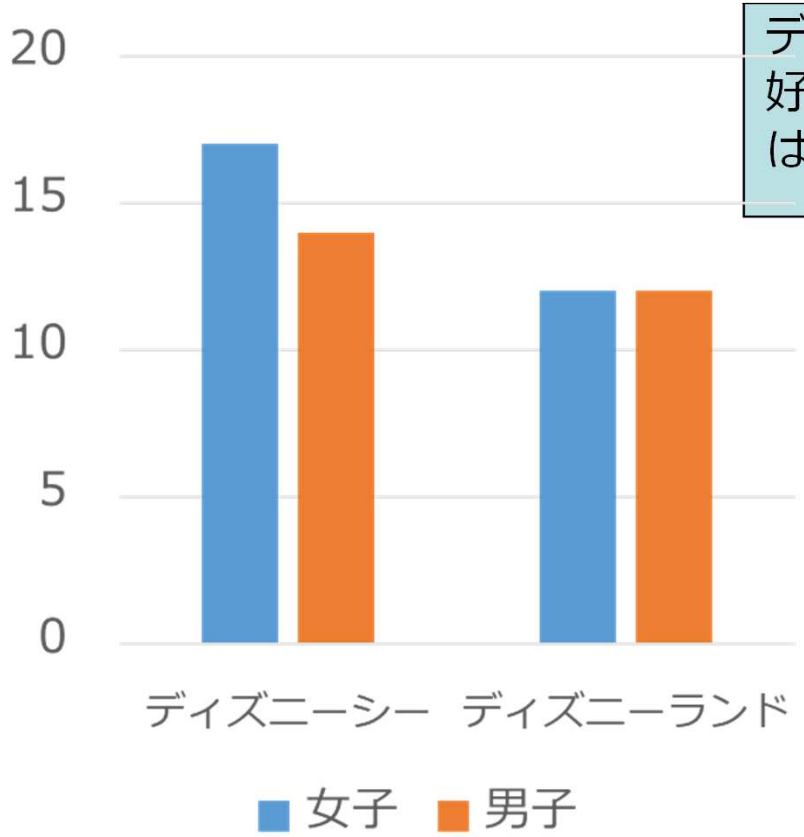
「アルゴリズムとプログラム」

三学期　第7回　袖高の生徒ってどうよ調査(3)

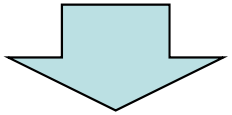
結果を統計的に正しく判断



本資料の信頼区間の図等は、統計学習オンライン(http://stat.doscience.jp)を使用して作成しています。

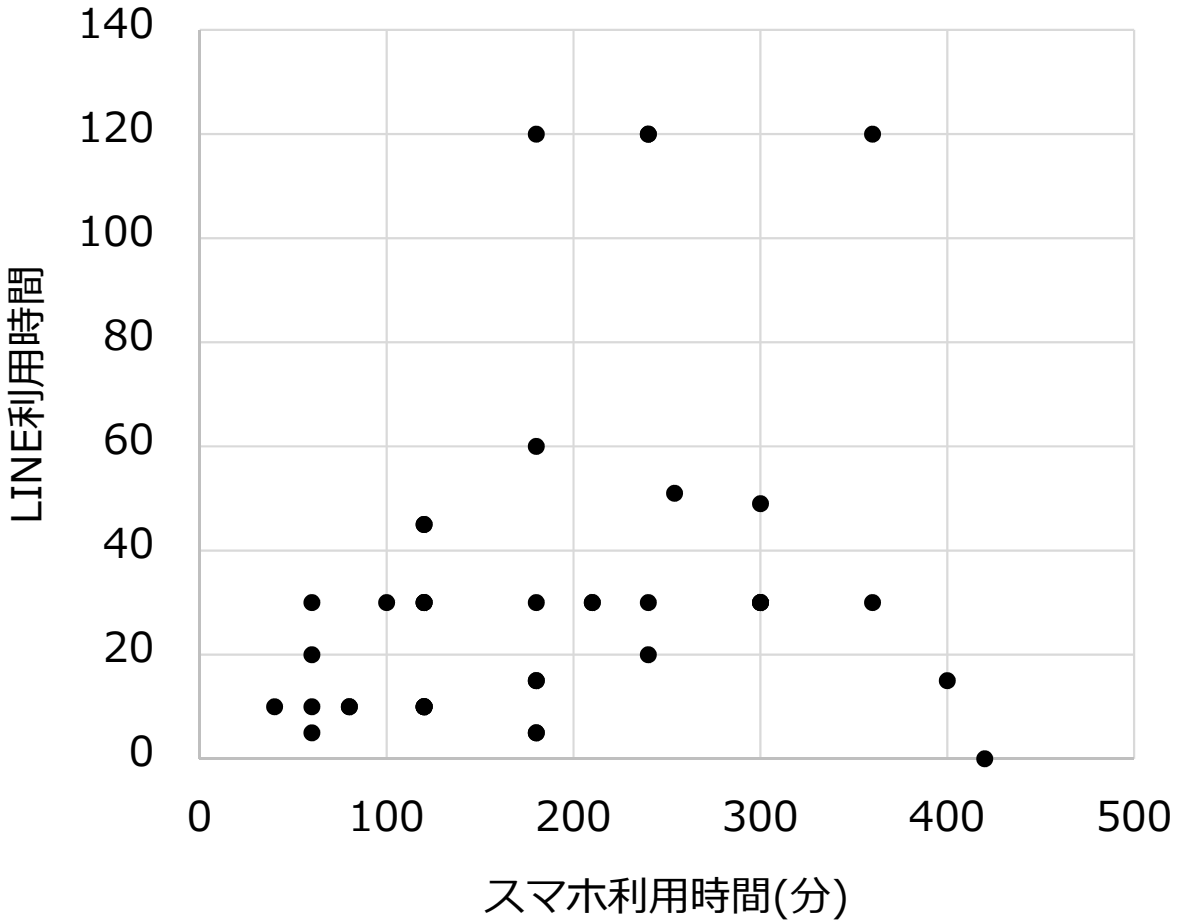
統計的推論とは

母集団から抽出された標本に基づき，母集団分布そのもの，あるいは母集団分布が想定されているときには，母集団における集団的特性値を引出すための方法をいう。その代表的なものは推定と検定である。



人間の感覚的な判断ではなく、ものごとが起こりうる確率をもとにして、データを判断したり、推測したりすること。

簡単なところから: 相関:ものごとの関連の強さ

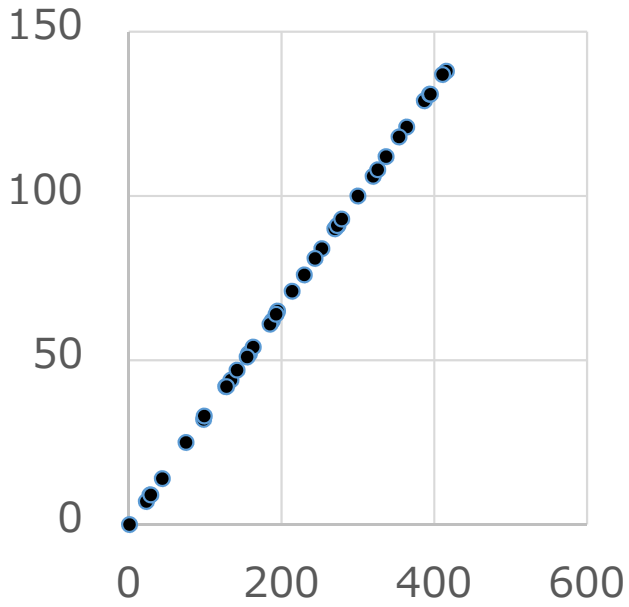


スマホの利用時間とLINEの使用時間は関係があるの?

関係の強さ
＝相関係数

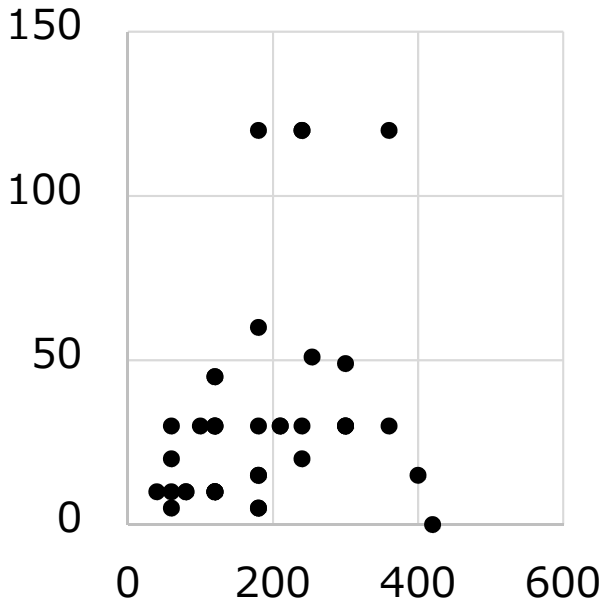
数学Bあたりでやっている?

相関係数 r



相関が完全にある
y = ax

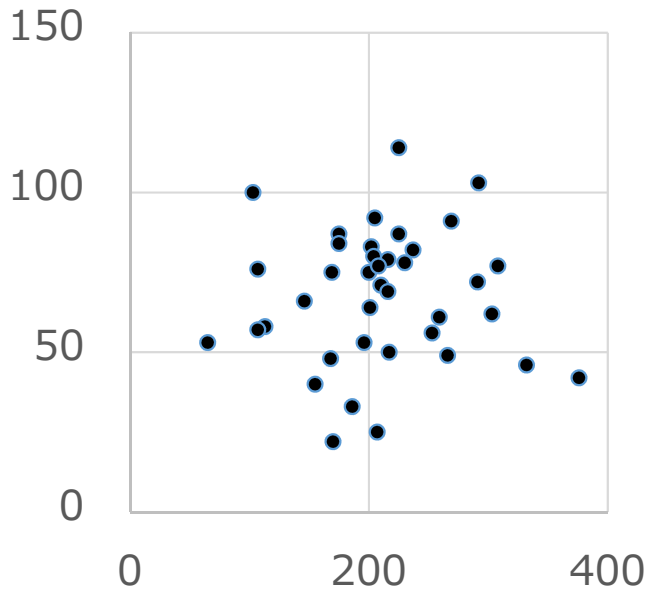
r = 1.00



相関の強さ

1.00 ≧ |r| ≧ 0
1.00 ≧ r ≧ -1.00

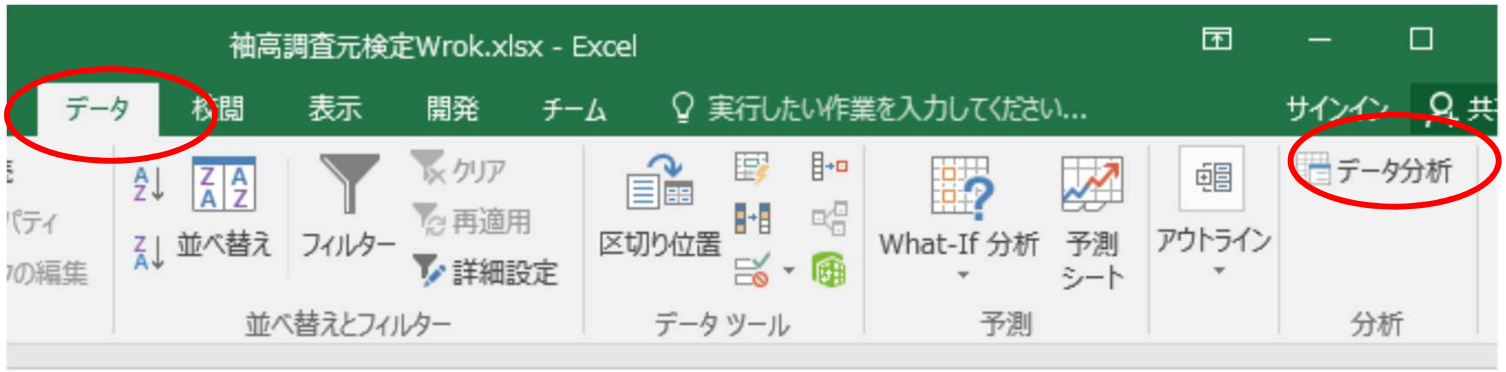
r が負の数の場合
は負の相関

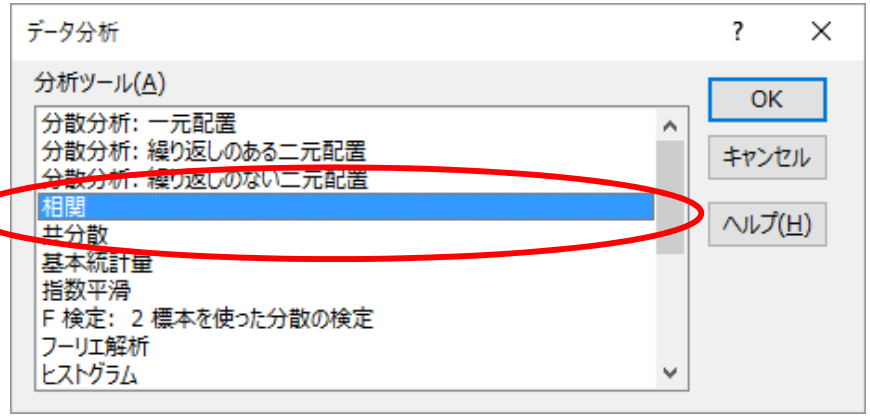


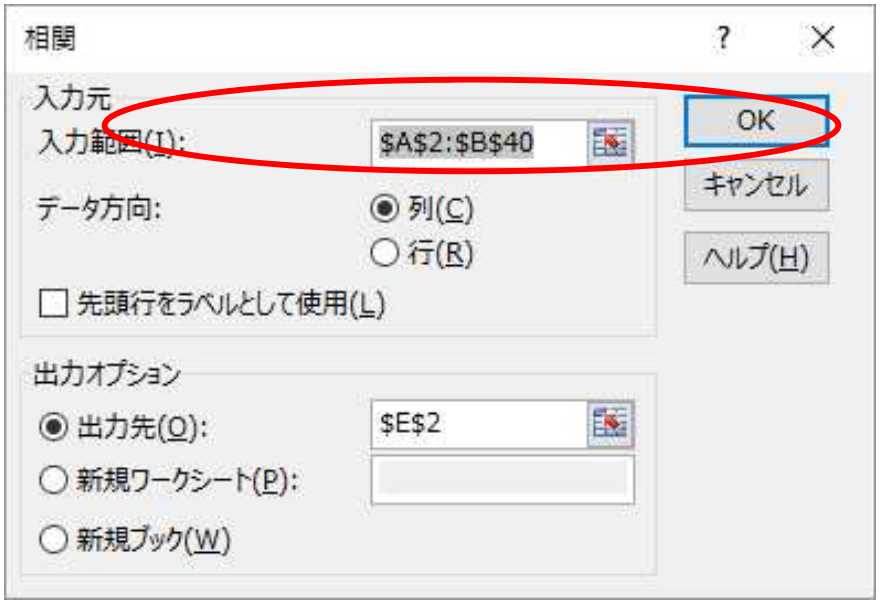
相関がない
バラバラ

r = 0

Excelを使った相関係数の求め方







Excel相関係数の結果

	列 1	列 2
列 1	1	
列 2	0.28162	1

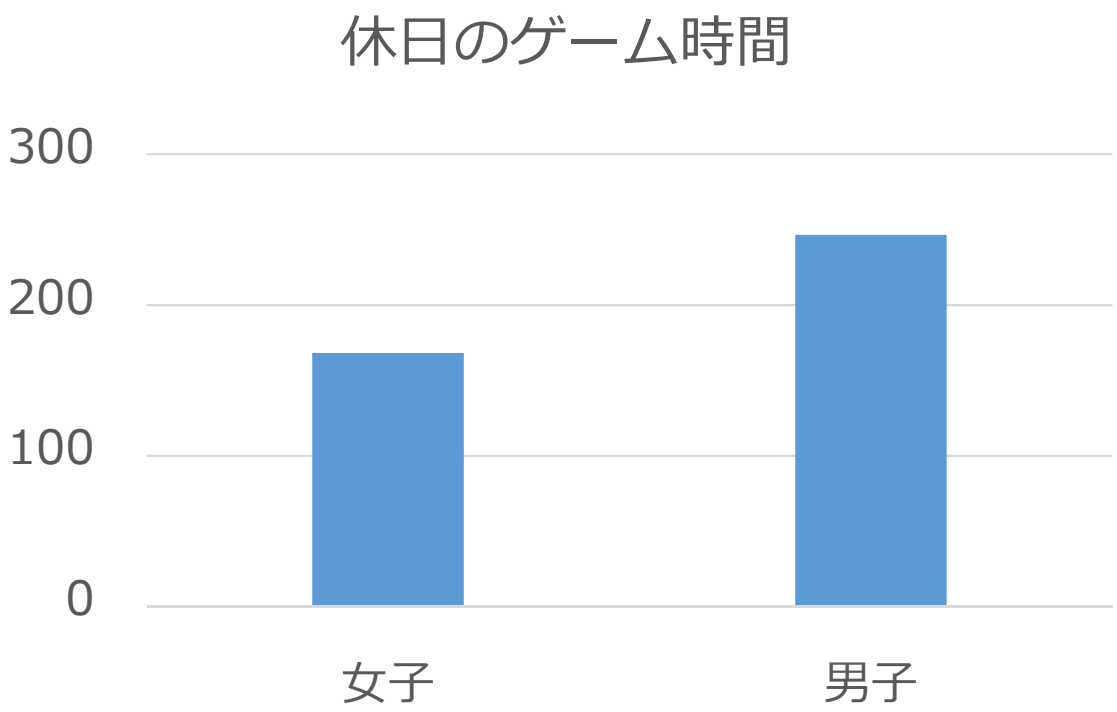
| r | =	0.7～1	かなり強い相関がある
| r | =	0.4～0.7	やや相関あり
| r | =	0.2～0.4	弱い相関あり
| r | =	0～0.2	ほとんど相関なし

書き方例:
xxxxとyyyyyの関係を見るために、相関分析を行った。その結果、
xxxxとyyyyの間には、高い正の相関が認められた （ r = .747)

xxxxとyyyyyの関係を見るために、相関分析を行った。その結果、
xxxxとyyyyの間には、ほとんど相関がなかった （ r = .011)

グループで違いがあるか判断する: t検定

グループ間の人数の違いを比較する(クロス集計)



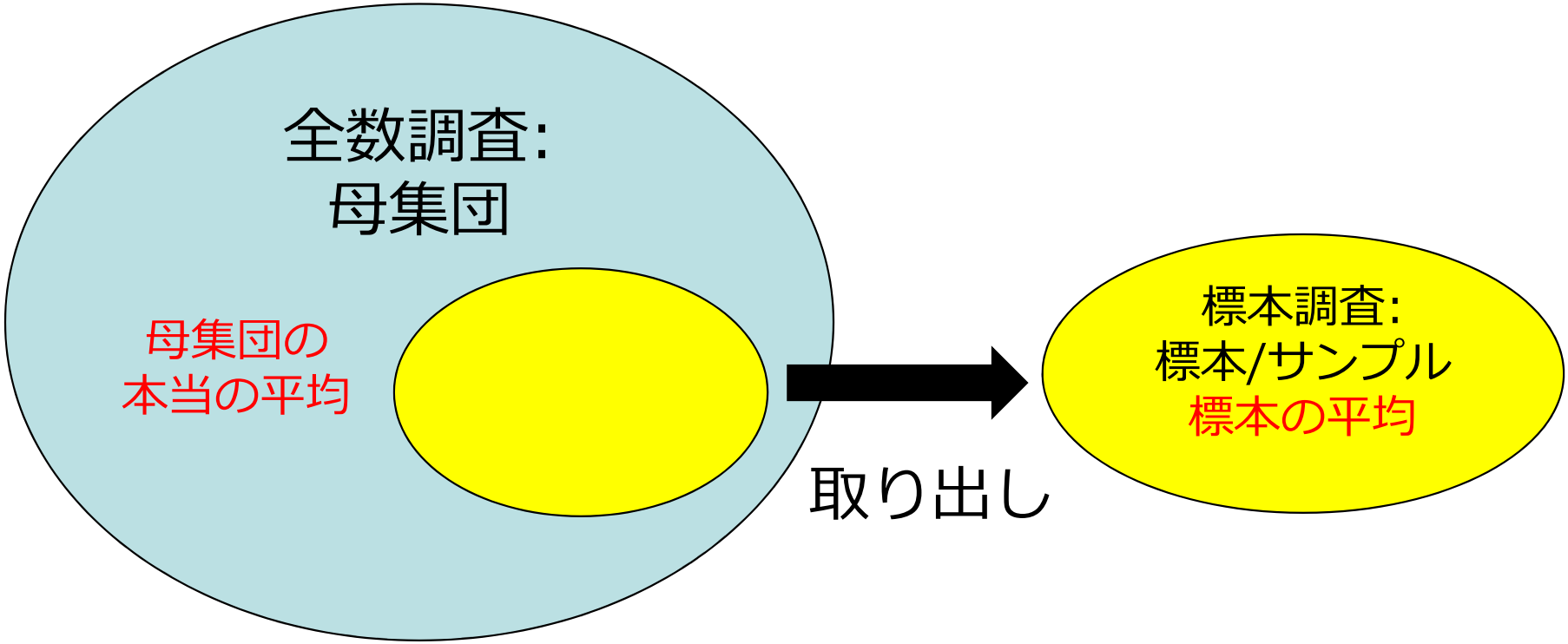
男女でゲーム時間に違いはあるの?

違いの証明
=t検定

数学Bあたりでやっている?

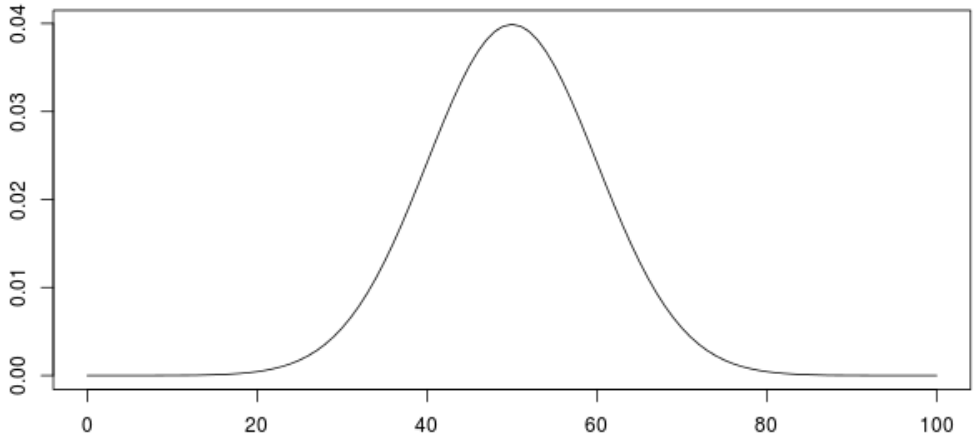
事前知識:サンプリング調査

全数調査は難しい ⇒ 一部を取り出して調査
　一部の調査から全体をどうやって推定するか?

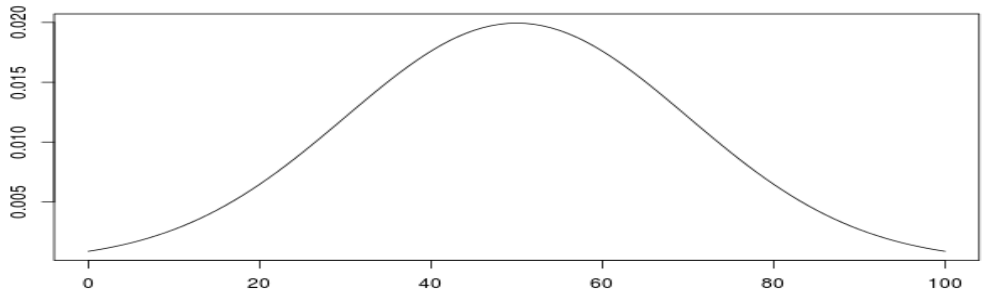


事前知識の事前知識

世の中の自然のものは正規分布する



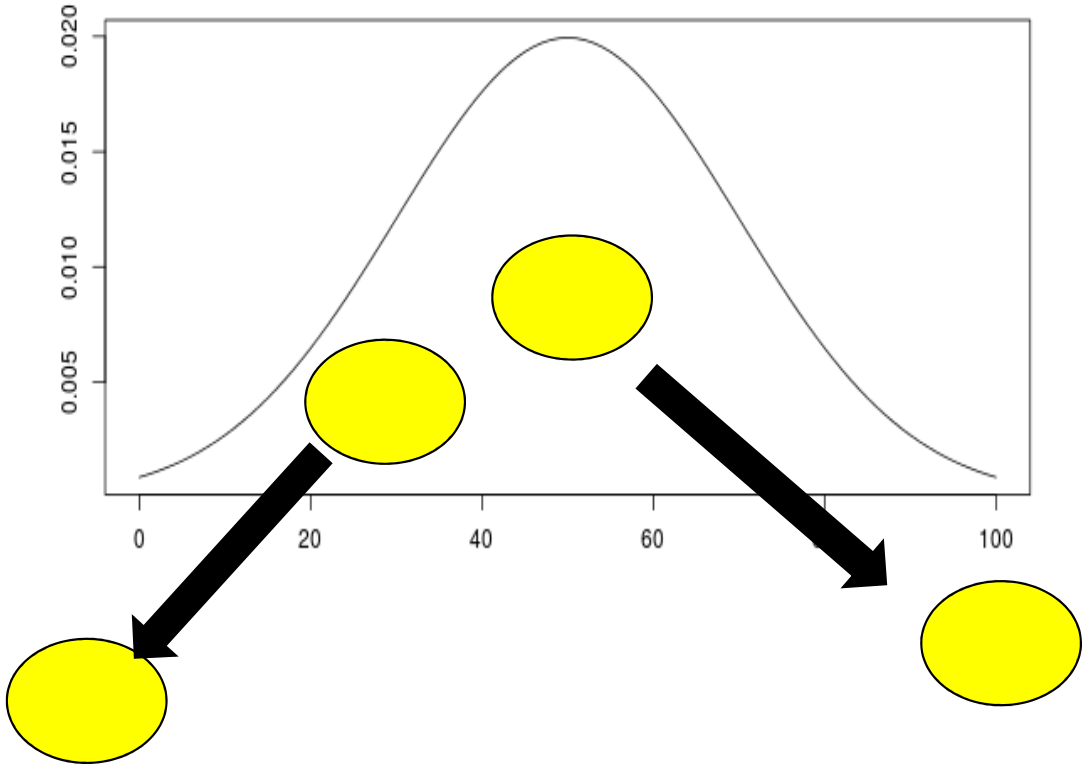
平均 = 50
標準偏差 = 10
の分布



平均 = 50
標準偏差 = 20
の分布

事前知識:標本の平均

世の中の自然のものは正規分布する

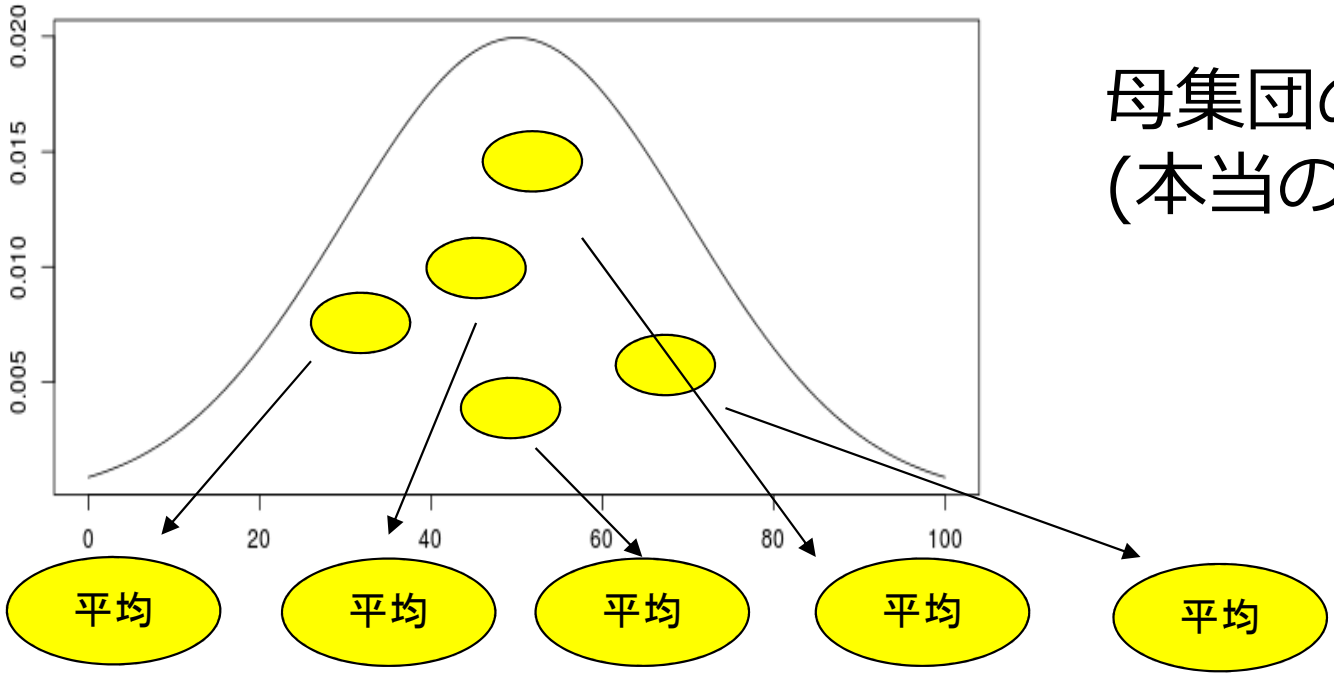


母集団の平均 = 50
(本当の平均)

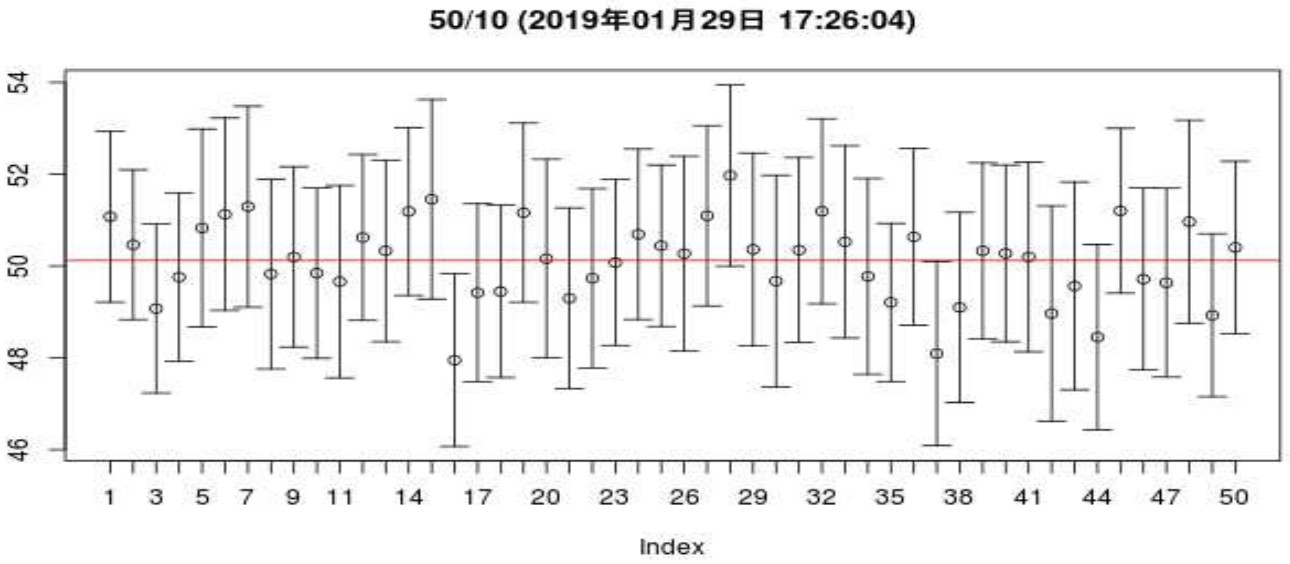
運良くこのあたり
からサンプル
標本の平均 = 50

運悪く、このあた
りからサンプル
標本の平均 = 27

事前知識:標本の平均の分布



母集団の平均 = 50
(本当の平均)

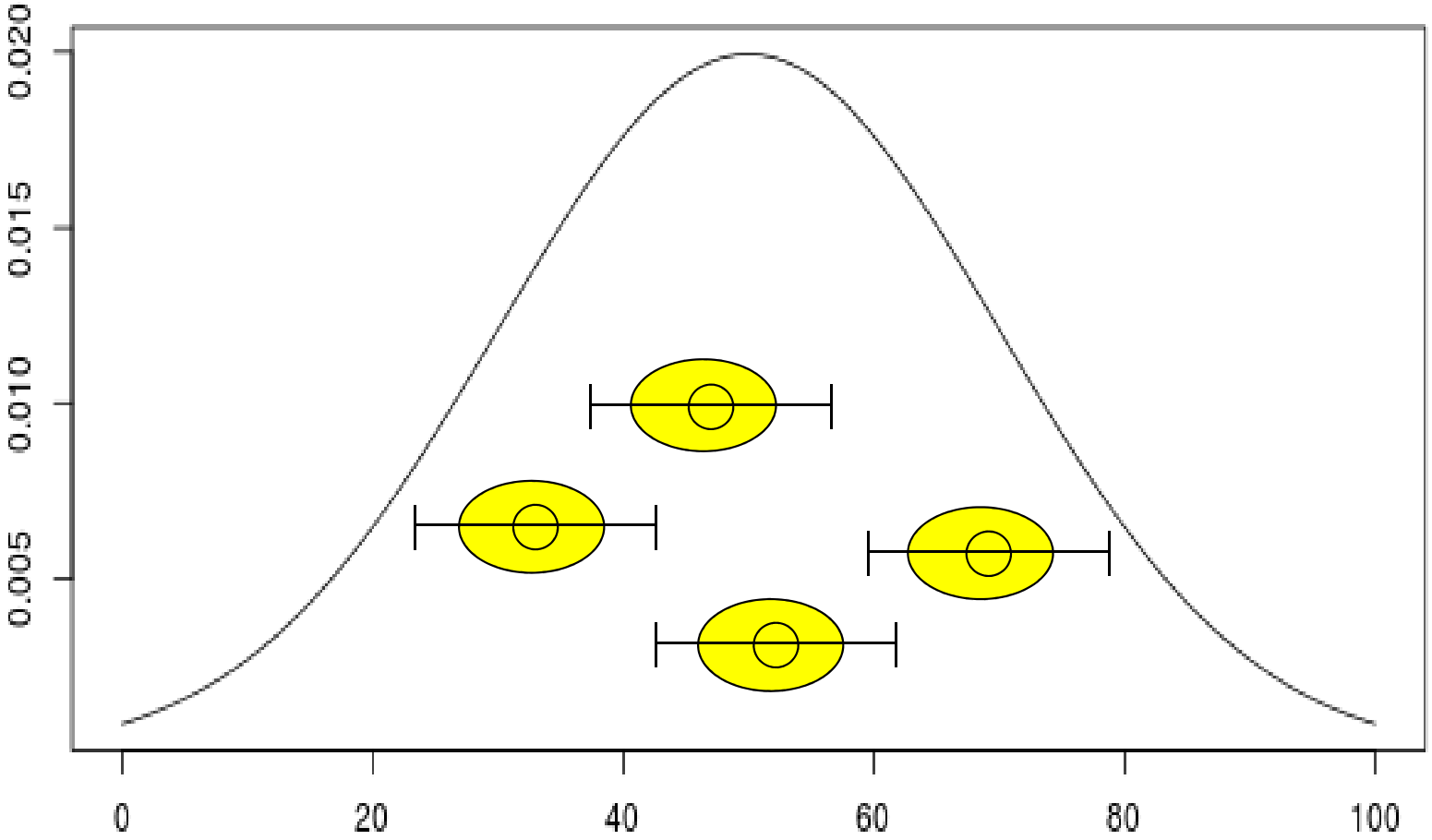


何回も調査すると
結構バラバラ

事前知識:信頼区間: 母集団の平均の推定

母集団の平均 = 50
(本当の平均)

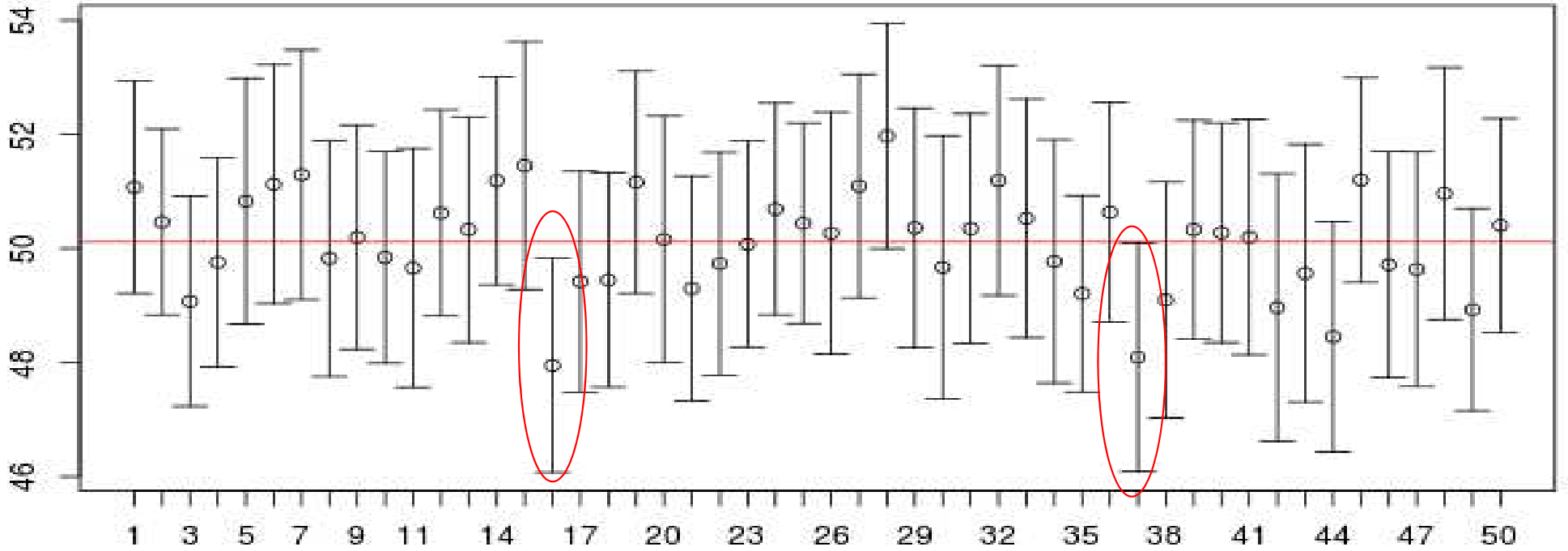
標本調査で出た平均に幅を持たせれば本当の平均が入っているだろう



事前知識:　どれだけ幅をもたせるか

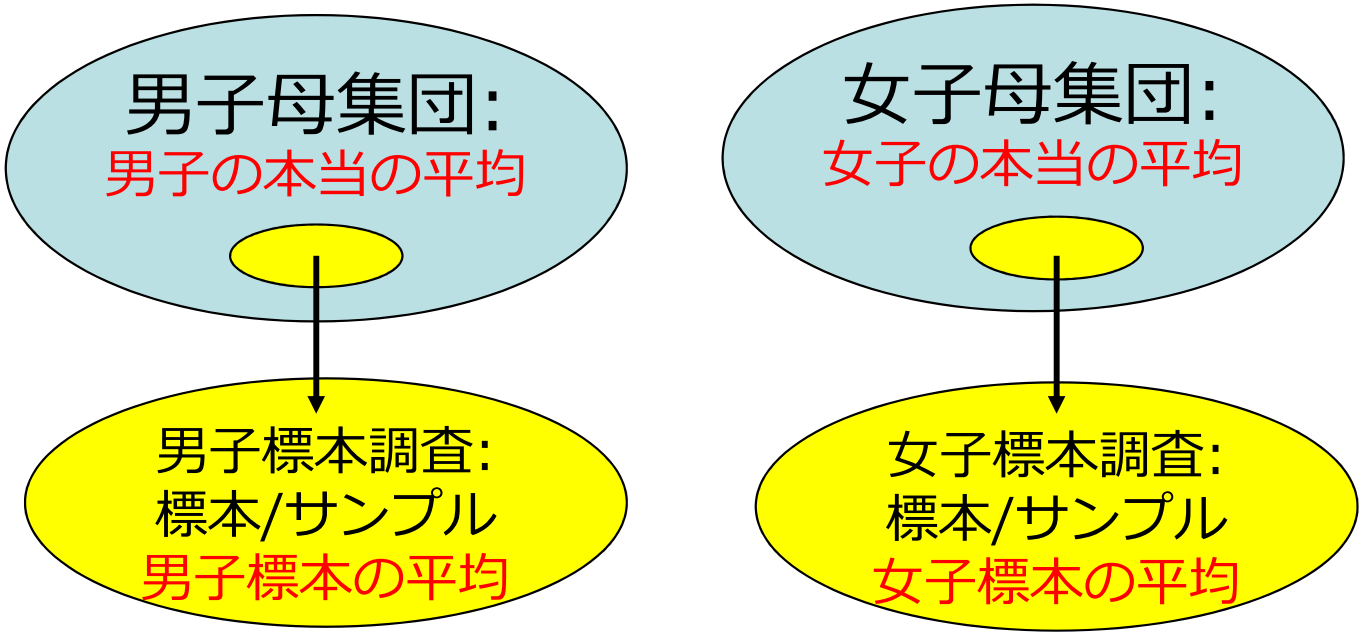
通常: 信頼区間 95% を使用。
おおよそ、この数式の範囲を
母集団の本当の平均と考えよう

$$-1.96 \leq \frac{\overline{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96$$



通常: 信頼区間 95%の意味
例えば、100回標本調査をすれば、95回は、正しい
母集団の平均を示すことができるが、運が悪ければ
5回は誤った母集団の平均を示す。

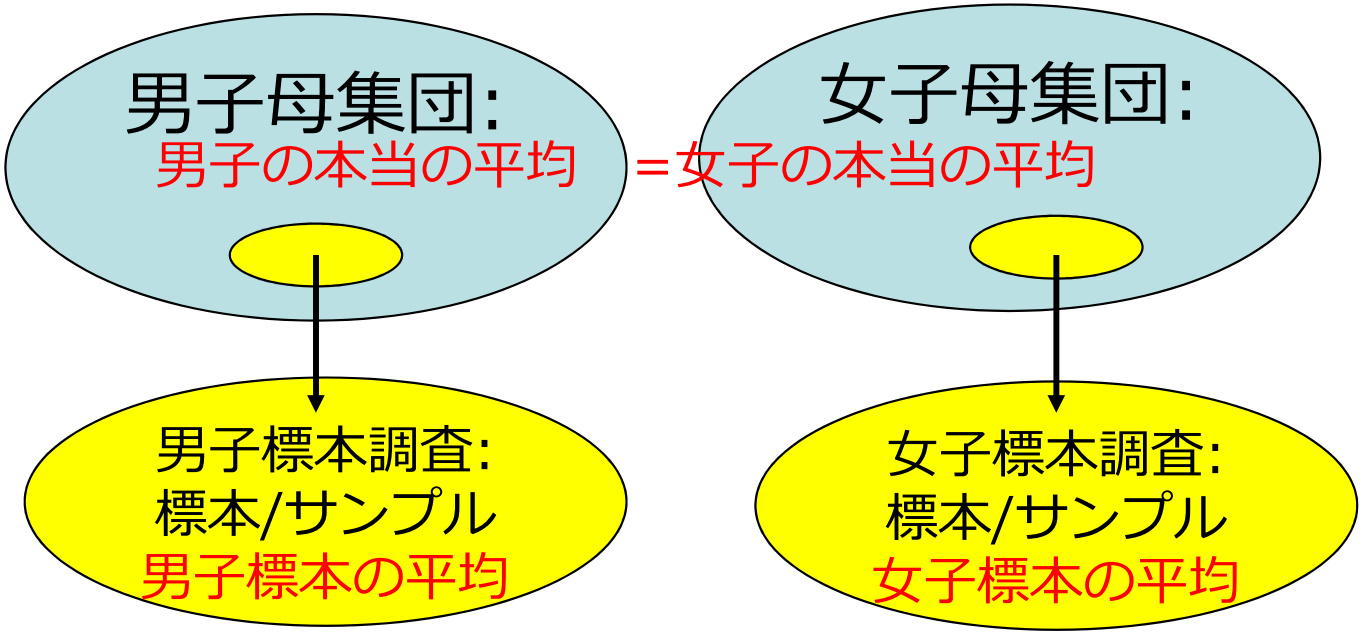
グループで違いがあるか判断する: t検定



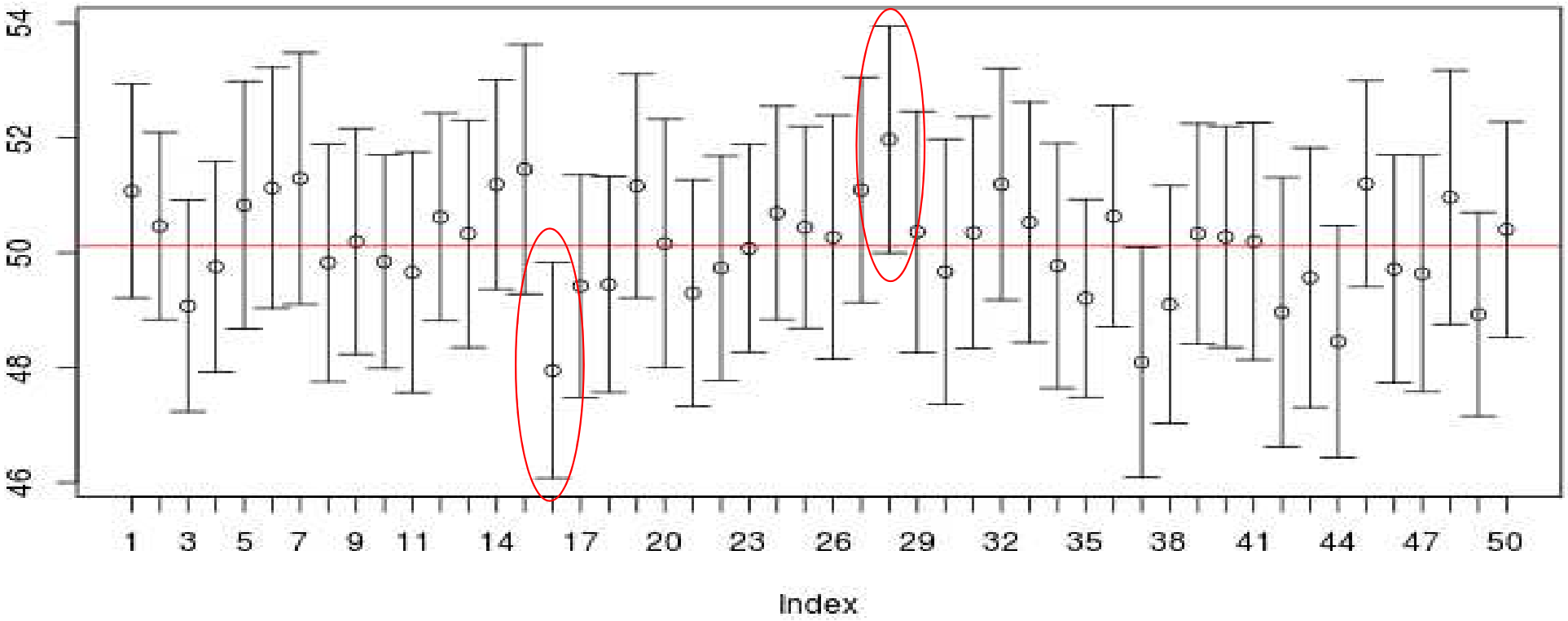
二つの標本調査の結果から、それぞれの母集団の平均が違っているか証明する

t検定:帰無仮説

実は、二つの集団に差が無く母集団の平均は一緒。
ただし標本調査した結果がたまたま違っていた。



t検定:帰無仮説:標本調査の結果がたまたま違う

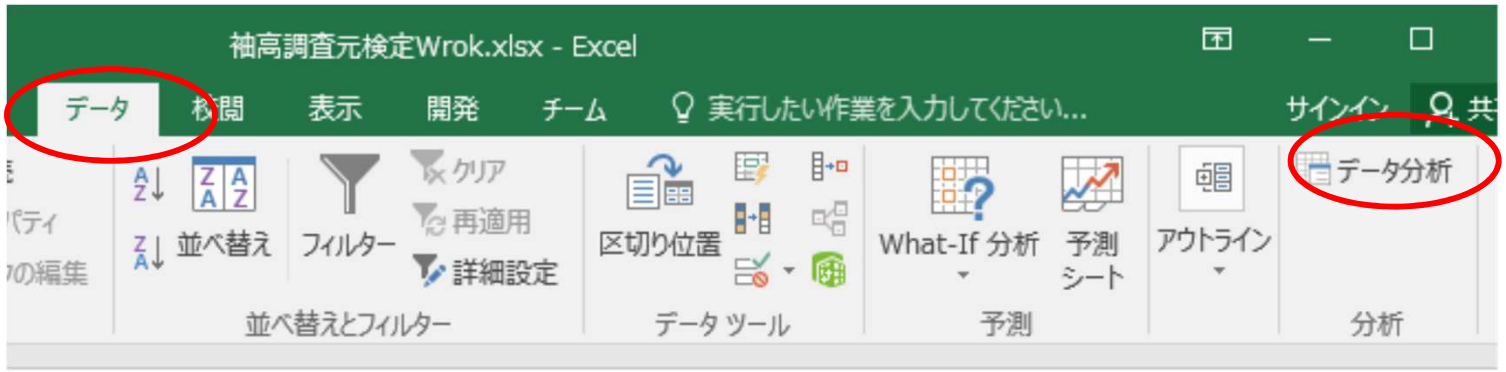


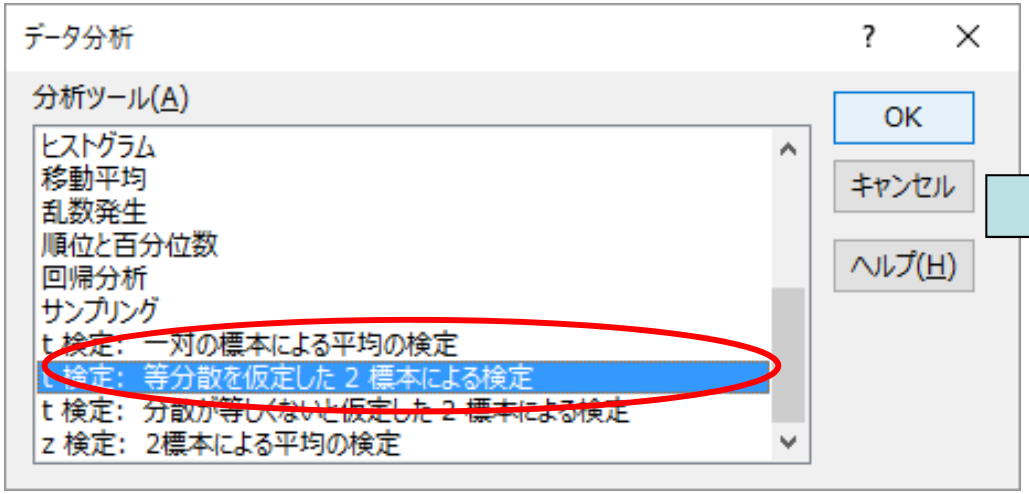
たまたま平均が大きく違う標本調査が2つあった。

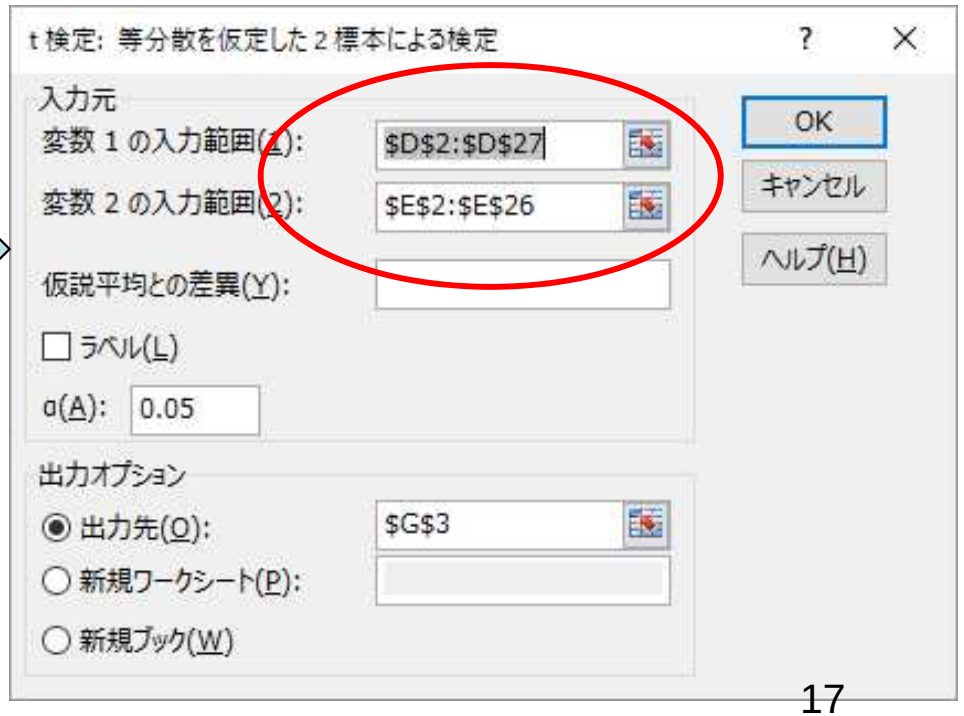
そのような標本が選ばれる可能性はどのぐらい

⇒ t検定

Excelを使ったt検定







Excelのt検定の結果

t-検定: 等分散を仮定した２標本による検定

	変数 1	変数 2
平均	41.73076923	105.4
分散	3299.884615	6166.5
観測数	26	25
プールされた分	4703.94113	
仮説平均との差	0	
自由度	49	
t	-3.31413545	
P(T<=t) 片側	0.000867211	
t 境界値 片側	1.676550893	
P(T<=t) 両側	0.001734422	
t 境界値 両側	2.009575237	

実は、平均が同じと仮定した母集団から、二つの標本調査をした場合に、このように平均が異なるものが選ばれる確率は

0.0017　＝　1.7 %

これは偶然選ばれることはないから、もともと母集団が平均が等しいという考えが間違っているんじゃないの

＝ 帰無仮説の廃棄

＝ もともと母集団に差があった。

Excelのt検定の結果の判断:信頼水準

二つの標本調査をした場合に、このように平均が異なるものが選ばれる確率は

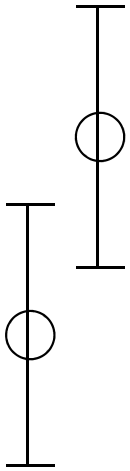
$p < 0.05$　5% より小さい場合

$p < 0.01$　1% より小さい場合

　このぐらいだと、偶然選んだといえないので、帰無仮説の廃棄

この確率が5%以下や、1%以下のことを信頼水準といいます。
5%以上又は10%以上だと、偶然でもそのぐらいの差はでるということで違わないという判断

信頼区間と信頼水準



信頼区間と信頼水準は全く別のものです。
一般的に計算すると、二つの標本データがあった時に、その時、その信頼区間が1/3以上重ならないようだと、その二つのデータの抽出される確率(信頼水準)は5%ぐらいになります。

Excelのt検定の結果の判断の書き方

xxxxとyyyyの間で、平均値間に差があるか対応のないt検定を行ったところ5%水準で有意な差が認められた。

xxxxとyyyyの間で、条件間で平均値間の差について対応のあるt検定を実施した。結果、統計的に有意な差は認められなかった。

注意:
t検定は、有意な差があるかどうか検定するだけであって、差が大きいや小さいを判断するものではありません。

ようやく今日の課題

少ない時間ですが

すでに作成した、袖高の生徒ってどうよパワポに
・相関計算できるもの
・t検定できるもの
について、相関係数又はt検定を行って、その確かさなどを追記してみよう。

来年度の課題研究のデータの分析に向けて(1)

いろいろなt検定

対応のない場合(今回説明したもの)	別々のグループの平均を比較する。	Excel 「t 検定: 等分散を仮定した 2 標本による検定」 又は 「t 検定: 分散が等しくないと仮定した 2 標本による検定」
対応のある場合	同じ人が、 ・時間の経過で変化 ・何か特別なことをした前後(特別なことが有効だったかの判断)	Excel 「t 検定: 一対の標本による平均の検定」

来年度の課題研究のデータの分析に向けて(2)

グループ間の人数などの違いを比較する

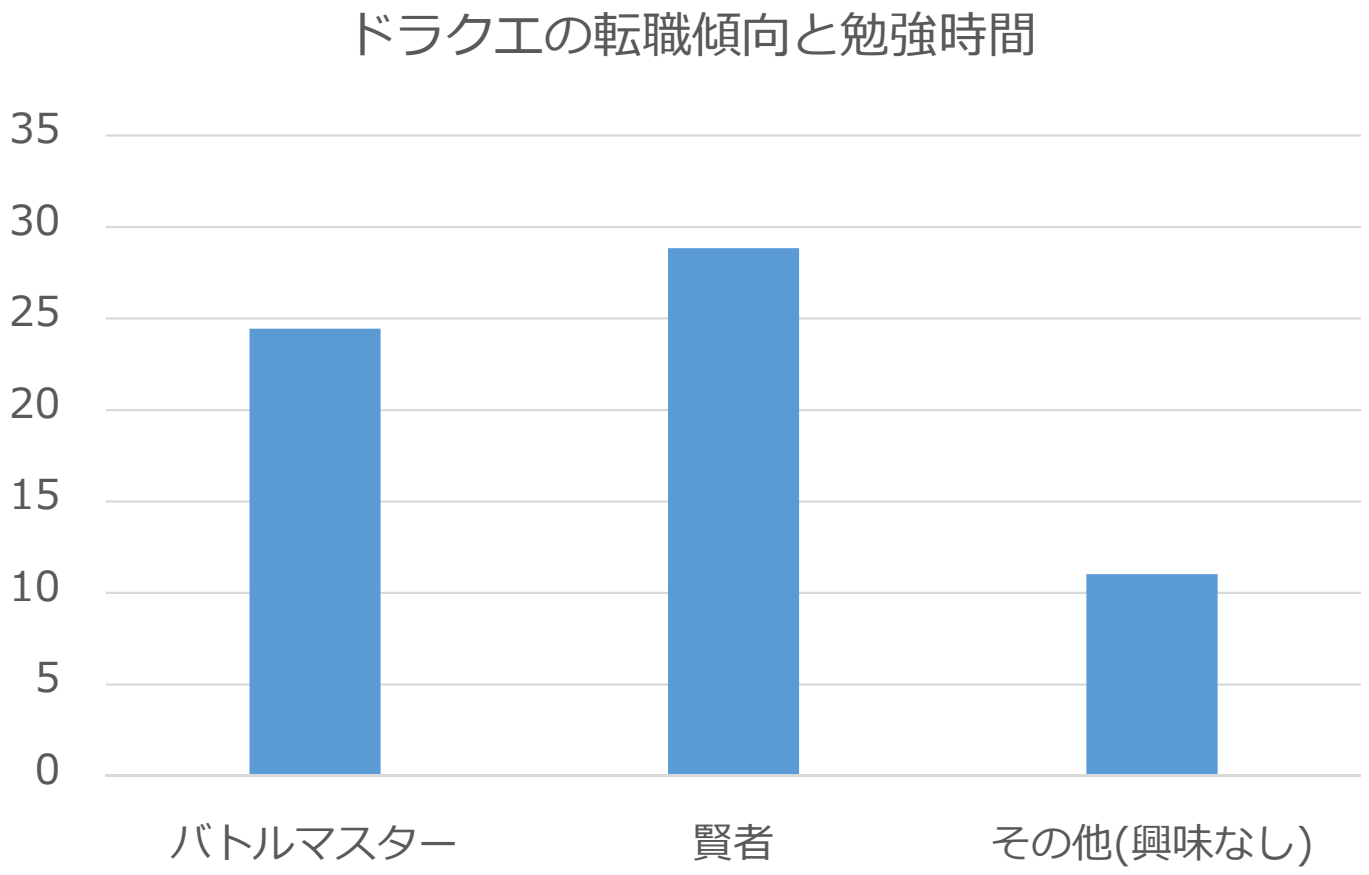
表: 男女別の好みの違いの人数

	ディズニーシー	ディズニーランド
女子	17	12
男子	14	12

カイ二乗検定使います。

来年度の課題研究のデータの分析に向けて(3)

3個以上のグループ間の平均などを判断する



t検定は2グループの比較、それ以上のグループの比較は分散分析使います。