平均値とｔ検定を考える

松尾太加志（北九州市立大学文学部）

１．統計分析をしなくていいことがある

２人の陸上競技の記録。どちらかを代表選手にしたい。

（データ例１）

A	B
35.5	36.0

（データ例２）

	A	B
1	34.5	35.9
2	35.2	36.1
3	35.6	36.1
4	35.8	36.0
5	35.9	37.9

（データ例３）

	A	B
1	34.4	34.2
2	34.5	35.4
3	35.9	35.6
4	36.0	36.5
5	37.2	37.3

データ例１　→　１対１で比較可能　→　統計分析不要
データ例２　→　複数個だが,比較可能　→　統計分析不要
データ例３　→　直接データを比較して結論が出せない　→　１個ずつにすればよい

２．代表値という考え方

誰もが考えつくのが平均値　→　代表値　→　１対１で比較可能

（データ例３）

	A	B
1	34.4	34.2
2	34.5	35.4
3	35.9	35.6
4	36.0	36.5
5	37.2	37.3
平均	35.6	35.8

統計分析とは？
「ある結論を導くために，一連のデータ群の特徴を上手に表現するやり方」

「○○選手を代表にする」という結論を決めるとき，平均値でよいのか？
最小値？　最小値と最大値を削除して，真ん中の３つの平均？

みんなが納得いく代表値で決定を。

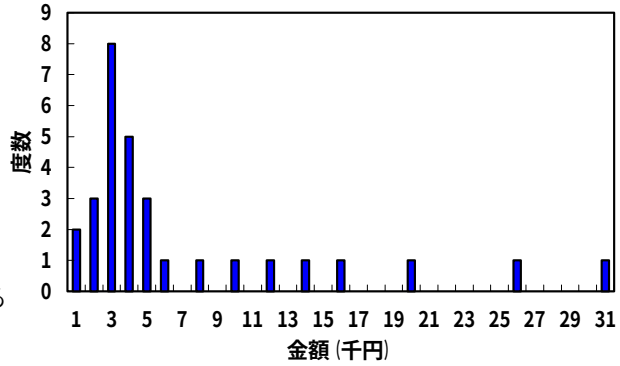
３．平均値は万能ではない

30人のお年玉の金額

（データ例４）

1,000	3,000	5,000
1,000	3,000	6,000
2,000	3,000	8,000
2,000	4,000	10,000
2,000	4,000	12,000
3,000	4,000	14,000
3,000	4,000	16,000
3,000	4,000	20,000
3,000	5,000	26,000
3,000	5,000	31,000

平均値　7,000　実感を反映していない？
最頻値　3,000　親が子どもにあげる金額の参考に？
中央値　4,000　自分の金額が多いか少ないか



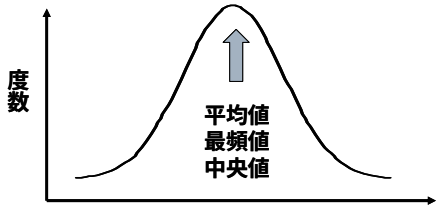
どのような目的に使うかによってかわる
代表値は勝手に決めてよい
データの分布に適切な代表値を
どのような代表値が適切か？

４．それでもやっぱり平均値

平均値は何かと便利

計算が簡単。検定に使いやすい。

正規分布では,平均値，最頻値，中央値がいっしょ。



平均値で比較

各クラスの５人の身長

（データ例５）

	A	B	C	D
1	157	161	150	151
2	158	162	154	157
3	159	163	157	165
4	160	164	161	170
5	161	165	173	172
平均	159	163	159	163

AとBでは,

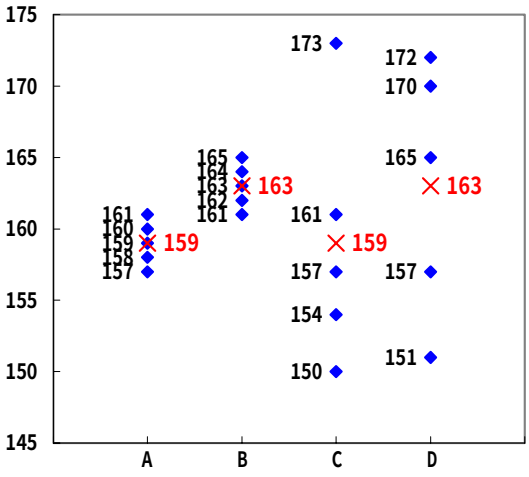
A<B

CとDでは,

C<Dだが…

単純に平均値だけでは

比較ができない



グラフに描いてみると

同じ平均の差でも，データのちらばり具合が異なる。

平均の差に比べて，ちらばりが十分に小さければよい。

ちらばりとの関係で判断すれば…

５．散らばりの指標が必要

範囲，平均偏差，分散，標準偏差…

（データ例５）

	A	B	C	D
1	157	161	150	151
2	158	162	154	157
3	159	163	157	165
4	160	164	161	170
5	161	165	173	172
平均	159	163	150	163
分散	2.0	2.0	62.0	62.8
標準偏差	1.41	1.41	7.87	7.92

分散：平均からの距離からの平方和

もっとも一般的に使われる指標

標準偏差：分散は，データを2乗しているので，ルートをとって元に戻してあげる。

分散に比べ，平均の差が十分に大きければ，２つの平均の間に差がありといってもいいのでは

次のような式は？　$\dfrac{(\text{群Bの平均}-\text{群Aの平均})}{(\text{群Bの分散}+\text{群Aの分散})}$

６．知りたいのは，真の世界

知りたいことの区別

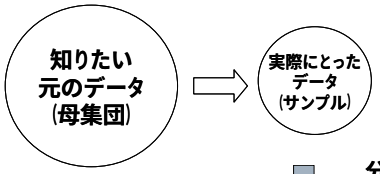
数値上の問題：各５つの平均値はどちらが大きいのか？　→　結論は出ている

知りたいこと：CとDのクラスではどちらが背が高いか？　→　サンプルから真の世界を推定する

サンプルから母集団の平均値と分散を推定する

平均値　→　そのままOK

分散　→　変換が必要　→　不偏分散



$$\text{分散}=\frac{(\text{各データ}-\text{平均})^2\text{の総和}}{\text{データの数}}$$

$$\text{分散}=\frac{(\text{各データ}-\text{平均})^2\text{の総和}}{\text{データの数}-1}$$

ｔ値

$$t = \frac{(\text{Bの平均} - \text{Aの平均})}{\sqrt{\frac{(\text{Bの各データ}-\text{Bの平均})^2\text{の総和} + (\text{Aの各データ}-\text{Aの平均})^2\text{の総和}}{(\text{Bの数}-1)+(\text{Aの数}-1)} \times \left(\frac{1}{\text{Bの数}}+\frac{1}{\text{Aの数}}\right)}}$$

前提：２つの分散は等しい

$$t_{B-A} = \frac{(163-159)}{\sqrt{\frac{(10+10)}{(5-1+5-1)} \times \left(\frac{1}{5}+\frac{1}{5}\right)}} = 4.0 \qquad t_{D-C} = \frac{(163-159)}{\sqrt{\frac{(314+310)}{(5-1+5-1)} \times \left(\frac{1}{5}+\frac{1}{5}\right)}} = 0.72$$

ｔ値が大きければ，２つのデータ（母集団）の間には違いがある
では,どの程度大きければ差があるのか？

７．こんなことを考えてみましょう

赤と青の２つのりんごの箱がある。２つの箱に入っているりんごの大きさは異なる。箱から２回ずつ５個のりんごを取り出した。

１回目：赤の箱　５個の平均250.0g
２回目：？の箱　５個の平均254.4g
２回目に取り出したのがどちらの箱かわからない。どちらの箱だろうか？

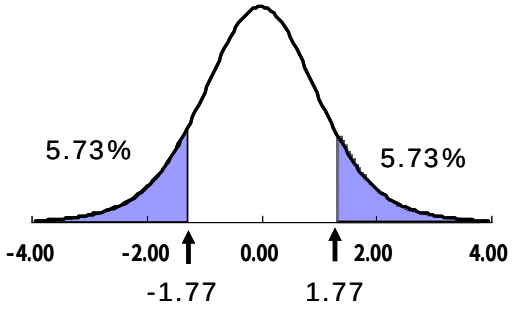
（データ例６）

	１個	２個	３個	４個	５個	平均	標準偏差	平均の差	ｔ値
赤の箱	250	253	258	252	259	250.0	3.94	4.4	-1.77
赤？青？	256	250	251	247	246	254.4	3.91		

２回目も赤の箱から取り出したとすれば，4.4gという差は生じるのかどうか？
実際に赤の箱の箱から，５個の２組を何度も取り出して，どのようになるかを調べよう。

	１個	２個	３個	４個	５個	平均	標準偏差	平均の差	ｔ値
1	255	244	254	253	244	250.0	5.52	2.2	-0.69
	250	253	247	252	259	252.2	4.44		
2	248	251	248	250	245	248.4	2.30	3.0	-1.52
	254	254	253	251	245	251.4	3.78		
3	252	247	253	250	246	249.6	3.05	2.0	-1.21
	253	252	252	253	248	251.6	2.07		
4	252	254	253	252	250	252.2	1.48	5.6	4.91
	245	250	247	246	245	246.6	2.07		
5	255	254	250	255	253	253.4	2.07	2.2	1.16
	247	249	254	256	250	251.2	3.70		

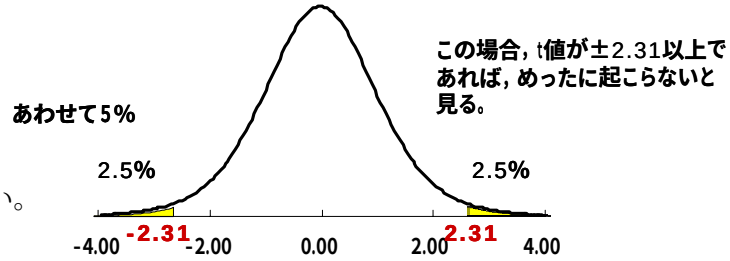
ばらつきも考慮すべきなので，平均の差ではなく，ｔ値でみる。
ｔ値をたくさん計算すると，下の図のような分布（ｔ分布）になる。
問題は，±1.77までずれる確率。 5.73+5.73=11.46%
11.46%は，起こる可能性が高いとみるか？



t値	確率
～ -4.0	0.20%
～ -3.5	0.40%
～ -3.0	0.85%
～ -2.5	1.85%
～ -2.0	4.03%
～ -1.5	8.60%
～ -1.0	17.33%
～ -0.5	31.53%
～ 0.0	50.00%
0.0 ～	50.00%
0.5 ～	31.53%
1.0 ～	17.33%
1.5 ～	8.60%
2.0 ～	4.03%
2.5 ～	1.85%
3.0 ～	0.85%
3.5 ～	0.40%
4.0 ～	0.20%

統計の世界では，５％を基準にする。　→　11.46%は十分に起こりうる。
赤い箱から取ったと結論づけられるか？

しかし，青い箱の可能性がある　→
青い箱は，大きさが異なるとしていたが，
「異なる」程度が小さければ，
２回目に取り出した５個の平均が 254.4g では，どちらの箱からであるかは区別がつかない。



８．ｔ検定とは

２つの母集団の平均に差があるかどうかをみる。
そのときに２つの母集団のサンプルの平均値を利用する。２つのサンプルの平均値からｔ値を算出して，確率的に２つの母集団の平均値に差があるかどうかをみる。

（データ例７）

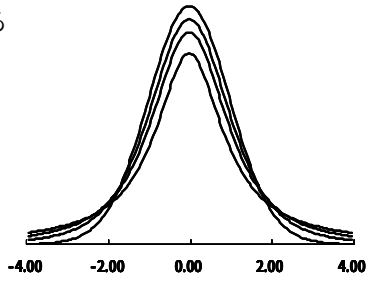
	１個	２個	３個	４個	５個	平均	標準偏差	平均の差	ｔ値
赤の箱	250	253	258	252	259	250.0	3.94	5.8	-2.72
黒の箱	254	255	260	253	262	256.8	3.96		

t=2.72　この場合，2.31 を超えているので，有意差あり。

確率の計算をどうするか？　→　先ほどのように，実際にサンプルをとることはできない。

ｔ値は，どんなデータでも同じ分布をする
ただし，条件がある。
　正規分布するデータであること
　データの数によって変わる

データ数が増えると，尖度が高くなる。
０付近になる確率が高くなる。



自由度	t値	自由度	t値
1	12.71	16	2.12
2	4.30	17	2.11
3	3.18	18	2.10
4	2.78	19	2.09
5	2.57	20	2.09
6	2.45	25	2.06
7	2.36	30	2.04
8	2.31	35	2.03
9	2.26	40	2.02
10	2.23	45	2.01
11	2.20	50	2.01
12	2.18	100	1.98
13	2.16	200	1.97
14	2.14	500	1.96
15	2.13	∞	1.96

ｔ表：ｔ値が「０である」と言える確率が５％であるときのｔ値を示す。
　　（実際には，確率が１％や 10%の場合も書いてある）

身長のデータ（データ例５）をｔ表でみると，
　t_8＝4.00＞2.31　　有意差あり
　t_8＝0.72＜2.31　　有意差なし

統計検定の考え方
　まず，「２つは同じ」と考えておいて（帰無仮説）
　同じだと考えるには，確率的に無理がある（普通，５％）　→　そこで差があるとする。

・危険率（有意確率）
「同じではない（有意な差がある）」という結論を出してしまったが，それが間違っている可能性，間違っている危険率
　５％　１％　0.1%などで決める。

・ｔ検定の結果の記述のしかた
　t (8)＝3.25，p<.05　　ｔ値，自由度，有意確率を記述
　t_8=1.56, ns　　ｔ値は，プラスもマイナスもあるが，一般にはプラスで示す。
　t =13.87，df=4，p<.01

次のような記述はよくない

> 条件間の比較を行うため，t 検定を行ったところ，有意な差が見られた（t_5=3.16，p<.05）。

↓　以下のように書く

> 条件間の比較を行うため，条件ごとに平均値を算出した（図○参照）。平均値を比較するため，t 検定を行ったところ，有意な差が見られた（t_5=3.16，p<.05）。

※平均値を代表値として使うことを記述すること。その平均値を示すこと。t 統計量などは，むしろ無くてもよい。

９．t 検定は万能ではない

母集団が平均を代表値としてとることが意味をもっていないといけない。
母集団が正規分布をしていることが必要。

（データ例３）

	A	B
1	34.5	35.9
2	35.2	36.1
3	35.6	36.1
4	35.8	36.0
5	35.9	37.9
平均	35.4	36.4
標準偏差	0.57	0.84

t_8=2.20，p>.05
有意差なし
これで「有意差なし」なのか？
はずれ値「37.9」が分散を大きくしてしまっている。

（データ例８）

	A	B
1	34.5	35.9
2	35.2	36.1
3	35.6	36.1
4	35.8	36.0
5	35.9	36.9
平均	35.4	36.3
標準偏差	0.57	0.40

はずれ値がなければ，
t_8=2.57，p<.05
有意差あり

10.統計検定よりもデータを自分の目でみる

重要なことは
Rawデータを十分に眺める
t検定に適しているデータであるかどうかを見極める
・平均値をとることに意味があるのか？
・分布が正規分布になりそうか

データを眺めることが大切
データ数が多くなると，眺められないので，平均値などの代表値に頼るだけにすぎない。
盲目的に，平均値をとって，t 検定や分散分析をすることはよくない。