが小さくなって正規分布に近づいたことが判る．

対数変換前

	ALT	AST
平均	3.81	3.55
標準偏差	0.59	0.25
変動係数		
ひずみ	1.23	0.54
とがり	1.57	-0.15
最小値	2.88	3.14
第1四分位値	3.42	3.34
中央値	3.72	3.52
第3四分位値	4.08	3.72
最大値	5.49	4.19

17

まとめ

- データが得られたらまず**基本統計量**を計算し，**グラフ**に表して丁寧に観察する．
- 得られたデータがどのような形をしているのか，を知る方法として
  ヒストグラム
  箱ひげ図
  を描くことは非常に有用である．
- 得られたデータに外れ値がある場合，外れ値の取扱いには慎重を期すべきである．
  外れ値は異常値とは違う．
- 臨床検査値のように，ひずみが正の値をとり分布が全体として上の方に裾を引く場合，測定値を**対数変換**すると正規分布に近づくことが多い．
  対数変換により正規分布に近づくような分布を対数正規分布という．

18

データ

	ALT	AST
1	49.4	41.7
2	41.7	34.4
3	63.0	23.4
4	17.8	28.9
5	66.0	46.6
6	36.3	37.3
7	34.2	33.4
8	30.1	32.2
9	35.2	23.5
10	62.6	47.6
11	28.7	33.4
12	51.5	46.3
13	39.6	37.6
14	35.1	25.5
15	23.1	44.8
16	33.1	35.9
17	41.1	28.0
18	31.8	45.7
19	51.7	35.5
20	42.8	34.7
21	45.6	29.3
22	24.2	36.5
23	30.5	32.2
24	23.8	25.6
25	226.0	63.4
26	49.0	29.1
27	44.8	28.1
28	37.3	33.4
29	38.5	27.1
30	242.4	47.1
31	23.8	30.7
32	27.5	28.6
33	56.2	33.8
34	50.0	24.8
35	43.6	28.0
36	63.7	37.4
37	161.0	65.7
38	90.3	48.4
39	30.6	26.6
40	26.2	23.1
41	71.5	45.9
42	31.2	40.8
43	64.9	35.1
44	21.4	31.1
45	162.2	35.5
46	89.5	52.8
47	28.7	27.8

基本統計量

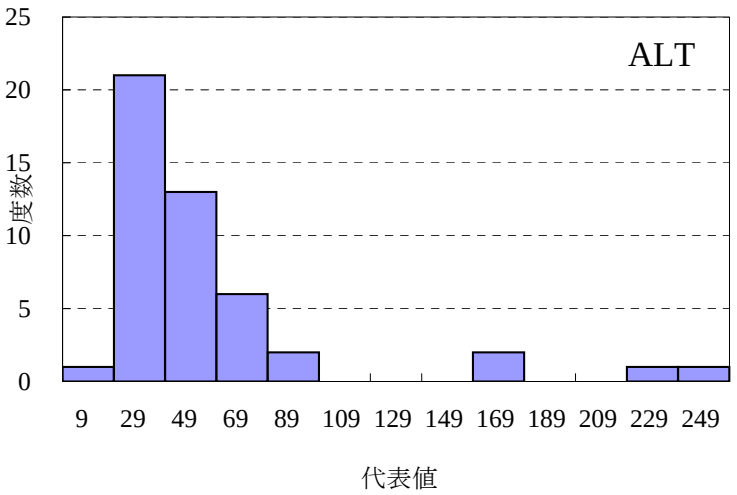
	ALT	AST
平均	55.73	35.84
標準偏差	48.08	9.82
変動係数	0.86	0.27
ひずみ	**2.74**	1.16
とがり	**7.56**	1.37
最小値	17.80	23.10
第1四分位値	30.55	28.35
中央値	41.10	33.80
第3四分位値	59.40	41.25
最大値	242.40	65.70
四分位値の差	28.85	12.90
差の1.5倍	43.28	19.35

ヒストグラム作成用データ

ALT

間隔　20

上限値	累積度数	代表値	度数
19	1	9	1
39	22	29	21
59	35	49	13
79	41	69	6
99	43	89	2
119	43	109	0
139	43	129	0
159	43	149	0
179	45	169	2
199	45	189	0
219	45	209	0
239	46	229	1
259	47	249	1



データ

	ALT	AST
1	49.4	41.7
2	41.7	34.4
3	63.0	23.4
4	17.8	28.9
5	66.0	46.6
6	36.3	37.3
7	34.2	33.4
8	30.1	32.2
9	35.2	23.5
10	62.6	47.6
11	28.7	33.4
12	51.5	46.3
13	39.6	37.6
14	35.1	25.5
15	23.1	44.8
16	33.1	35.9
17	41.1	28.0
18	31.8	45.7
19	51.7	35.5
20	42.8	34.7
21	45.6	29.3
22	24.2	36.5
23	30.5	32.2
24	23.8	25.6
25	226.0	63.4
26	49.0	29.1
27	44.8	28.1
28	37.3	33.4
29	38.5	27.1
30	242.4	47.1
31	23.8	30.7
32	27.5	28.6
33	56.2	33.8
34	50.0	24.8
35	43.6	28.0
36	63.7	37.4
37	161.0	65.7
38	90.3	48.4
39	30.6	26.6
40	26.2	23.1
41	71.5	45.9
42	31.2	40.8
43	64.9	35.1
44	21.4	31.1
45	162.2	35.5
46	89.5	52.8
47	28.7	27.8

基本統計量

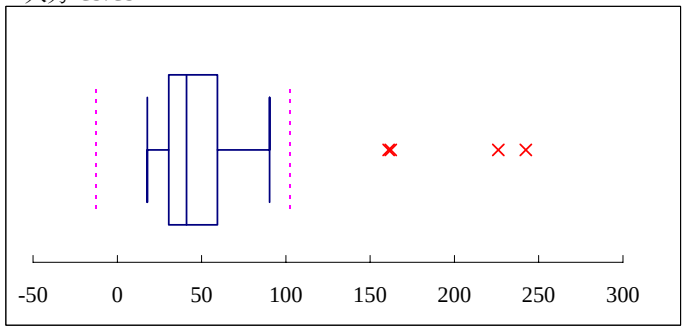
	ALT	AST
平均	55.73	35.84
標準偏差	48.08	9.82
変動係数	0.86	0.27
ひずみ	2.74	1.16
とがり	7.56	1.37
最小値	17.80	23.10
第1四分位値	30.55	28.35
中央値	41.10	33.80
第3四分位値	59.40	41.25
最大値	#####	65.70
四分位値の差	28.85	12.90
差の1.5倍	43.28	19.35

箱ひげ図作成用パラメータ

ALTデータ

出力 R5

入力 C5:C51

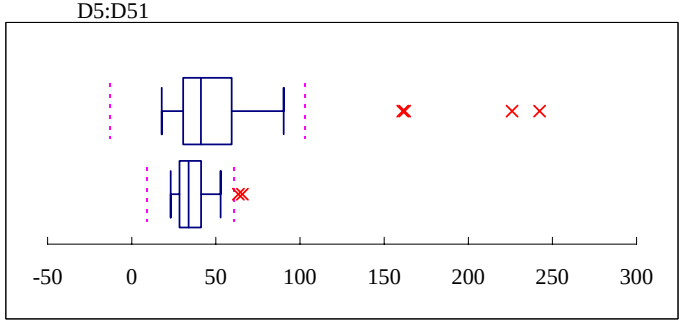


ALT･ASTデータ

出力 V5

入力 C5:C51

D5:D51



箱ひげ図作成用計算表

242		15	65.7		15
226		15	63.4		15
162		15	9	23	
161		15	9	7	
-13	23		23.1	22	
-13	7		23.1	8	
17.8	22		23.1	15	
17.8	8		28.4	15	
17.8	15		28.4	25	
30.6	15		41.3	25	
30.6	25		41.3	15	
59.4	25				
59.4	15		33.8	25	
			33.8	5	
41.1	25				
41.1	5		28.4	15	
			28.4	5	
30.6	15		41.3	5	
30.6	5		41.3	15	
59.4	5		52.8	15	
59.4	15		52.8	22	
90.3	15		52.8	8	
90.3	22		60.6	23	
90.3	8		60.6	7	
103	23				
103	7		242		40
			226		40
			162		40
			161		40
			-13	48	
			-13	32	
			17.8	47	
			17.8	33	
			17.8	40	
			30.6	40	
			30.6	50	
			59.4	50	
			59.4	40	
			41.1	50	
			41.1	30	
			30.6	40	
			30.6	30	
			59.4	30	
			59.4	40	
			90.3	40	
			90.3	47	
			90.3	33	
			103	48	
			103	32	

表示 3.1

データ

	ALT	AST
1	49.4	41.7
2	41.7	34.4
3	63.0	23.4
4	17.8	28.9
5	66.0	46.6
6	36.3	37.3
7	34.2	33.4
8	30.1	32.2
9	35.2	23.5
10	62.6	47.6
11	28.7	33.4
12	51.5	46.3
13	39.6	37.6
14	35.1	25.5
15	23.1	44.8
16	33.1	35.9
17	41.1	28.0
18	31.8	45.7
19	51.7	35.5
20	42.8	34.7
21	45.6	29.3
22	24.2	36.5
23	30.5	32.2
24	23.8	25.6
25	226.0	63.4
26	49.0	29.1
27	44.8	28.1
28	37.3	33.4
29	38.5	27.1
30	242.4	47.1
31	23.8	30.7
32	27.5	28.6
33	56.2	33.8
34	50.0	24.8
35	43.6	28.0
36	63.7	37.4
37	161.0	65.7
38	90.3	48.4
39	30.6	26.6
40	26.2	23.1
41	71.5	45.9
42	31.2	40.8
43	64.9	35.1
44	21.4	31.1
45	162.2	35.5
46	89.5	52.8
47	28.7	27.8

基本統計量

	ALT	AST
平均	55.73	35.84
標準偏差	48.08	9.82
変動係数	0.86	0.27
ひずみ	2.74	1.16
とがり	7.56	1.37
最小値	17.80	23.10
第1四分位値	30.55	28.35
中央値	41.10	33.80
第3四分位値	59.40	41.25
最大値	#####	65.70
四分位値の差	28.85	12.90
差の1.5倍	43.28	19.35

データの対数変換

	ALT	AST
1	3.9	3.7
2	3.7	3.5
3	4.1	3.2
4	2.9	3.4
5	4.2	3.8
6	3.6	3.6
7	3.5	3.5
8	3.4	3.5
9	3.6	3.2
10	4.1	3.9
11	3.4	3.5
12	3.9	3.8
13	3.7	3.6
14	3.6	3.2
15	3.1	3.8
16	3.5	3.6
17	3.7	3.3
18	3.5	3.8
19	3.9	3.6
20	3.8	3.5
21	3.8	3.4
22	3.2	3.6
23	3.4	3.5
24	3.2	3.2
25	5.4	4.1
26	3.9	3.4
27	3.8	3.3
28	3.6	3.5
29	3.7	3.3
30	5.5	3.9
31	3.2	3.4
32	3.3	3.4
33	4.0	3.5
34	3.9	3.2
35	3.8	3.3
36	4.2	3.6
37	5.1	4.2
38	4.5	3.9
39	3.4	3.3
40	3.3	3.1
41	4.3	3.8
42	3.4	3.7
43	4.2	3.6
44	3.1	3.4
45	5.1	3.6
46	4.5	4.0
47	3.4	3.3

対数変換後の基本統計量

	ALT	AST
平均	3.81	3.55
標準偏差	0.59	0.25
変動係数		
ひずみ	1.23	0.54
とがり	1.57	-0.15
最小値	2.88	3.14
第1四分位値	3.42	3.34
中央値	3.72	3.52
第3四分位値	4.08	3.72
最大値	5.49	4.19
四分位値の差	0.66	0.38
差の1.5倍	1.00	0.56

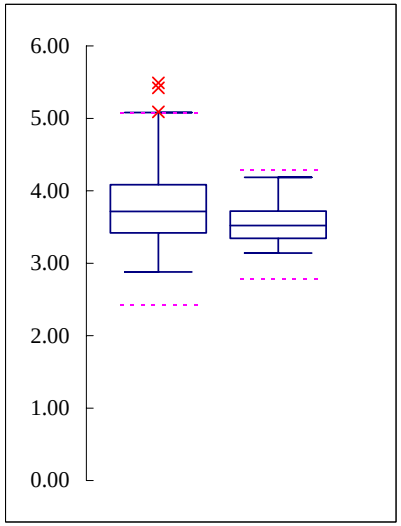
箱ひげ図作成用パラメータ

対数変換 ALTデータ

出力 X5 T

入力 C5:C51 L

D5:D51



箱ひげ図作成用計算表

15		5.49
15		5.42
15		5.09
23		2.42
7		2.42
22	2.88	
8	2.88	
15	2.88	
15	3.42	
25	3.42	
25	4.08	
15	4.08	
25	3.72	
5	3.72	
15	3.42	
5	3.42	
5	4.08	
15	4.08	
15	5.08	
22	5.08	
8	5.08	
23		5.08
7		5.08
48		2.78
32		2.78
47	3.14	
33	3.14	
40	3.14	
40	3.34	
50	3.34	
50	3.72	
40	3.72	
50	3.52	
30	3.52	
40	3.34	
30	3.34	
30	3.72	
40	3.72	
40	4.19	
47	4.19	
33	4.19	
48		4.28
32		4.28

基本に戻ろう
基本統計量とデータの比較

株式会社　バイオスタティスティカルリサーチ
杉本　典子

1

はじめに

グリーン本の紹介

「じっくり勉強すれば身につく統計入門」がテーマ。
医薬の分野で統計的方法を適用する際の基本的な考え方を説明しており、入門者向け。

ダウンロード先
http://www.scientist-press.com/12_278.html

2

イエロー本

浜田知久馬「学会・論文発表のための統計学― 統計パッケージを誤用しないために―」, 真興交易医書出版部(1999)

と

ピンク本

吉村功編著「毒性・薬効データの統計解析― 事例研究によるアプローチ―」, サイエンティスト社(1987)

両者の間を結びつける役割

「基本に戻るには・・・・」
今回は浜田先生のイエロー本
２章とグリーン本の一部から紹介です

3

内容

要約統計量（イエロー本から）

- 平均
- 中央値
- SD, SE

２組のデータの比較（グリーン本から）

- ｔ検定
- Welchの検定

4

平均

平均に関する皮肉

“Statistic has been described as the science which tells you that if you lie with your head in the oven and your feet in the refrigerator, on average you’ll be comfortably warm.”

「統計学とは、頭がオーブンの中にあって、足が冷蔵庫の中にあれば、平均して暖かいと記述する科学である」

この皮肉はある意味で統計学の本質を捉えたもの
統計の重要な役割の１つはデータの要約

5

平均

データ解析の要約統計量の代表
→平均、標準偏差(SD)

平均　：　分布の中心
標準偏差(SD)　：　バラツキを示す指標

＜分布の中心の指標＞
ASTやALTのような生化学検査では、
機械的に平均値を計算するまえに
データをグラフ化する習慣が大事

6

血中甲状腺刺激ホルモン(TSH)の濃度

TSH							
0.06	0.13	5.22	0.72	1.14	0.05	0.09	2.41
3.36	0.45	2.99	3.42	2.74	0.37	0.87	1.15
0.84	0.83	3.16	0.05	0.09	0.12	0.08	12.66
0.12	6.94	3.42	8.44	0.07	1.29	1.35	0.07
2.90	6.64	0.05	4.42	1.85	0.29	1.11	1.00
0.20	1.84	1.26	0.05	18.37	9.17	1.48	5.85
1.62	0.65	0.16	0.06	1.73	0.10	13.25	1.30
0.07	1.81	2.36	1.51	1.29	24.30	0.08	0.78
6.01	0.47	0.38	0.07	0.07	0.09	0.76	1.02
1.89	2.08	0.69	0.08	0.78	0.06	0.51	1.01
3.40	0.34	3.67	1.10	0.50	2.44	12.30	0.14
0.61	2.43	1.42	0.34	0.05	0.07	57.00	0.05
1.02							

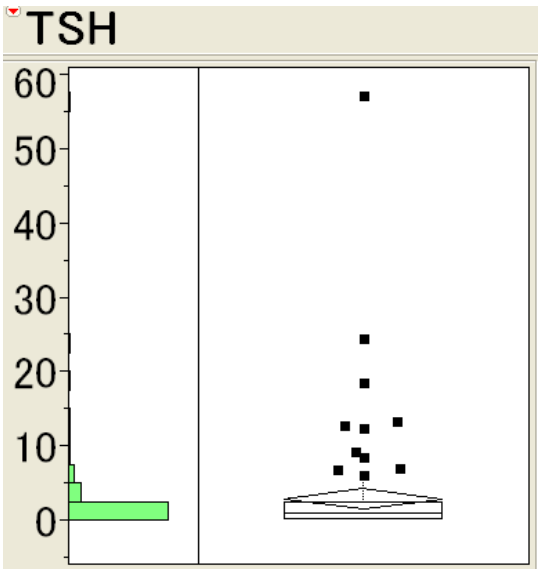
N：97　平均：2.84　　SD：6.79

これだけみると2.84を中心としてTSHが分布してると思ってしまいそう。

7

血中甲状腺刺激ホルモン(TSH)の濃度

グラフ化



・57や24.3のような外れ値が平均値をひっぱっている
・データは1前後をとる値が多いことがわかります。
⇒こういった場合、橘田さんのところであったように平均値より中央値を用いたほうが適切です。
（ちなみに中央値=1.01）

8

バラツキを示す指標

では分布の中心が決まったとして
次は中心位置からデータが
どの程度ばらつくかを示す代表がSDです。

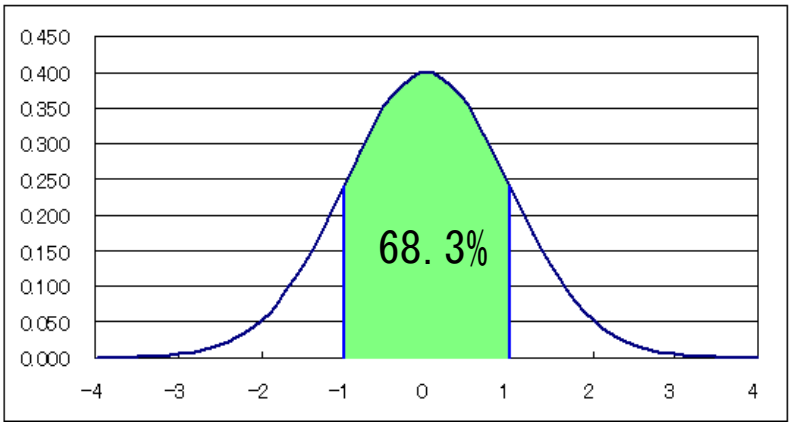
なぜSDか？

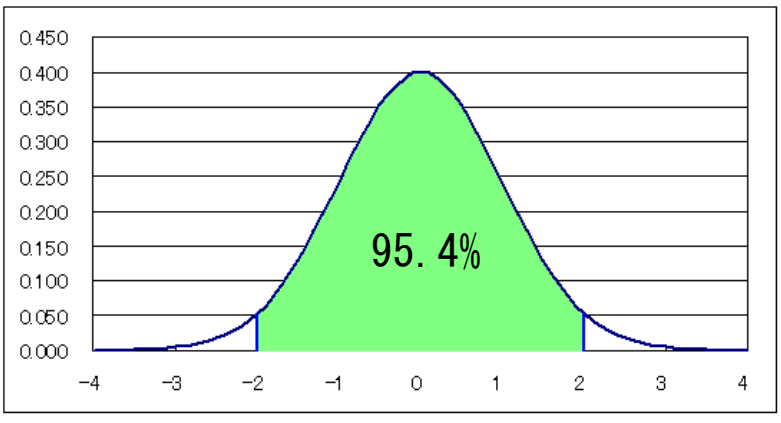
正規分布であれば
平均±SDの区間にデータの約68%、
平均±2SDの区間にデータの約95%が
入るから。

正規分布の場合、平均値とSDでばらつく範囲を示せる

9

正規分布の±1ＳＤ、２ＳＤに入るデータの割合

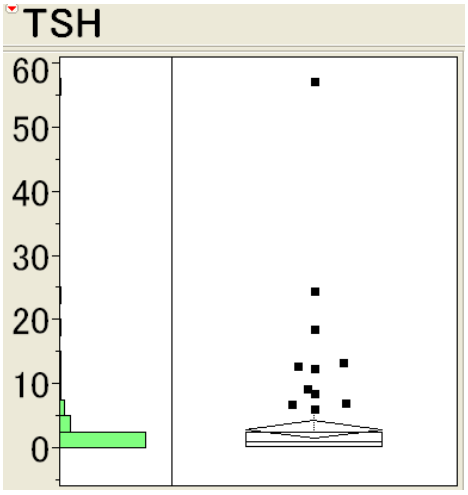




10

バラツキを示す指標

でも・・・
先ほどのTSHでは
平均±SD　-3.95～9.63
平均±2SD　-10.74～46.42
とりえない負の値を含んでいます。



原因は
正規分布とかけはなれているから

この例だと
平均±SDでは91/97個で94%、
平均±2SDでは95/97個で98%
のデータを含んでいて
バラツキを過大に評価しています。

11

バラツキを示す指標

正規分布とかけはなれているなら、
バラツキ具合はどう示す？

平均値でなく中央値でみてみよう
そして、中央値を拡張して分位点を定義しよう！

分位点は
例：100個データがあれば、大きい順に並び替えて
　　下から10番目が10%点、下から20番目が20%点

TSHの例だと
15%～85%点　　0.07～3.67
2.5%～97.5%点　0.05～18.37
となり、平均値±SDと比べて負の値も含みません

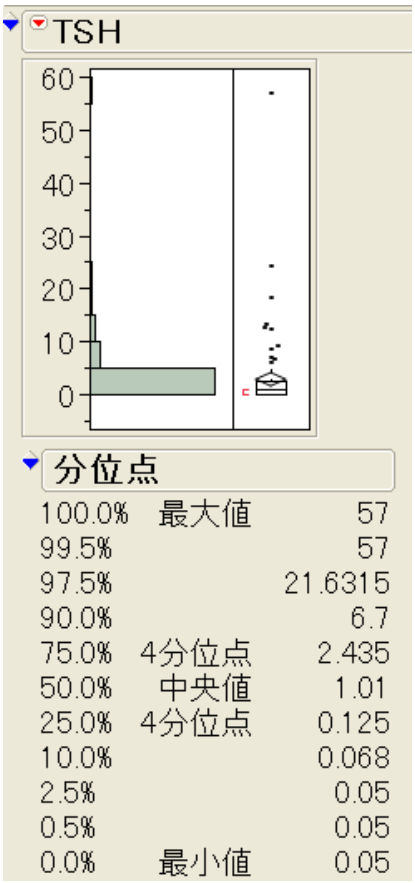
12

バラツキを示す指標

＜箱ひげ図＞
箱の下が25%点、上が75%点
箱の長さ＝四分位偏差です

TSHの要約は
　中央値　　1.01
　25～75%点　0.13～2.43
　四分位偏差　2.3

⇒平均2.84とSD6.79
　データの分布の特性を
　適切に表している



13

分布の中心とバラツキを示す指標

＜まとめ＞
・分布の位置とバラツキを表す指標として
　平均値とSDがよく用いられる

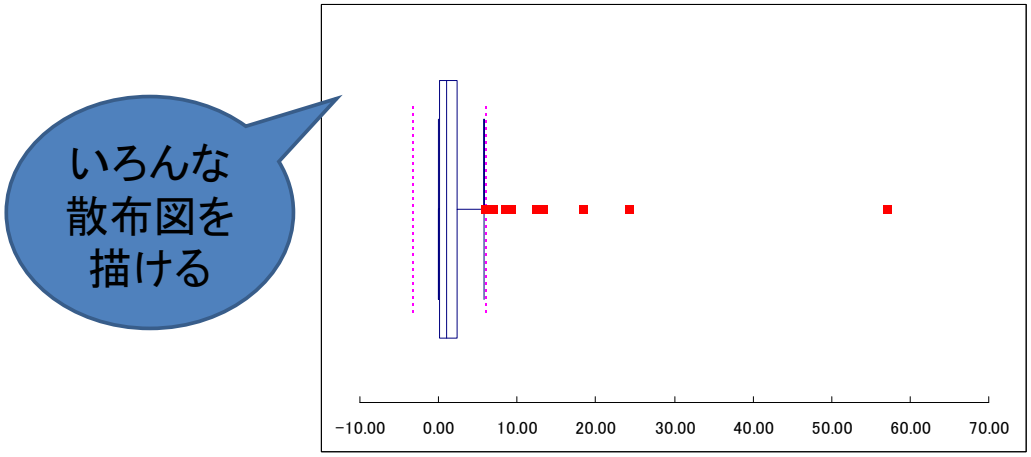
・外れ値を含んでいる場合、正規分布でない場合は
　誤解をまねかないように中央値と分位点の利用
　を考慮する必要がある

14

補足：エクセルで箱ひげ図

箱ひげ図をエクセルで描くには
ツールが「無料」で手に入ります
　　詳しくは・・・グリーン本　第1部　基礎

＜ツールで作図した例＞



15

SDとSE

SD　標準偏差　Standard　Deviation
SE　標準誤差　Standard　Error
この２つは意味が全く違うので混同しては
いけません。
＊グラフでどちらを利用しているか明記が大事

＜SDとは＞
　生データのバラツキの大きさを表す指標ですが、
　医薬データでよくみられる右にすそを引いた分布
　では適切な指標とはいえません

＜SEとは＞
　推定値（例：平均値）のバラツキの大きさを示す指標
　SE＝SD／√Nで計算されます

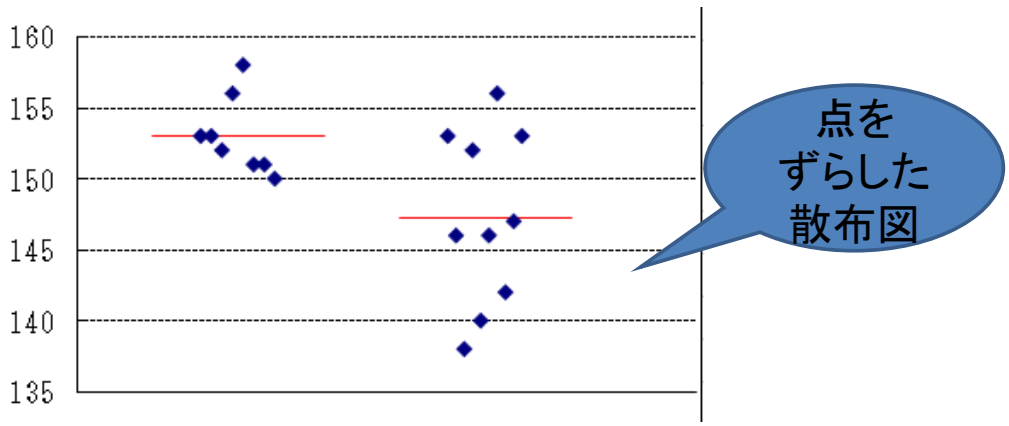
16

2 組のデータの解析

17

2 組のデータの解析：目的

2組のデータ解析では，
2つの群の間に平均値に違いがあるかを
検討したい



18

２群のデータを検定するとき

バラツキが２群で同じか？
（F検定）
↓

同じ場合	違う場合
ｔ 検定	Welchの検定

この流れで機械的にやっていませんか？

ここからは

- t検定とWelchの検定の違い
- 上記の流れで機械的にする以外にできること
  - ー対数変換による等分散化
  - ーLevene の検定

を紹介したいと思います。

19

t検定とWelchの検定の違い：データ

事例で ｔ 検定とWelchの検定の違いをみます

対照群と投与群のマウスの体重を測定した結果

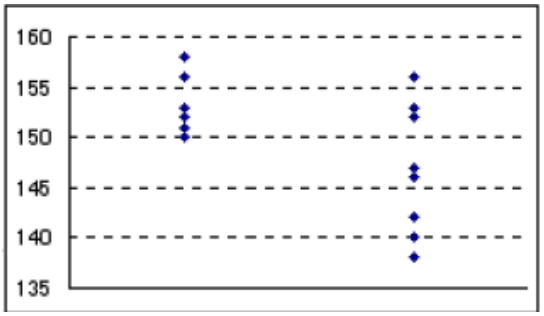
	対照群	投与群
1	153	153
2	153	146
3	152	138
4	156	152
5	158	140
6	151	146
7	151	156
8	150	142
9		147
10		153

まずはデータをグラフ化して眺めてみましょう

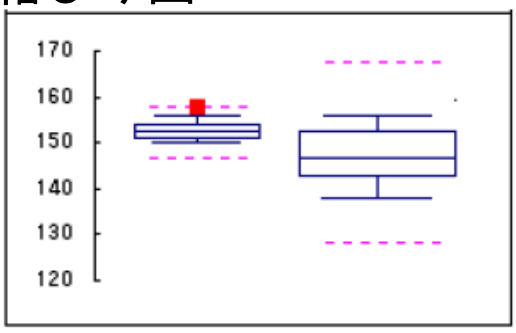
20

グラフ化

散布図



箱ひげ図



- 投与群のばらつきは対照群に比べて大きいな
- 平均値にも差があるように見える
- 対照群には外れ値が見られる.

正規分布から外れているか
どうかは±1.5 を目安

ひずみ	1.02	−0.13
とがり	0.19	−1.26

「正規分布を仮定して大丈夫」⇒ t 検定

t検定の流れ：t統計量→t分布→P値

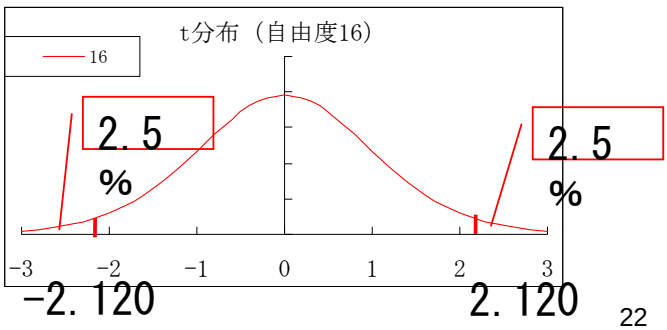
t 統計量→ t 分布を利用してP値

$$t = \frac{\text{平均値の差}}{\text{平均値の差の標準誤差}}$$

**ばらつきに比べて
自分が確認したい差が
十分大きいかどうか**

	対照群	投与群
1	153	153
2	153	146
3	152	138
4	156	152
5	158	140
6	151	146
7	151	156
8	150	142
9		147
10		153
n	8	10
平均	153.0	147.3
平方和	52.0	334.1
自由度	7	9
標準偏差	2.73	6.09

平均値の差の標準誤差
どう求める？



平均値の差の標準誤差

定式化すると・・・

各群の平均値 $\bar{x}_1.$, $\bar{x}_2.$ の分散

対照群の平均値の分散＝$V[\bar{x}_1]=V_1$、

投与群の平均値の分散＝$V[\bar{x}_2]=V_2$

平均値の差 $d=\bar{x}_2.-\bar{x}_1.$ の分散

$$V[\bar{x}_1-\bar{x}_2]=V[\bar{x}_1]+V[\bar{x}_2]=\frac{V_1}{n_1}+\frac{V_2}{n_2}$$ 分散の加法性

平均値の差の標準誤差

$$SE=\sqrt{\frac{V_1}{n_1}+\frac{V_2}{n_2}}$$

推定値のバラツキの大きさを示す指標 はSEと呼びました

t 検定とWelchの検定の違い：①分散

平均値の差の標準誤差

$$SE=\sqrt{\frac{V_1}{n_1}+\frac{V_2}{n_2}}$$

t 検定　２群とも分散が同じという仮定なので
２群のデータ全て使って分散を求める

$$V_1=V_2$$

$$SE=\sqrt{\frac{V}{n_1}+\frac{V}{n_2}}=\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)V}$$

t 検定とWelchの検定の違い：② t 分布の自由度

t 検定の自由度

$$(n_1 - 1) + (n_2 - 1)$$

Welchの検定の自由度

ここでは，t 値を求めるのに２つの平均平方を用いたので，その自由度は，ν_1，ν_2，$\nu_1 + \nu_2$ のいずれでもなく，定まらない．正確にいうと，式(3.5)で計算した t の値は t 分布とはならない．

Welch は式(3.5)で計算した t の近似分布として，自由度を調整した t 分布を用いることを提案した．その調整した自由度を **等価自由度** と呼ぶ．これを ν^* で表わすことにする．

等価自由度は

$$\frac{(V_1/n_1 + V_2/n_2)^2}{\nu^*} = \frac{(V_1/n_1)^2}{\nu_1} + \frac{(V_2/n_2)^2}{\nu_2} \qquad (3.6)$$

$$\nu^* = \frac{(V_1/n_1 + V_2/n_2)^2}{\dfrac{(V_1/n_1)^2}{\nu_1} + \dfrac{(V_2/n_2)^2}{\nu_2}} = \frac{(0.929 + 3.712)^2}{\dfrac{0.929^2}{7} + \dfrac{3.712^2}{9}} = 13.02$$

先ほどのデータ： t 検定とWelchの検定

	Y	Z	AA	AB	
14	n	8	10		
15	平均	153.0	147.3	5.7	平均の差
16	平方和	52.0	334.1	386.1	合計
17	自由度	7	9	16	合計
18	平均平方	7.43	37.12	24.13	
19	平均平方/n	0.929	3.712		
20					
21		t検定	Welch		
22	se	2.330	2.154		
23	t	2.446	2.646		
24	自由度	16	13.02	=(Z19+AA19)^2/(Z	
25	p値(両側)	0.026	0.020	=P_TDIST(AA23,AA	

SEと自由度が異なっています。
このデータはバラツキがF検定で有意
→Welchの検定の方を利用したほうがよいデータ

2群のデータを検定するとき

バラツキが２群で同じか？
（F検定）
↓

同じ場合	違う場合
t 検定	Welchの検定

この流れで機械的にやっていませんか？

ここからは

・t検定とWelchの検定の違い

この説明にはいります

・上記の流れで機械的にする以外にできること

－対数変換による等分散化

－ Levene の検定

を紹介したいと思います。

データ

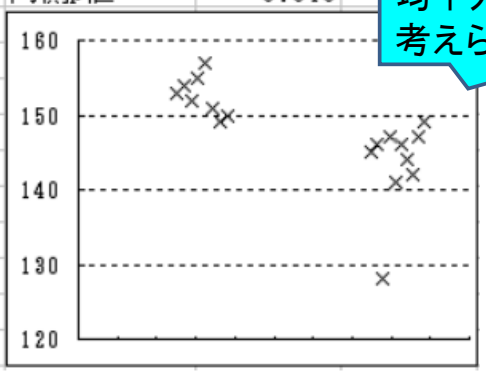
手元のデータをF検定したら有意。

F検定で有意＝バラツキが異なると即断するのは危険

⇒ データをグラフ化して良く観察してみると

	対照群	投与群
1	153	145
2	154	146
3	152	128
4	155	147
5	157	141
6	151	146
7	149	144
8	150	142
9		147
10		149
n	8	10
平均	152.6	143.5
平方和	49.9	318.5
自由度	7	9
平均平方	7.13	35.39

分子の自由度	9
分母の自由度	7
F比	4.967
上側p値	0.023
下側p値	0.977
両側p値	0.046



F検定は有意。
グラフからバラツキきが全体で違うというより，投与群に外れ値が含まれているために投与群の平均平方が大きくなったと考えられる．

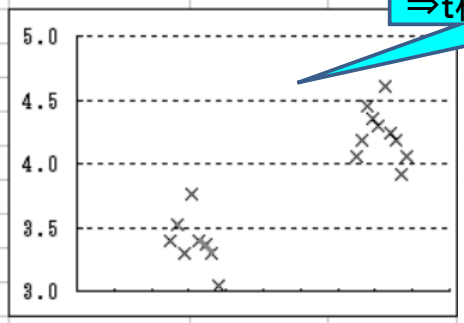
対数変換による等分散化

・対数変換による等分散化

⇒変数変換によって等分散に近づく検討する

	対照群	投与群
1	3.40	4.06
2	3.53	4.19
3	3.30	4.45
4	3.76	4.36
5	3.40	4.29
6	3.37	4.61
7	3.30	4.23
8	3.04	4.19
9		3.91
10		4.06
n	8	10
平均	3.4	4.2
平方和	0.3	0.4
自由度	7	9
平均平方	0.04	0.04
標準偏差	0.205	0.203

分子の自由度	9
分母の自由度	7
F比	0.985
上側p値	0.520
下側p値	0.480
両側p値	0.959



標準偏差は0.205, 0.203とほぼ等しくなった.

F 比は1 に近づき，有意ではなくった

⇒t検定で検定

Levene の検定

Levene の検定

- F検定のように分散の違いの検定
- 外れ値の影響を受け難いという利点

なぜなら、
データに外れ値が含まれているとき,

F 検定は偏差の2 乗を用いていてその影響を強く受ける.

Levene の検定は偏差の絶対値を用いるため
Levene の検定の方が外れ値の影響を受け難い.

ちなみにF検定とは

	対照群	投与群
1	153	153
2	153	146
3	152	138
4	156	152
5	158	140
6	151	146
7	151	156
8	150	142
9		147
10		153
n	8	10
平均	153.0	147.3
平方和	52.0	334.1
自由度	7	9
平均平方	7.43	37.12

2群の分散が同じか検定したいので，
各群について平均平方を計算する．

2群の分散が等しいなら
2つの平均平方の比（= F比）は

$$F = V_2 / V_1 \fallingdotseq 1$$

となるはず．

偏差の2 乗を用いている

2 つの平均平方の比

$$F = \frac{V_2}{V_1} = \frac{37.12}{7.43} = 5.00$$

データとF検定の結果

	対照群	投与群
1	153	153
2	153	146
3	152	138
4	156	152
5	158	140
6	151	146
7	151	156
8	150	142
9		147
10		153
n	8	10
平均	153.0	147.3
平方和	52.0	334.1
自由度	7	9
平均平方	7.43	37.12

分子の自由度	9
分母の自由度	7
F比	4.997
上側p値	0.023
下側p値	0.977
両側p値	0.045

偏差の絶対値

	対照群	投与群	
1	0.0	5.7	
2	0.0	1.3	
3	1.0	9.3	
4	3.0	4.7	
5	5.0	7.3	
6	2.0	1.3	
7	2.0	8.7	
8	3.0	5.3	
9		0.3	
10		5.7	
n	8	10	
平均	2.0	5.0	3.0
平方和	20.0	88.1	108.1
自由度	7	9	16
平均平方	2.86	9.79	6.76
平均値の差の標準誤差			1.23

平均値の差のt 検定をするのがLevene の検定
⇒このデータはLeveneの検定でも有意

平均値の差の有意差検定		
t	2.401	
p値(片側)	0.014	*
p値(両側)	0.029	*

Levenの検定はJMPでも標準出力される

⇒ JMP による等分散の検定の出力例

検定	F値	分子自由度	分母自由度	p値
O'Brien[.5]	5.1058	1	16	0.0382*
Brown-Forsythe	4.8860	1	16	0.0420*
Levene	5.7645	1	16	0.0289*
Bartlett	4.1086	1	.	0.0427*
両側F検定	4.9972	9	7	0.0455*

F検定で2群のばらつきに違いがみられた場合
グラフ化による吟味を省略してWelchの検定を機械的に適用する前に・・・

- 等分散が成立しないその原因を確認
- どのような違いかをみるためにグラフ化して検討
- Leveneの検定

⇒その結果により処理方法を考える

- 例えば対数変換などの変数変換により等分散に近づけてからt検定する方法
- Welch の検定をする

※正規分布でないときは対数変換などの変数変換で
正規分布を仮定できるか検討するのも重要

33

まとめ

- グラフ化がやはり大事。

  正規分布か？　等分散か？

  →外れ値を省く

  →対数変換など変数変換を実施する

  →F検定のみでなくデータによってLeveneの検定なども考慮する

- F検定⇒ t 検定　or　Welchの検定　の手順だけじゃなく

  →グラフ化などの吟味が大切

34