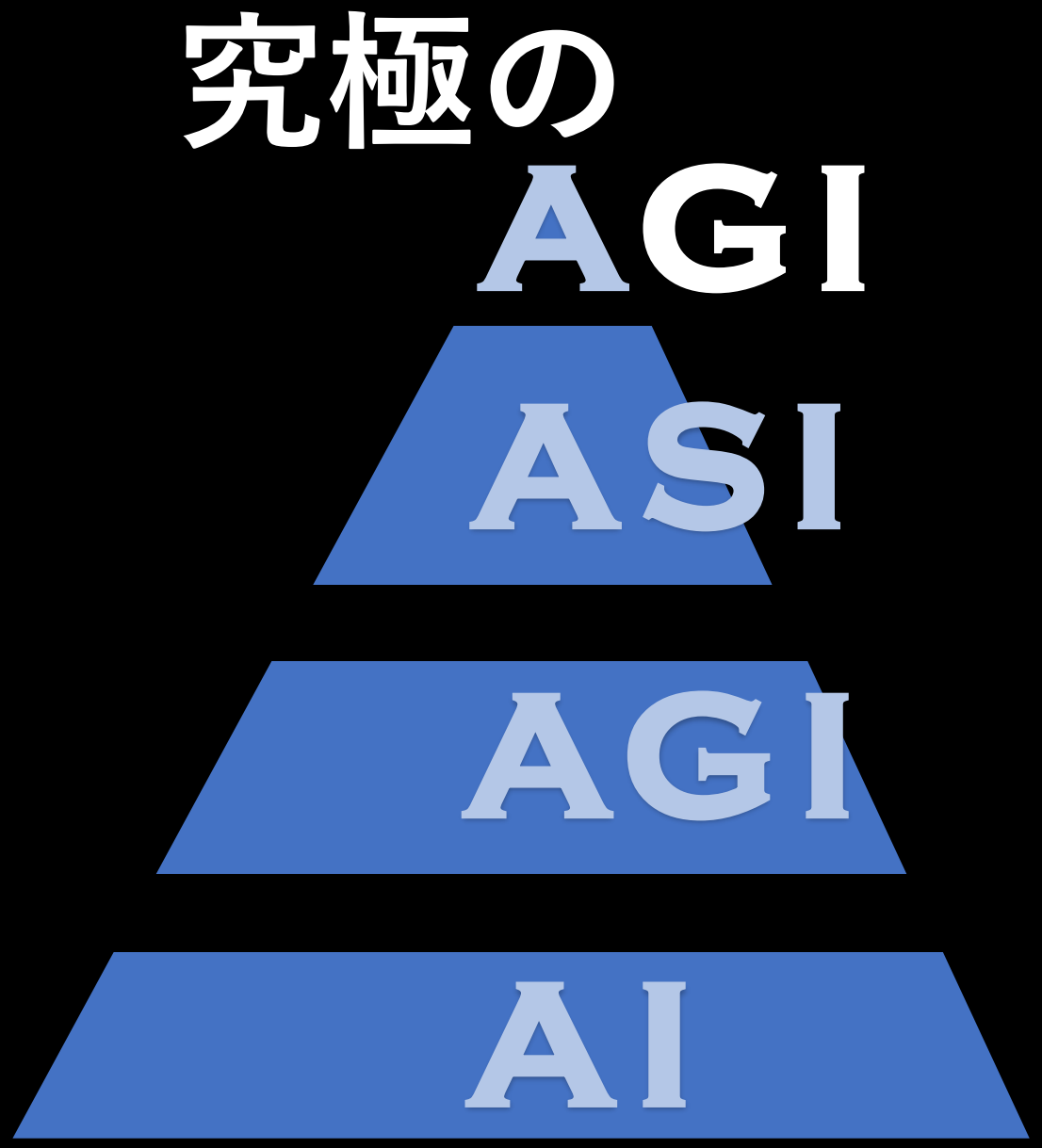


理論 ②

統計 2.0

目次 (1/2)

おさらい 4
統計2.0の構成 5
頻度主義統計の問題 (1) 6
頻度主義統計の問題 (2) 7
ベイズ主義統計の問題 (1) 8
ベイズ主義統計の問題 (2) 9
仮説検定 10
文字認識NN 11
不明の定義 12
不明な初期値 13
不明な標本の追加 14
不明な標本の解釈 15
矛盾 16
大数の法則 17
従来の標本の選択 18
統計2.0の定義 19
標本選択戦略 20
標本 21
論理和の禁止 22
論理否定の禁止 23
標本の区別 24
部分的な重複 25
単位標本 26
標本の情報 27
目的変数と説明変数 28
鍵の写像 29
標本サイズ関数 (1) 30
標本サイズ関数 (2) 31
相対標本サイズ 32
推論に使用する標本 33
連続した単一の範囲 34
クラスタリング 35
解像度 36
論理和の禁止 (2) 37
推論に使う集合 38
恣意的なクラスタリング 39
自由度 40
標本選択の自由 41
座標変換 42
直感的に良い推論 43
推論の優劣と分散 44
最良な帰納の極限 (1) 45

目次 (2/2)

最良な帰納の極限 (2) 46
最良な帰納の極限 (3) 47
理想帰納 48
順位の混合 49
段階的な選択 50
順位の組み合わせ 51
理想率 (1) 52
理想率 (2) 53
バラツキの数値化 54
距離の勝率 55
非理想距離の勝率 56
戦略評価法 (1) 57
戦略評価法 (2) 58
非理想距離の解釈 59
ブラインド推論 60
名義尺度の距離 61
推論の層 62
不明瞭な分離 63
一致率 (名義) 64
一致率 (連続) 65
交換回数 66
不明瞭な分離のウェイト 67
統計学の目的 68
統計における情報 69
生の目的変数 70
1標本からの推論 71
2標本からの推論 (1) 72
2標本からの推論 (2) 73
n標本からの推論 74
重複値の推論 75
名義の推論 76
0標本からの推論 77
推論の合成 (1) 78
推論の合成 (2) 79
回答の量子化 (1) 80
回答の量子化 (2) 81
回答の量子化 (3) 82
統計2.0のまとめ 83
統計2.0と究極のAGI 84
参考文献 85
あとがき 86
次回予告 87

おさらい

①(前回)

・究極のAGIは、どんな質問にでも最適解を答えられるものと定義しました。

・そのためには、最適な帰納推論の方法を決める必要がありました。

・また、究極のAGIが満たすべき条件がいくつか明らかになりました。

②(今回)

・それらの条件を満たすように、最適な帰納推論の方法を決めていきます。

・従来の統計の概念の拡張した「統計2.0」について説明していきます。

前回までのおさらいです。
究極のAGIは、どんな質問にでも最適解を答えられるものと定義しました。
そのためには、最適な帰納推論の方法を決める必要がありました。
また、究極のAGIが満たすべき条件がいくつか明らかになりました。
その条件を満たすように、最適な帰納推論の方法を決めていきます。
それを可能にする、従来の統計の概念の拡張した「統計2.0」について説明していきます。

統計2.0の構成

(0)　統計2.0の従来の統計との違いを定義

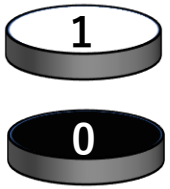
統計2.0
- (1)　有効な標本および、標本の有効な選択方法の定義
- (2)　標本の選択方法の優劣の評価方法の定義
- (3)　選択された標本から推測される値を定義

統計2.0は、三つの部分から構成されます。
まず、統計2.0の従来の統計との違いを定義します。
それから、三つの構成部分を説明していきます。

一つ目は、有効な標本および、標本の有効な選択方法の定義です。
二つ目は、標本の選択方法の優劣の評価方法の定義です。
三つ目は、選択された標本から推測される値を定義です。

頻度主義統計の問題 (1)

コインの
表と裏の数字



過去のコイントス

平均値 = 1

標準偏差 = 0

未来のコイントス

推測値 = 1

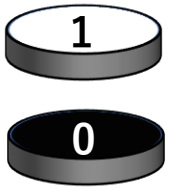
(100%)

100%未来を予測できてしまうのは、明らかに誤り。

はじめに従来の頻度主義の統計の問題点を見ていきます。
簡単な例として、コイントスを2回行い、表(1)が2回、裏(0)が0回出たとします。
このとき、平均値は1で、標準偏差は0です。
2回の試行結果は、100%表が出るという母集団から抽出された標本と推定します。
次の3回目の試行は、同じ母集団から抽出されるため、100%表が出ると推定します。
しかし、100%未来を予測できてしまうのは、明らかに誤りです。
この問題は演繹ではなく帰納です。
標本が増えるほど確証が増すとしても、100%確証することは許されません。

頻度主義統計の問題 (2)

コインの
表と裏の数字



過去のコイントス

平均値 = 1

標準偏差= ∞

(計算不能)

未来のコイントス

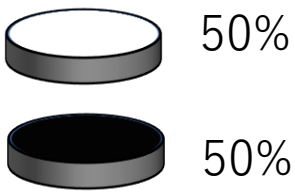
(不明)

標本が1つあるのに、標本が0の場合と同様に、完全な不明なのは誤りです。

次の例として、コイントスを1回行い、表(1)が1回、裏(0)が0回出たとします。
このとき、標準偏差が∞になってしまい計算できません。
しかし、計算できないから何も推定できないというのは、おかしいです。
少なくとも100%裏が出るコインではないことは分かります。
直感的には、どちらかといえば、表が出やすいかもしれないと推測できます。
原理的にも、標本が1つでもあれば、何もないよりは確証が増すはずです。
標本が1つあるのに、標本が0の場合と同様に、完全な不明なのは誤りです。

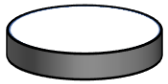
ベイズ主義統計の問題 (1)

主観的な
事前確率



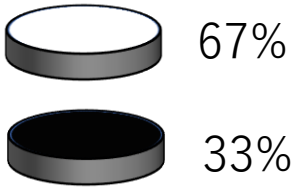
×

客観的な
観測



=

事後確率

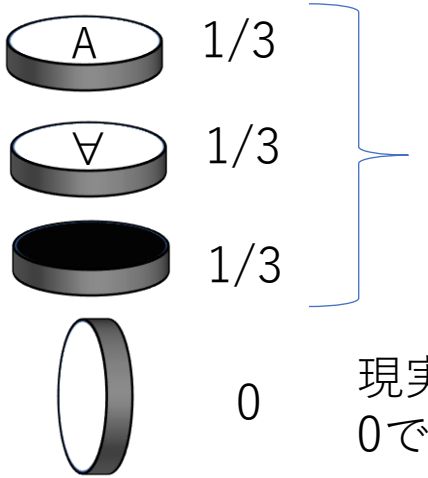


不明だからといって、事前確率を主観で決めてしまうのは誤りです。

ベイズ主義の統計で解釈する場合も考えています。
表50%、裏50%という事前分布を仮定すれば、標本が1つでも推定できます。
無差別の原理に基づいて、取りうる区別可能な2つの状態の一様分布を仮定しています。
不明だからといって、事前確率を主観で決めてしまうのは誤りです。

ベイズ主義統計の問題 (2)

主観的な
事前確率



1/3
1/3
1/3
0

一様分布

現実には
0ではない

3つの状態の頻度についての
情報はない。

取りうる状態がこの3つのだという、
主観的な情報が存在する。

真の無情報なら、
事前確率は「不明」としかいえない。

コインの表側だけは絵柄「A」があり、上向き(A)と下向き(∀)を区別できると仮定してみます。
すると、「表(上向き)、表(下向き)、裏」の1/3ずつが、一様分布になります。
また、コインが倒れずに立つ可能性を0と決めつけていますが、現実には0ではありません。
この一様分布は、無情報事前分布と呼ばれますが、真の無情報ではありません。
3つの状態の頻度についての情報はありません。
ですが、取りうる状態がこの3つのだという、主観的な情報が存在します。
真の無情報なら、事前確率は「不明」としかいえません。

仮説検定

値の情報　：恣意的に追加しない (事前確率)

分布の情報：恣意的に追加する (正規分布など)

質問: ある薬が効くか効かないか?

結果:　~~「効く」または「効かない」~~

「効く」なたは「不明」

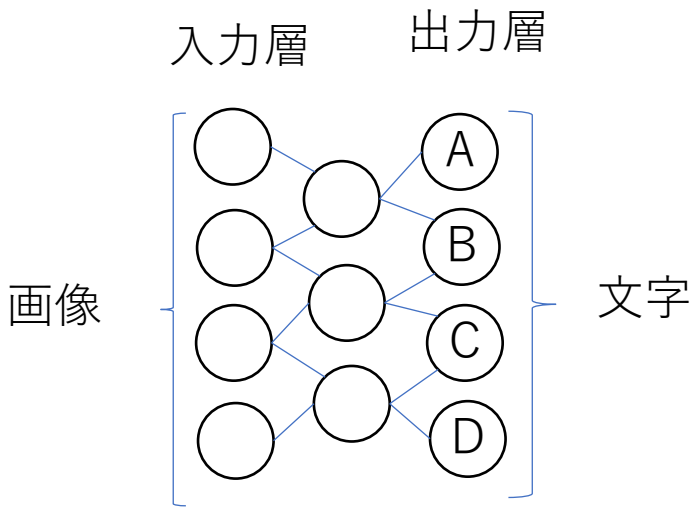
↑

標本が極端に少ない場合

一方、統計主義統計での情報について考えます。
ある薬の効果があるかどうかの仮説検定を考えます。
事前確率のように、勝手に新しい値の情報を追加することはありません。
ですが、正規分布のように、勝手に値の分布に関する情報を追加します。
値の情報を追加しないので、標本が極端に少ないと、まともな結果は得られません。
誤解しやすいのが、仮説検定では、「薬効あり」または「薬効なし」のどちらなのかを判断するのではありません。
「薬効あり」または「不明」のどちらかを判断するものです。
標本が極端に少ないと「分からない」という結果になります。

文字認識NN

文字認識ニューラルネットワーク



初期値
(事前確率)

：ランダムな
情報

初期値と学習に使った情報は混ざってしまい、区別できない。

次に、文字認識のニューラルネットワークが行っている統計的推論について考えます。
出力層のニューロン1つが、1つの文字に対応しているとします、
ニューロンの初期値はランダムに設定され、それは事前確率に相当します。
1つの標本でしか学習していなければ、ほとんどランダムな初期値が出力されます。
ランダムに決めた初期値にも、情報量があります。
また、初期値と学習に使った情報は混ざってしまい、区別できません。

不明の定義

統計 2.0では

ランダムな初期値のような、
質問者が「利用可能な情報」と指定していない情報の使用は禁止

「不明」の定義:

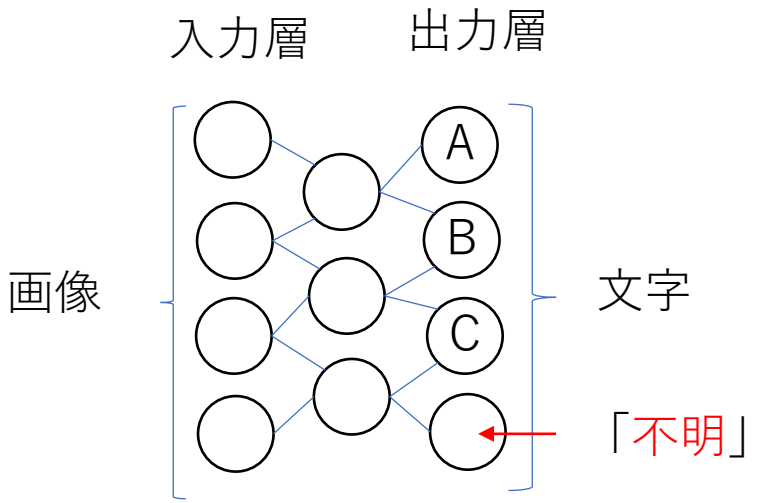
情報がないことを表す特別な値

	「不明」の扱い
頻度主義統計	「不明」ではない値だけを扱う
ベイズ主義統計	「不明」に主観的な値を代入する
統計 2.0	「不明」は「不明」のまま扱う

ここから、新しい「統計2.0」について考えていきます。
ランダムな初期値のような、質問者が「利用可能な情報」と指定していない情報の使用は禁止します。
情報がないことを表す、「不明」という特別な値を定義します。
「不明」の扱いを表にまとめました。
頻度主義統計では、「不明」ではない値だけを使って推定します。
ベイズ主義統計では、「不明」に主観的な数値を代入して推定します。
統計2.0では、「不明」は「不明」を保ったまま推定します。

不明な初期値

文字認識ニューラルネットワーク



初期値
(事前確率)　：「不明」
(100%)

「不明」以外の値を学習するほど、
「不明」の出力が小さくなる。

「不明」という初期値を保ったまま推定する方法を考えます。
文字認識のニューラルネットワークの例で考えます。
出力ニューロンの一つが「不明」を表すと決めます。
学習前は、ランダムではなく、100%「不明」が出力されるようにします。
「不明」以外の値を学習するほど、「不明」の出力が小さくなります。

不明な標本の追加

標本サイズ		一般的な学習率		出力	
n = 9	×	0.1%	=	0.9%: 学習データ	
				99.1%: 初期値 「不明」	

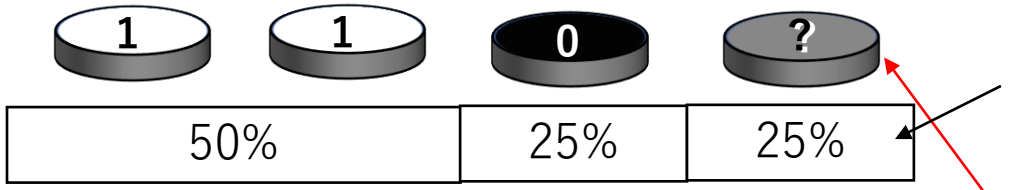
標本サイズ		最適な学習率			n
n = 9	×	10%	=	90%: 学習データ	(9×10%)
				10%: 初期値「不明」	(1×10%)

「不明」という標本を一つ(n=1)追加した上で、
全ての標本が均等に出力に影響するのが最適。

ニューラルネットワークの学習率は0.1%程度が一般的です。
出力の内、学習されていない割合の分は、初期値の「不明」のままです。
標本サイズがn=9の場合、学習率の合計は0.9%なので、残りの99.1%は初期値の「不明」です。
標本サイズが小さいことを知っているなら、学習率を上げるべきです。
n=9なら、学習率は10%が最適と考えられます。
9個の標本が、10%ずつ出力に影響を与え、残りの10%が初期値の「不明」です。
つまり、「不明」という標本を一つ追加した上で、全ての標本が均等に出力に影響するのが最適です。

不明な標本の解釈

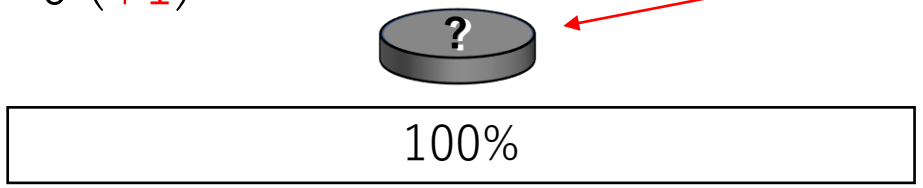
n = 3 (+1)



Impact on output
(probability distribution)

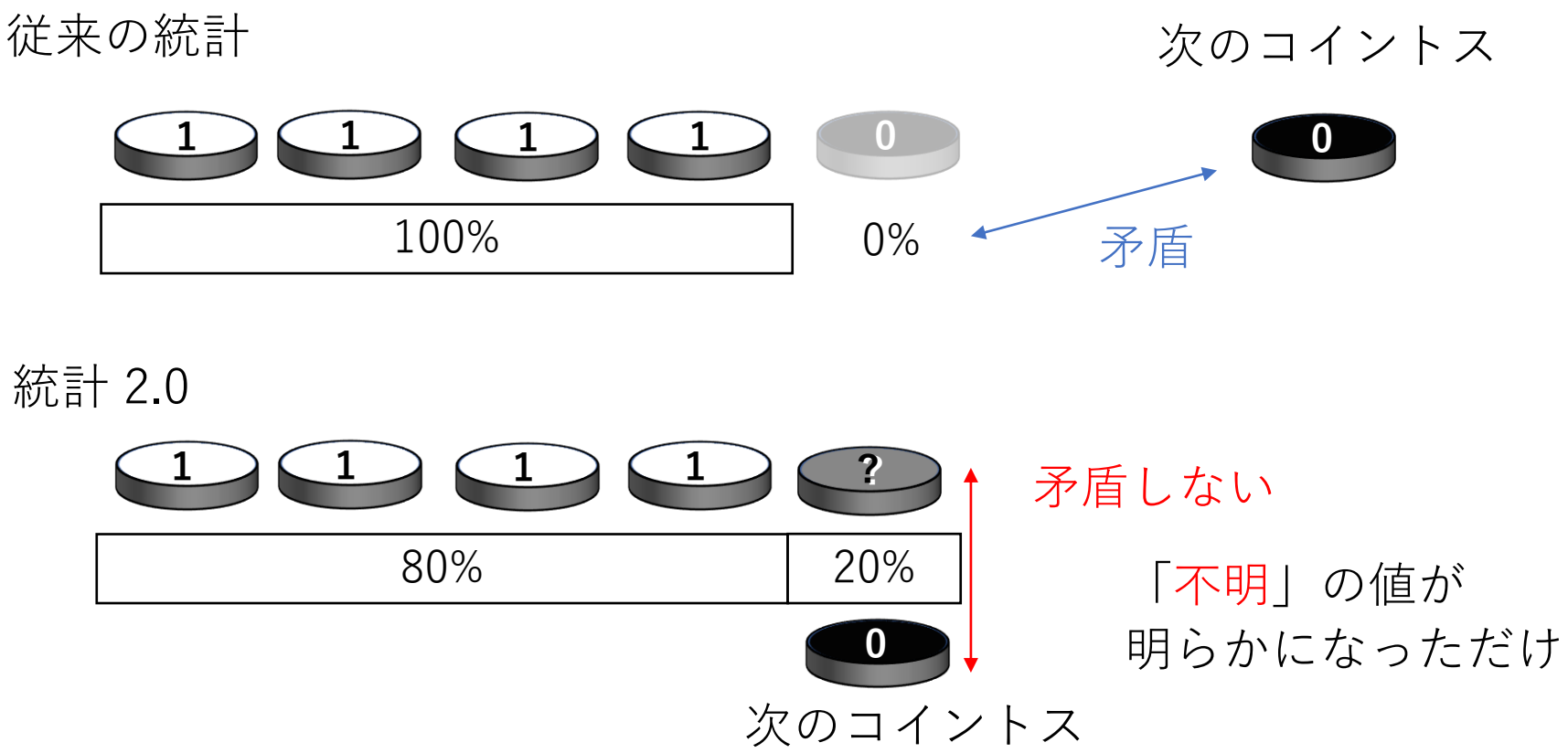
「不明」な標本は、
推測しようとしている標本
だと解釈できる

n = 0 (+1)



コイントスで表が2回、裏が1回出た場合を考えてみます。
「不明」な標本を一つ追加します。
表が50%、裏が25%、「不明」が25%、出力に影響します。
出力への影響の大きさの確率分布は、そのまま推測値とみなすこともできます。
ここで、「不明」の標本は、推測しようとしている標本そのものと解釈できます。
n=0なら、推測しようとしている標本そのもののみから推測します。
推測しようとしている標本の値は「不明」なので、推測値もまた「不明」=100%になります。
なぜ、1個追加するのかというと、推測対象が1個だからです。

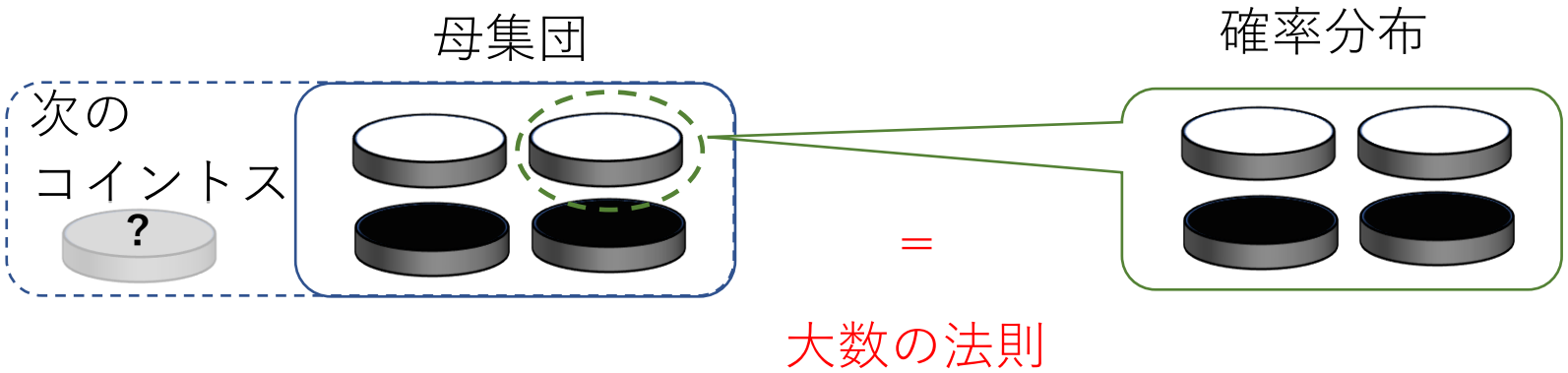
矛盾



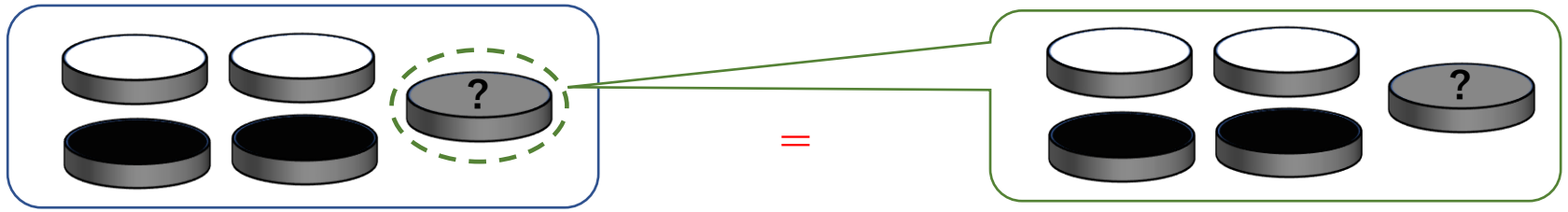
従来の統計では、値が既知の標本だけで、母集団を推定します。
「統計2.0」では、推測しようとしている「不明」という値の標本を含めて母集団を推定します。
コイントスで、表が4回、裏が0回出た場合、従来の統計では、表が100%出ると推測します。
もし、次のコイントスで裏が出た場合、推測値と矛盾します。
一方、統計2.0では、表が80%、「不明」が20%出ると仮定します。
次のコイントスで裏が出ても、それは「不明」の値が明確になっただけで、矛盾しません。

大数の法則

従来の統計

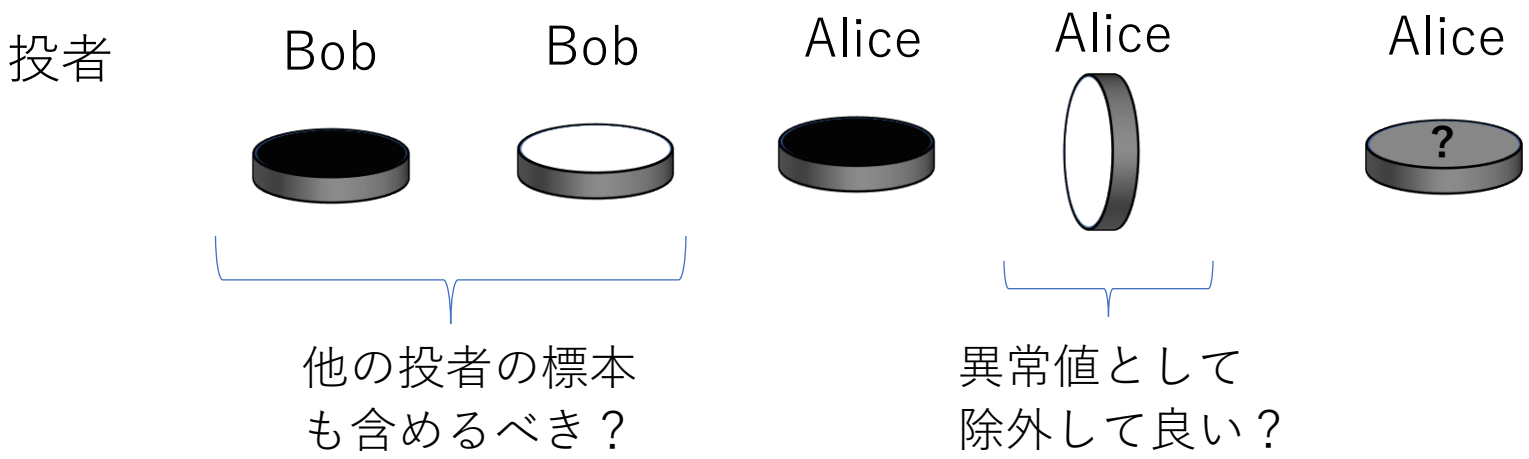


統計 2.0



次回のコイントスの推測値が、どのように正当化されているのか考えてみます。
まず、標本は母集団から無作為抽出されたと仮定しています。
すると、一つの標本の値の確率分布は、母集団内の各標本の値の分布と等しくなります。
いわゆる大数の法則です。
従来の統計では、次回のコイントスも同じ母集団に入っているだろうと仮定します。
もし、その仮定が誤りならば、矛盾する結果が得られます。
一方で、「統計2.0」では、次回のコイントス結果を「不明」な標本として含んでいます。
「不明」な標本の値の確率分布もまた、母集団内の各標本の値の分布と等しくなります。

従来の標本の選択



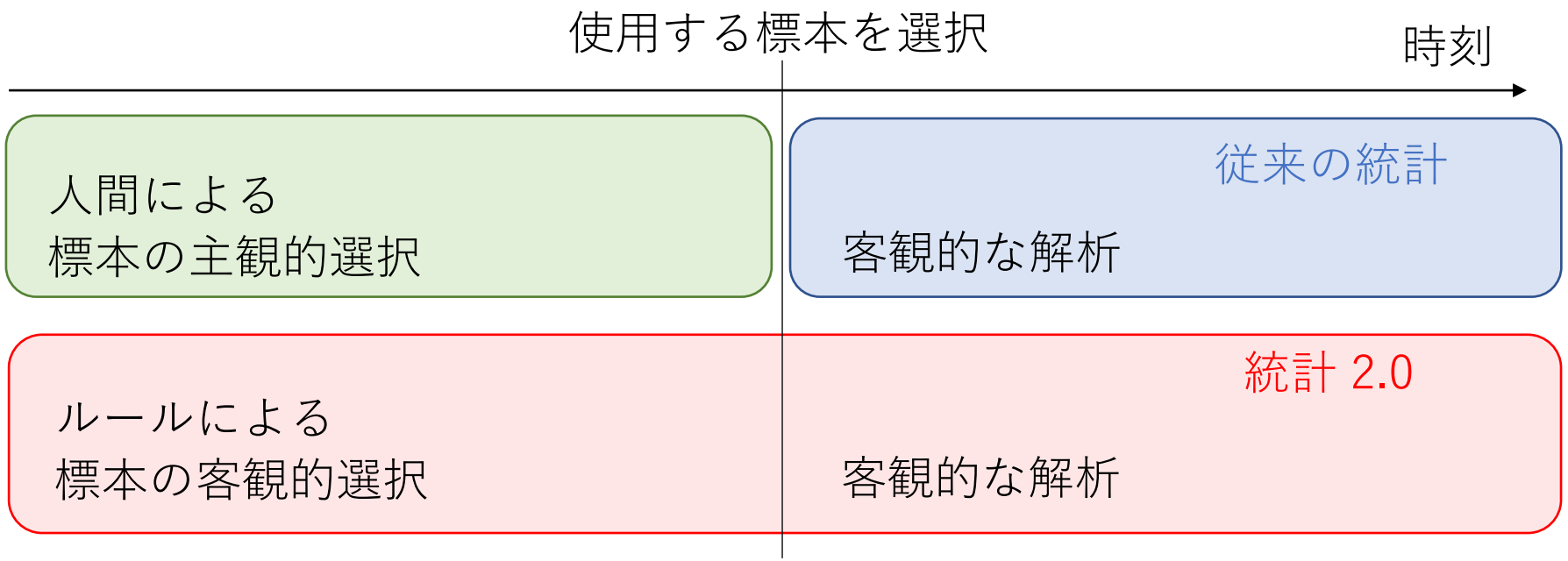
どの標本を使用するかは、人間が主観で判断する。

推測結果は、人間の判断に大きく依存する。

統計処理プログラムは、入力された標本はすべて同じ重みだと見做す。

次に、従来、統計に使う標本はどうやって決めたのか考えます。
次のコイントス結果を推測する場合は、過去のコイントス結果を標本とするでしょう。
その中には、違う人物が投げたデータもありますが、それも含めて良いでしょうか？
また、異常だと感じたデータがあれば、それは除外しても良いでしょうか？
どの標本を使用するかは、人間が主観で判断します。
推測結果は、人間の判断に大きく依存します。
統計処理するプログラムは、入力された標本はすべて同じ重みだと見做します。

統計2.0の定義



「統計 2.0」の定義:

客観的な議論の範囲を、
使用する標本を選択する段階まで拡げた統計

ここで、ようやく「統計2.0」が何なのかを定義します。
統計とは、標本を使用して客観的な解析を行うものです。
ですが、従来の統計では、使用する標本は人間が主観的に決めていました。
統計2.0では、ルールに従って客観的に使用する標本を決めます。
客観的な議論の範囲を、使用する標本を選択する段階まで拡げた統計が、「統計2.0」の定義です。

標本選択戦略

「標本選択戦略」:　どの標本を推論に使用するかという戦略

統計2.0の目的:　出来るだけ良い「標本選択戦略」の探索

「標本選択戦略」のルール:

- 必須条件
  有効な「標本選択戦略」を規定
- 評価基準
  複数の「標本選択戦略」の優劣を規定

どの標本を推論に使用するかという戦略を、「標本選択戦略」と呼ぶことにします。
統計2.0の目的は、出来るだけ良い「標本選択戦略」の探索です。
ルールとして、「必須条件」と「評価基準」があります。
「必須条件」は、有効な「標本選択戦略」を規定します。
「評価基準」は、複数の「標本選択戦略」の優劣を規定します。
「必須条件」の方から決めていきましょう。
その前に、「標本」とは何かを決めていきます。

標本

“標本“:　　同じ条件を満たすが、区別ができる個

“同じ条件“:　　論理和(OR)と論路否定(NOT)を含まない条件式

OR “A“　　：無効な標本条件

OR 　　：無効な標本条件

果物　AND　赤　　：有効な標本条件

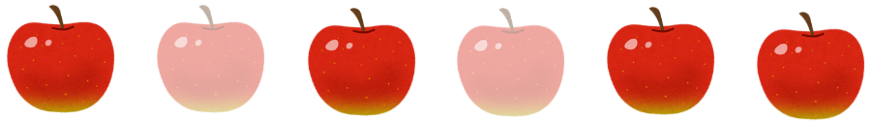
標本は、「同じ条件を満たすが、区別ができる個」と言い換えることができます。
「同じ条件」とは、「論理和と論路否定を含まない条件式」だと規定します。
論理和を認めると、「(リンゴ1個)または(アルファベット1文字)」といった定義も可能になってしまいます。
確証性の原理では、関係があるものを観測するほど、確証性が増すとされています。
論理和を認めると、全く共通点がなくても標本として使えてしまうのは、原理に反します。
例えば、「(リンゴ1個)または(イチゴ1個)」という条件も認められません。
「(フルーツ1個)かつ(赤色)」という条件にすれば認められます。

論理和の禁止

質問: あるリンゴの糖度は?

無効な標本条件:　左から1番目または3番目または5番目のリンゴ

糖度: 8.8　6.3　9.5　5.9　9.2　?



恣意的な選択が可能なため、標本条件に論理和を認めるのは禁止。

有効な標本条件: リンゴ

標本の条件が決められたら、
条件に一致するものは全て列挙しなければいけない。

標本の条件に論理和を認めてしまうと、恣意的な選択が許されてしまいます。
例えば、あるリンゴの糖度を推測したいとします。
「左から1番目または3番目または5番目のリンゴ」が標本として認められる条件だとします。
恣意的に、糖度が高いものだけを標本としても認めれば、推測値の糖度も高くなります。
この例では、単純に「リンゴ」を標本の条件とするべきです。
標本の条件が決められたら、条件に一致するものは全て列挙しなければいけません。

論理否定の禁止

ド・モルガンの法則

標本条件: $A\ OR\ B = NOT(\bar{A}\ AND\ \bar{B})$

禁止　禁止

論理否定を標本条件に含めるのも禁止。

A: $x = a$ ：有効な標本条件

$\bar{A}$: $NOT(x = a)$ ：無効な標本条件

$x \neq a$ ：有効な標本条件

論理否定の代わりに、不等号の反転で表現できれば、
有効な標本条件。

論理否定が、標本の条件として認められない理由を説明します。
ド・モルガンの法則を使うと、論理和を論理積の論理否定に変換できます。
そのため、論理否定も標本条件に含めるのを禁止します。
ただし、論理否定の代わりに、不等号の反転で表現できれば、有効な標本条件になります。

標本の区別

質問： ある人の体重は？

人	Alice	Bob (昨年)	Bob (今年)
体重	44.5kg	55.5kg	55.6kg

成長して変化しているので、標本2個？

同一人物のデータが重複しているだけなので、標本1個？

どちらが良いかは、何を推測したいのかによって異なる。

標本の区別の仕方は、自由に設定することができる。

標本は、「区別ができる個」であることも必要です。
例えば、ある体重を推測したいとします。
しかし、Bobのみ、昨年と今年で2つデータがあります。
二人のBobは、成長して変化しているので標本が2個と考えて良いでしょうか？
そのとも、同一人物なので、データが重複しているだけで、標本は1個でしょうか？
どちらが良いかは、何を推測したいのかによって異なります。
標本の区別の仕方は、自由に設定することができます。
ただ、質問者が意図した通りに設定しないと、意図した通りの結果が得られないだけです。

部分的な重複

質問: 数列の次の数値は?

標本: 1で始まり5で終わる数列

情報: 1 3 5 7 9 1 3 5 7 ?

部分的に重複しているけど、標本は2個？
それとも、標本は1個？

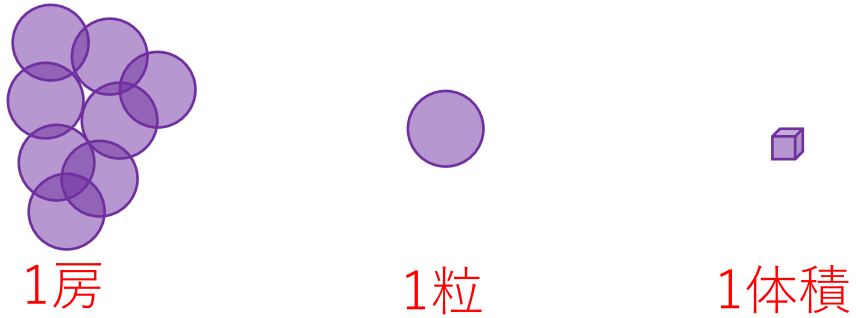
重複の定義、あるいは標本個数の定義は、
自由に設定することができる。

標本が部分的に重複しているような場合、個と認めて良いのか迷います。
数列の次の数字を推測する場合を考えます。
例えば、1で始まり5で終わる数列を、一つの標本とします。
このような定義だと、部分的に重複する標本が現れます。
部分的に重複してますが、標本が2つあるといってよいでしょうか？
それは、ケースバイケースで異なります。
重複の定義、あるいは標本個数の定義は、自由に決めることができます。

単位標本

「単位標本」：　標本サイズが1の標本

質問: あるブドウの糖度



重複の判別ができて、数えられさえすれば、
「単位標本」は自由に定められる。

標本サイズが1の標本は、「単位標本」と表現できます。
単位標本は、質問者に指定されなければ、自由に決めることができます。
例えば、あるブドウ一粒の味を推測したいとします。
一房を、標本1個と定めることができます。
一粒や、単位体積を、標本1個と定めることもできます。
重複の判別ができ、数えることさえできれば、「単位標本」は自由に決められます。
とはいえ、一粒の味を推測したいなら、一粒を単位にするのが良いでしょう。

標本の情報

単位標本 (1人):　手足のある一塊の体積?

「単位標本」の情報:　~~人体の立体形状データ~~

人体と衣服の境界は
必ずしも明確ではない。

区別するための「鍵」となる情報
(例: 体の中心座標)

他人と区別さえできれば良いのなら、
体の中心座標を知っているだけで十分。

さまざまな概念を「標本」と定義することが可能ですが、情報としてどう表現できるのか考えてみます。
例えば、「手足のある一塊の体積」を「一人」だと定義するとします。
しかし、人体と衣服の境界は、必ずしも明確ではありません。
そのため、1標本ごとに、人体の立体形状データを保持するのは困難です。
ですが、他人と区別さえできれば良いのなら、体の中心座標を知っているだけで十分です。
各標本は、区別するための「鍵」となる情報を保持しているだけで十分です。

目的変数と説明変数

目的変数: 推測しようとしている値

説明変数: 標本の絞り込みなどに使う値

これらは、標本の「鍵」の写像

一意に決まる値ならば何でも、説明変数や目的変数として使用できる

標本は、標本同士を区別する「鍵」以外の情報も持っています。
「目的変数」は、推測しようとしている値です。
「説明変数」は、標本の絞り込みなどに使う値です。
これらの値は、標本に紐づけられた値です。
すなわち、「鍵」の写像です。
つまり、一意に決まる値ならば何でも、説明変数や目的変数として使用できます。

鍵の写像

質問: 数列の次の数値は?

座標:	1	2	3	4	5	6	7	8	9	10
	1	3	5	7	9	1	3	5	7	?

単位標本: (座標Aから座標Bの間の数列)
AND (座標Aにある数値=1)
AND (座標Bにある数値=5)

鍵: (座標A, 座標B) =(1, 3), (1, 8)

鍵の写像

目的変数: ○ 座標Bから二つ先にある数値

説明変数: □ 座標Aにある数値 ◇ 座標Bにある数値

「鍵」の写像について、数列を予測するときの例を示します。
座標Aから座標Bの間の数列を単位標本とします。
座標Aと座標Bが、標本を区別する鍵です。
目的変数は、座標Bの二つ先にある数値です。
説明変数は、座標Aと座標Bにある数値です。
これらの変数は、鍵の写像です。
二つの説明変数には、条件が設定されています。

標本サイズ関数 (1)

```
FUNCTION 標本数(任意の情報){
    …
    RETURN 標本数;
}
```

(引数) 任意の情報: 鍵となりうる任意の情報

(戻り値) 標本サイズ : 引数内に含まれる標本サイズの合計　　≧ 0

標本サイズ(鍵) = 0:　　有効な標本が存在しない

標本サイズ(A ∪ B) =
標本サイズ(A)+標本サイズ(B):　　鍵Aの標本と鍵Bの標本は重複しない

次に、標本サイズや標本の区別の取り決めが、どういう情報で表されるのか考えます。
それらは、「標本サイズ関数」として表現できます。
引数は任意の情報で、「鍵」となりうる情報を何個でも渡せます。
戻り値は、引数内の標本サイズの合計を表す0以上の実数です。
1つの「鍵」を関数に渡せば、その標本サイズが分かります。
戻り値が0ならば、有効な標本ではないと分かります。
2つの「鍵」を関数に渡すことで、重複を検査できます。
2つの「鍵」の標本サイズが、一つずつの標本サイズの和と等しければ、重複はありません。

標本サイズ関数 (2)

引数の例: (x1, x2, y1, y2, z1, z2)

有効な戻り値の例: (x2-x1)(y2-y1)(z1>0)(z2<0)/C

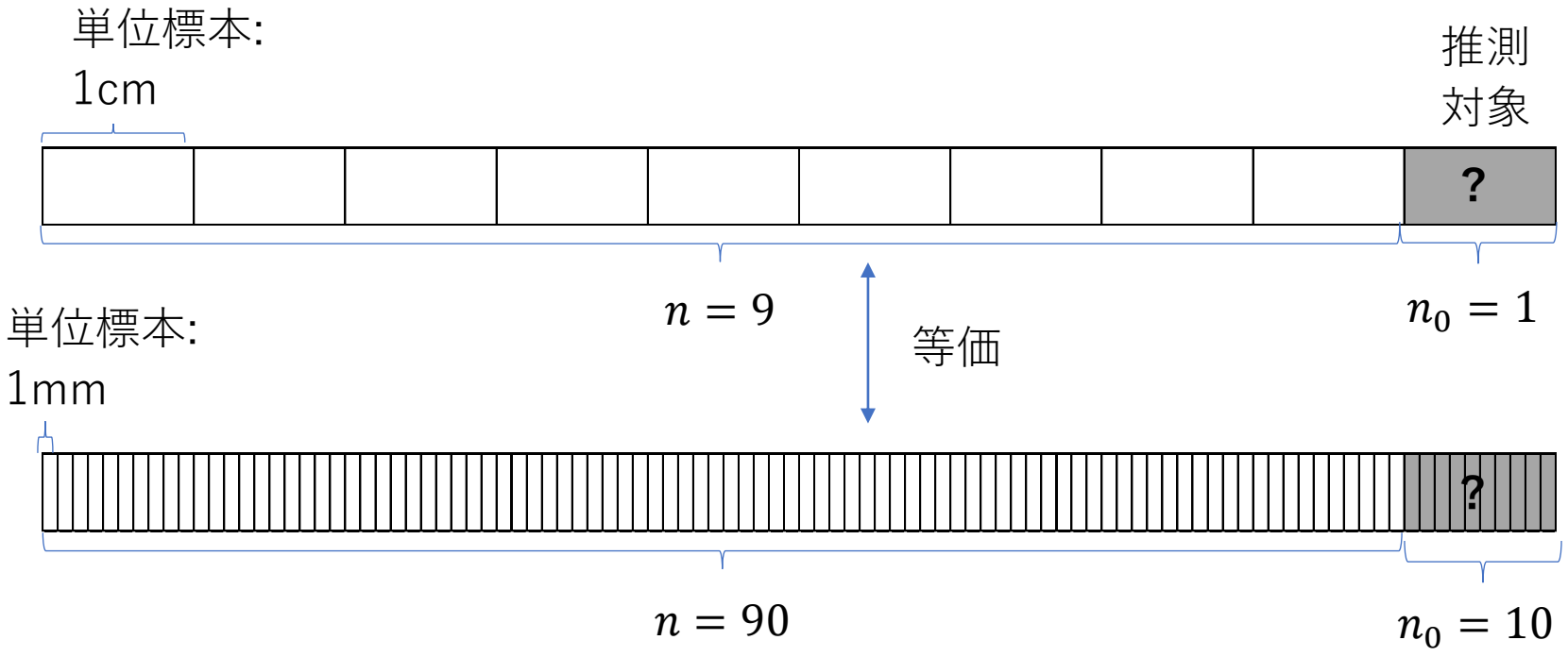
無効な戻り値の例: (x2-x1)+(y2-y1)+(z1>0)+C

- 「鍵」の写像の何らかの量と、単位標本サイズCの比
- 「論理和」「論理否定」を使用してはいけない
- 異なる量や定数の和や差であってはいけない

標本サイズ関数() = 0

「標本サイズ関数」は、一意な値を返すものなら何でも良いわけではありません。
「同じ条件」を満たす「単位標本」の個数を返さなければいけません。
引数としては、例えば、x1,x2,y1,y2,z1,z2を、標本を特定する「鍵」として渡します。
標本サイズは、例えば、「(x2-x1)(y2-y1)(z1>0)(z2<0)/C」と定義できます。
「鍵」の写像の何らかの量と、単位標本サイズCの比が、戻り値になります。
不等号を使った条件式は、満たせば1、さもなくば0だとします。
「同じ条件」であるために、論理和と論理否定の使用は許されません。
また、「(x2-x1)+(y2-y1)+(z1>0)+C」のように、異なる量や定数の和や差であってもいけません。
必然的に、引数が空なら、戻り値は0になります。

相対標本サイズ



帰納推論に効くのは相対標本サイズ

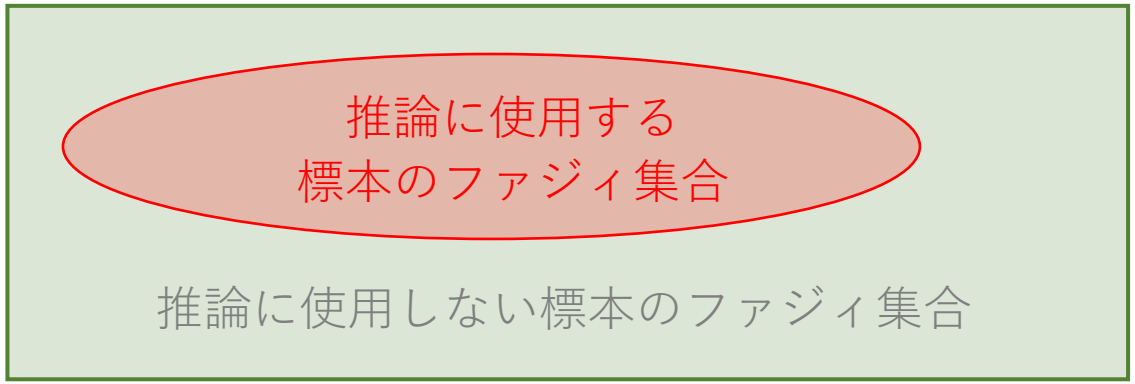
$$\frac{n}{n_0} = 9$$

標本サイズは、「単位標本」の個数です。
推測対象は「単位標本」であるとは限らず、複数の「単位標本」をまとめて推測する場合があります。
例えば、10cmの帯の、9cm部分の色から、残りの1cm部分の色を推測するとします。
1cmの帯を単位標本とすれば、推測対象は1個です。
1mmの帯を単位標本とすれば、推測対象は10個です。
これは、解像度が変わっただけで、実際に推測する内容は変わりません。
推測対象の「標本サイズ」を1とした比率を「相対標本サイズ」とします。
帰納推論に効くのは絶対値ではなく、「相対標本サイズ」です。

推論に使用する標本

標本サイズ関数で定義されるすべての標本

（すべての標本を推論に使用する必要はない）



メンバシップ関数: 推論に使用する度合いの条件式

推論に使用する集合は、連続した単一の範囲

＝条件式に「論理和」「論理否定」を使用してはいけない

標本サイズ関数で定義される標本は、標本として認められる必須条件です。
すべての標本を使用する必要はなく、推測対象と似た標本だけを使用した方が、より良い推論ができます。
ある条件に従って、全ての標本を、使用する標本と、使用しない標本のファジィ集合に分けます。
メンバシップ関数が、推論に使用する度合いを表す条件式です。
推論に使用する集合は、連続した単一の範囲内の全ての標本でなければなりません。
言い換えると、条件式に「論理和」と「論理否定」の使用は許されません。

連続した単一の範囲

(1≦x≦2)AND(3≦y≦4): 連続した単一の範囲
「論理和」「論理否定」を含まない
推論に使える

(1≦x≦2)OR(3≦x≦4): 連続した単一の範囲ではない
「論理和」「論理否定」を含む
推論に使えない

Should be divided into three groups

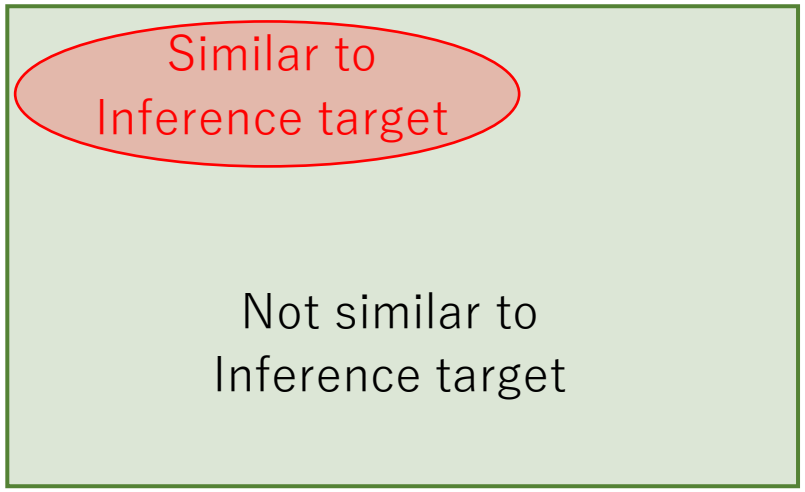
(1≦x≦2)
(3≦x≦4)
: 連続した単一の範囲

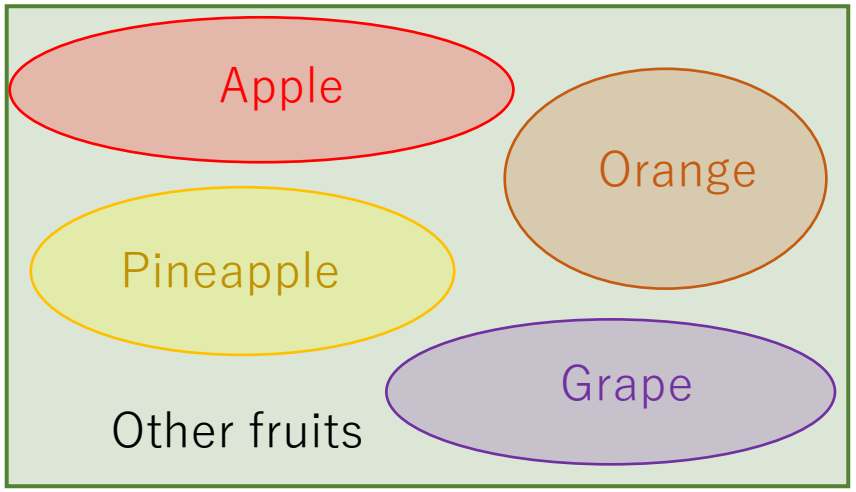
(補集合) : 連続した単一の範囲ではない

連続した単一のな範囲と条件式の関係を説明します。
例えば、「(1≦x≦2)AND(3≦y≦4)」は連続した単一の範囲です。
一方、「(1≦x≦2)OR(3≦x≦4)」のように、論理和が含まれる場合、単一な範囲とはいえません。
その場合は、「(1≦x≦2)」と「(3≦x≦4)」と「その他」の3つの集合に分けるべきです。
「その他」は補集合で、論理和と論理否定を使用せずに表現できません。
ですが、推論に使用しないのなら、単一の範囲である必要はありません。
「(1≦x≦2)」と「(3≦x≦4)」のどちらかは、推論に使用することができます。

クラスタリング

2つの集合に分ければ十分



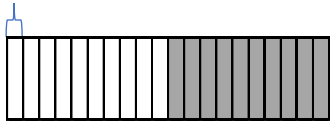


あらかじめ3つ以上の集合にクラスタリングしておけば、
素早く推論ができる

標本の集合は、推論に使うものと使わないものの2つに分かれていれば十分です。
ですが、3つ以上の集合に分ける利点もあります。
例えば、ある果物の糖度を推測したいとします。
標本は、外観的特徴が一致するものとしないものの2集合に分ければ十分です。
ですが、あらかじめ3つ以上の集合にクラスタリングしておくこともできます。
果物の種類別にクラスタリングできます。
推測したい果物が示されてから、各クラスタの代表と、外観的特徴の一致を比較します。
一致するクラスタのみを採用すれば、すばやく推論ができます。

解像度

単位標本:
1mm



単位標本:
10mm

クラスタリング ＝解像度を減らす

「特定のリンゴ」 ⊂ 「リンゴ」 ⊂ 「果物」

初めは粗く分類し、
徐々に細かく分類していくのが、効率的

クラスタリングと単位標本の解像度の関係を説明します。
ある帯の色を推測する例では、単位標本が1cmの場合と1mmの場合を考えました。
単位標本が1cmの場合は、1mmの場合の標本が10個ずつクラスタリングされていると解釈できます。
一方で果物のクラスタリングの例では、「特定のリンゴ」⊂「リンゴ」⊂「果物」という関係があります。
これは、大きくクラスタリングするほど、解像度が粗くなっていると解釈できます。
つまり、クラスタリングと解像度は同じです。
解像度は細かくするほど良い推論ができますが、計算量が増えます。
最終的に最適解を出そうというのなら、解像度をできるだけ高める必要があります。
ですが、初めは粗く分類し、徐々に細かく分類していくのが、効率的です。

論理和の禁止 (2)

仮説:「(ボタンAを押す) OR (ボタンBを押す) → ランプC点灯」
観測:「ボタンAを押した → ランプC点灯した」
「ボタンAを押した → ランプC点灯した」
「ボタンAを押した → ランプC点灯した」
…
繰り返しの観測で、ほぼ確実に仮説が成り立つと検証された
(ボタンBを押してないけどいいの?)

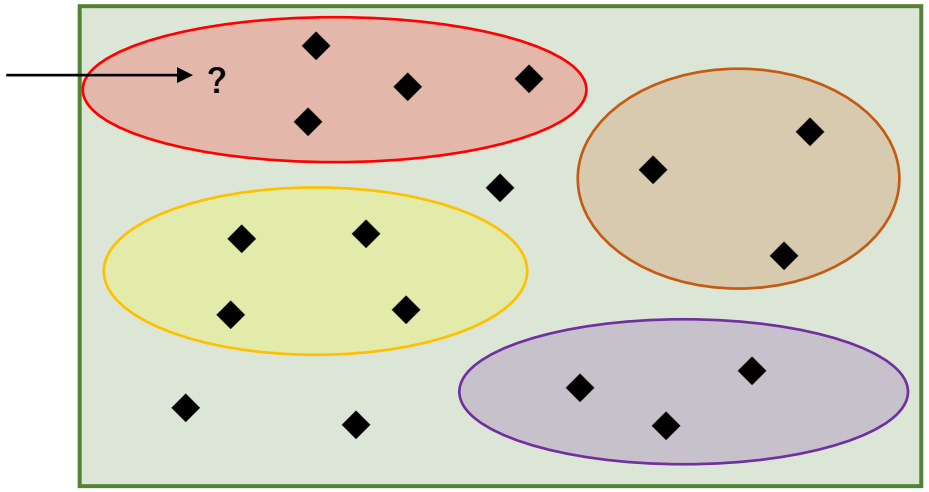
仮説:「ボタンAを押す → ランプC点灯」　…検証された
仮説:「ボタンBを押す → ランプC点灯」　…検証されていない

複数の条件を論理和にしてはならず、
個別に検証しなければならない。

論理和を条件とした集合を帰納推論に使用してはいけない理由を、例題から説明します。
ある機械の「(ボタンAを押す) OR (ボタンBを押す) → ランプC点灯」という仮説を検証します。
繰り返しボタンAを押すと、必ずランプCが点灯し、仮説がほぼ確実に成り立つことが確認できました。
しかし、ボタンBを一度も押さずに、ボタンBを押してもランプCが点灯するといえてしまうのはおかしいです。
仮説を二つに分け、「ボタンAを押す → ランプC点灯」という仮説のみの確証性が増したと考えるべきです。
複数の条件を論理和にしてはならず、個別に検証しなければなりません。

推論に使う集合

推測対象の標本も、他の標本と
同じルールで帰属する集合を決める



推測対象が帰属する集合が、
推論に使う集合

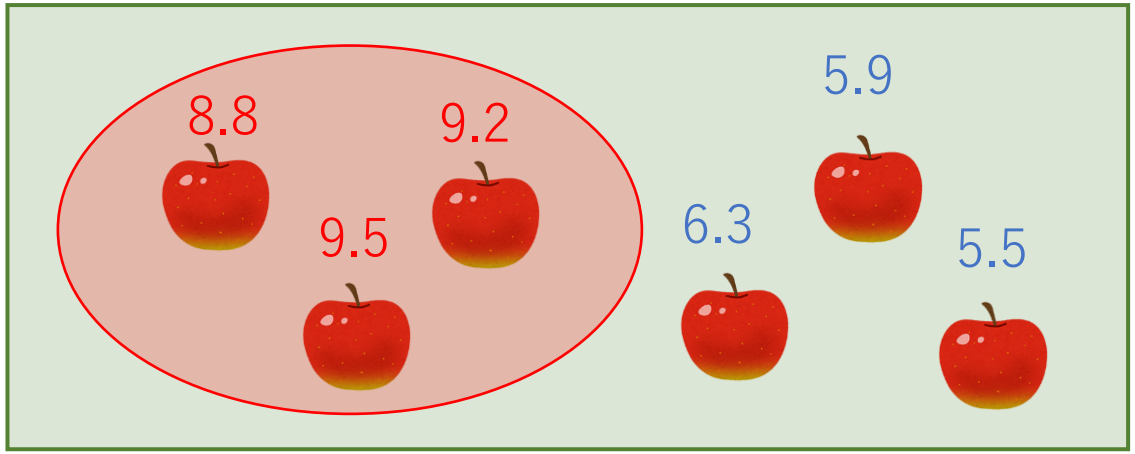
0.5ずつ、ファジィに帰属してもよい

全ての標本を任意の数の集合に分類できますが、推論に使う集合はどうやって決めればよいでしょうか？
全ての標本は、同じルールで、所属する集合を決めます。
推測対象の標本もまた、同じルールで、所属する集合を決めます。
ちょうど境界線上の場合などに、二つにファジィ集合に、0.5ずつ帰属することも許されます。
推測対象の帰属する集合が、推論に使う集合になります。
すべての集合に均等に帰属する場合は、すべての標本を推論に使うのと同じです。

恣意的なクラスタリング

糖度: ?

糖度でクラスタリング

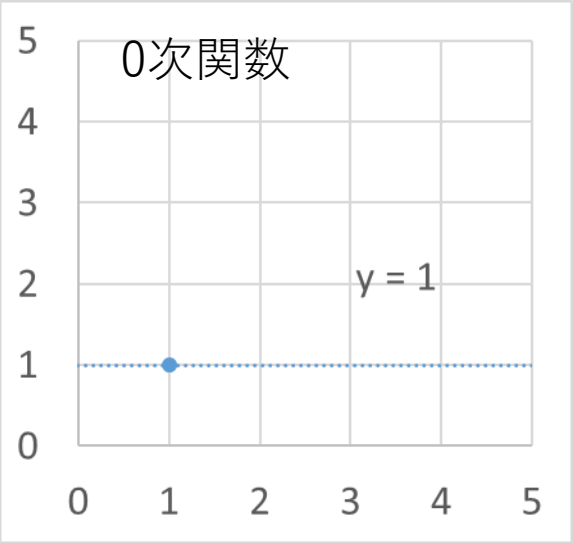


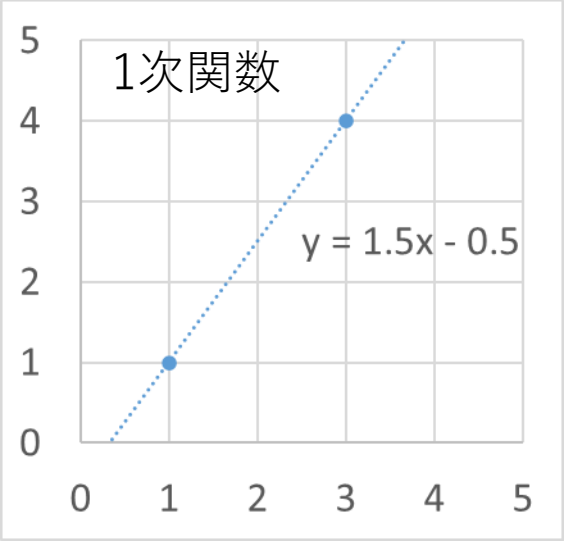
推測対象がどちらに帰属するのか分からない

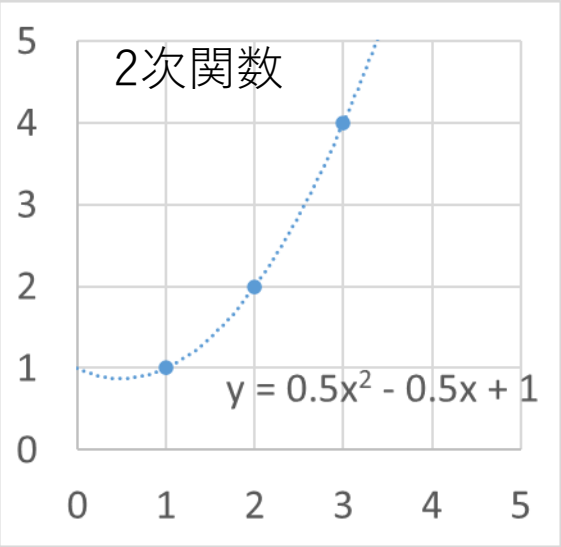
未知の値を基準にしてクラスタリングすることはできない

恣意的な標本の選択が自動的に禁止される例を示します。
例として、あるリンゴの糖度を推測する場合を考えます。
恣意的に、糖度の推測値を大きくする方法を考えます。
そこで、糖度が大きい標本と、糖度が小さい標本にクラスタリングしました。
しかし、糖度のみを推論に使うことはできませんでした。
推測対象のリンゴの糖度は「不明」なため、どちらに帰属するか分からないからです。
未知の値を基準にして、クラスタリングすることはできません。

自由度







N次関数

変数の数 ＝ N+1

$$残差平均 = \frac{\sum|Redisual|}{標本サイズ - (N+1)}$$

ここで、自由度の話をしておきます。
標準偏差の計算では、標本サイズから自由度1を引いて計算します。
回帰分析では、回帰式の変数の数が自由度になります。
例えば、一次式は、傾きと切片の2つの変数があります。
N次式なら、変数はN+1個あります。
プロットが2点なら、残差が0になるように、直線を引くことができます。
なので、残差の平均を計算するときは、標本サイズからN+1を引いて計算します。
ここまでは従来の統計の話です。

標本選択の自由

標本選択の条件: “0.0<x<6.0”

標本選択の変数: “0.0”, “6.0”

y： 目的変数	?	1.0	0.8	1.2	1.1	0.9	2.5	0.5
x： 説明変数	0	1	2	3	4	5	6	7

境界付近の1標本を含めるかどうか、
残差平均が小さくなるように、恣意的に決められる。

残差平均が全く減らない場合もある。

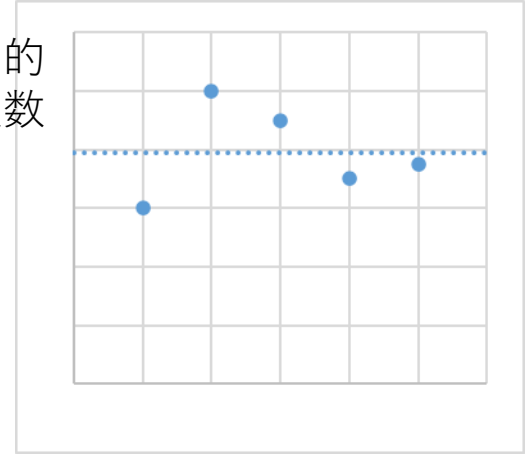
標本選択の自由（推定） $= 0.5 \times$ 標本選択の標本数

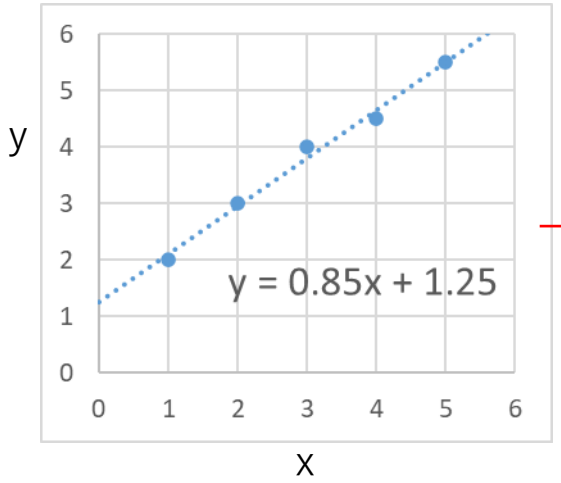
$$\text{残差平均} = \frac{\sum|\text{残差}|}{\text{標本サイズ} - (N+1) - \text{標本選択の自由}}$$

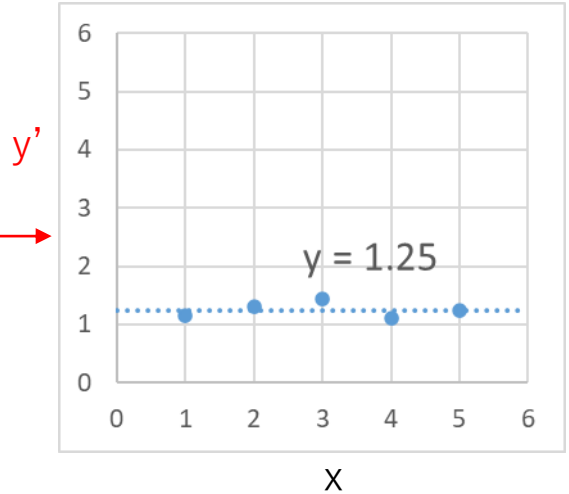
統計2.0では、標本を選択するための変数があるため、その自由度も考慮する必要があります。
例えば、「0.0<x<6.0」という条件の標本のみを使う場合、「0.0」「6.0」の2つの変数があります。
境界を自由に決めれられるので、残差平均が小さくなるように決めることができます。
つまり、変数1個につ付き、境界付近の標本1個を含めるか含めないか、恣意的に選択できます。
ただし、必ずしも、境界の決め方で、標本1個分の残差平均が減らせるとは限りません。
境界を変えても、全く残差平均が減らない場合もあります。
平均的には、変数一個で、自由度0.5と推定できます。
残差平均の計算では、標本選択の自由度を考慮します。

座標変換

目的
変数







目的変数が定数ではなく、何らかの関数だと推測したい場合、
目的変数が定数になるように座標変換すればよい

$$y' = y - ax \qquad = b\ (\text{定数})$$

他の推論と比較する場合、残差は元の座標に戻す。
また、自由度として、座標変換に使用した変数の数を考慮する。

ここまでは、説明変数が近ければ、目的変数も一定値に近いだろうと帰納推論していました。
目的変数が説明変数の関数の場合は、どうやって推論すればよいか考えます。
例えば、説明変数xと、目的変数yが、直線「y=ax +b」に近い関係にあったとします。
そこで、任意の傾きaで、「y'=y-ax」のように座標を変換します。
「y'=切片b」なので、一定値になります。
y'を目的変数とすれば、一定だろうという帰納推論ができます。
どんな関数であっても、目的値が一定値になるように座標変換すれば、推論ができます。
他の推論結果と残差を比較したいなら、元の座標に戻す必要があります。
また、自由度として、座標変換に使用した変数の数を考慮する必要があります。

直感的に良い推論

ここまでは、有効な「標本選択策」について述べました。
ここからは、「標本選択策」の優劣について考えます。

評価の案

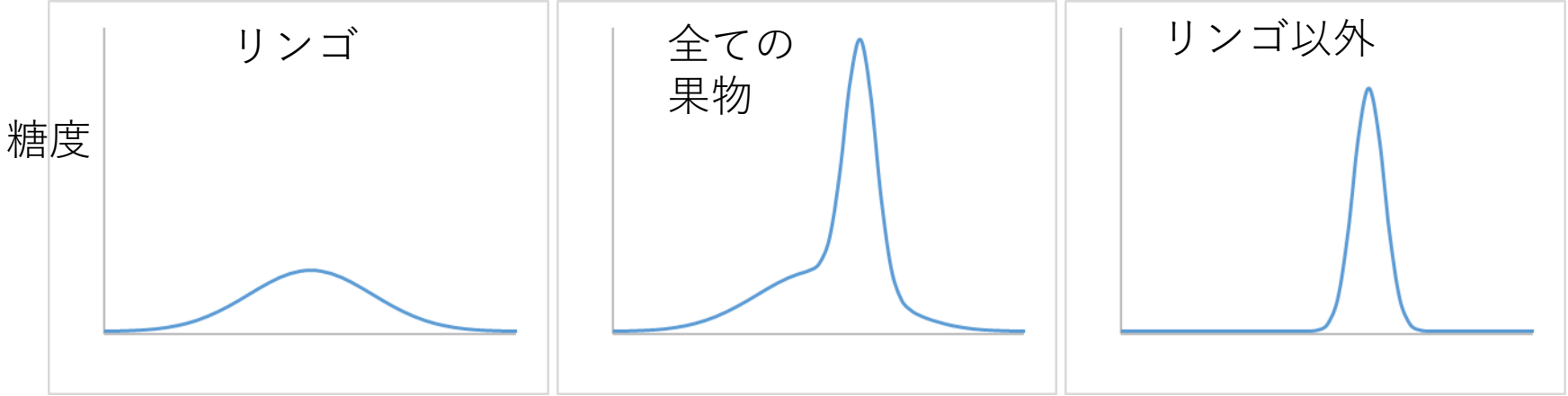
選択された標本の目的変数の分散が小さければ、良い推論?

しかし、分散で推論の優劣を評価するのは誤り。

ここまでは、有効な「標本選択策」について述べてきました。
ここからは、「標本選択策」の優劣について考えていきます。
直感的には、選択された標本の目的変数の分散が小さければ、良い推論だと感じます。
もし、100%推測できているのなら、分散は0です。
分散が小さいシャープな分布なら、選択肢が絞り込めていると解釈できます。
残っている選択肢の数が近いほど、たった一つの最適解に近づいていると感じます。
しかし、分散で推論の優劣を評価するのは誤りです。

推論の優劣と分散

質問: の糖度は?



	リンゴ	全ての果物	リンゴ以外
分散:	大	小	小
バイアス:	大	大	大

分散がいくら大きくなろうとも、リンゴのみの方がバイアスが小さいため、
良い推論だと感じる。

あるリンゴの糖度を推測する例を考えます。
全ての果物の糖度の分布ではなく、リンゴのみの糖度の分布を見た方が良い推論です。
ここで、リンゴ以外の果物の糖度の分散が非常に小さかったとします。
すると、果物全体の糖度の分散も小さくなります。
リンゴのみに絞り込むと、かえって分散が大きくなってしまいます。
分散が大きくても、リンゴのみを見た方が良いと思うのはなぜでしょうか？
推論の良さは、精度(分散)、正確度(バイアス)に分けられます。
つまり、分散がいくら大きくなろうとも、リンゴのみの方がバイアスが小さいため、良い推論だと感じたのです。

最良な帰納の極限 (1)

理論的に最も良い帰納推論の極限とは？

使用する標本の、質と量の両方がベストな状態

質問: あるリンゴの糖度は?

最も品質の良い標本: 推測対象のリンゴと区別できないほど似たリンゴ

最も品質の良い標本が無限個ある状態がベスト?

ただし、標本を数えられるということは、何らかの差で区別することはできる。

ここで、理論的に最も良い帰納推論の極限を考えてみます。
演繹ではないので、答えをそのもの持っている標本はありません。
使用する標本の、質と量の両方がベストな状態を考えます。
直感的には、最も品質の良い標本が、無限個あればベストだと思われます。
あるリンゴの糖度を推測する例を考えます。
推測対象のリンゴと区別できないほど似たリンゴなら、最も品質の良い標本といえます。
ただし、標本を数えられるということは、何らかの差で区別することはできます。

最良な帰納の極限 (2)

一位タイの標本が大量にあり状態が、本当に最適?

標本
品質
ランキング: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 … ∞

差があるのに気が付いていないだけ。

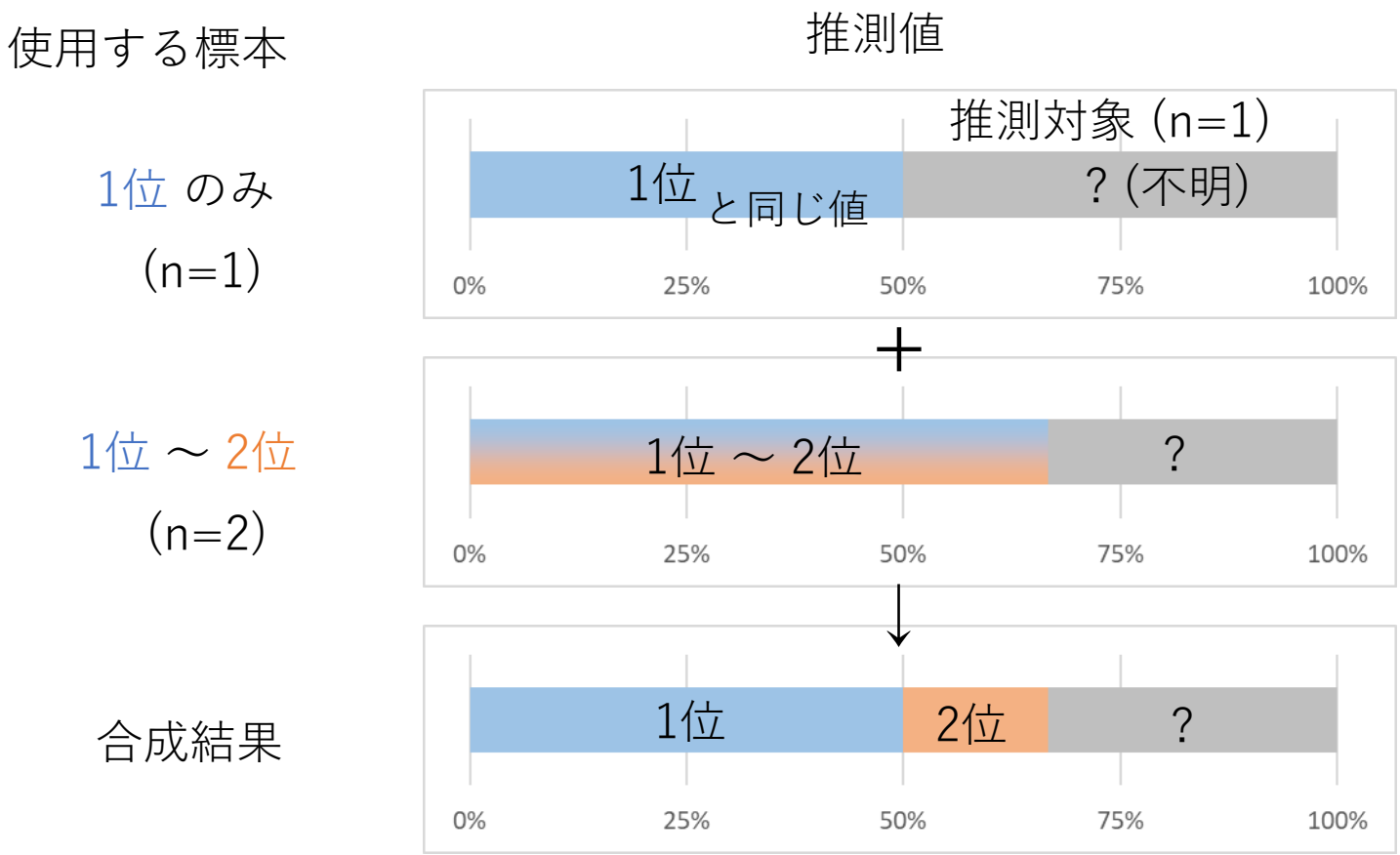
標本
品質
ランキング: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 … ∞

質と量がベストな標本

n位の標本は、n+1位の標本よりも、
明らかに推測対象の標本に類似している状態

しかし、類似度が一位タイの標本が大量にある状態が、本当に最適なのでしょうか？
気が付いていないだけで、一位タイの中でも、何らかの差があるかもしれません。
何らかの差があるなら、順位に差が付けられるはずです。
そのため、同じ順位の標本がない状態が、最適だといえます。
n個の標本があれば、1位から、n位まで、順位が付けられた状態です。
n位の標本は、n+1位の標本よりも、明らかに推測対象の標本に類似している状態です。
このような標本が無限個ある状態が、標本の質と量がベストな状態です。

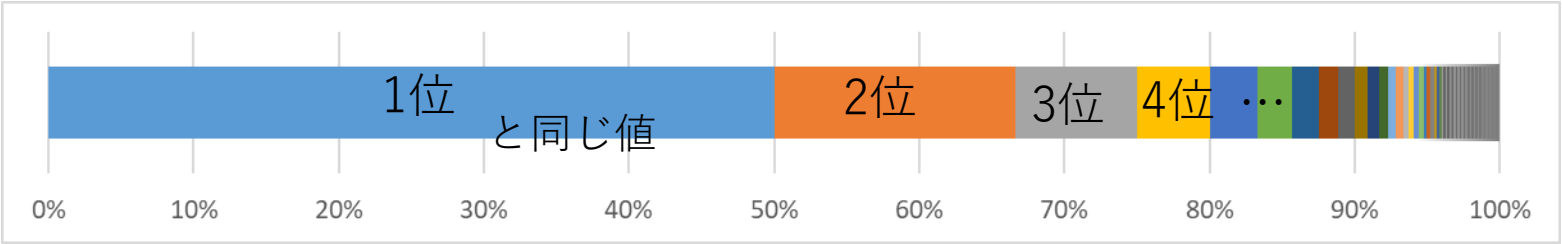
最良な帰納の極限 (3)



標本の質と量がベストな状態では、推測値がどうなるか考えてみます。
1位の標本1つのみで推論する場合を考えてみます。
推測対象の標本を1個加えるため、50%は1位の標本と同じ値、残りの50%は「不明」となります。
2位以上の標本2つで推測する場合も考えてみます。
67%は2位以上の標本の2つと同じ値、残りの33%は「不明」となります。
ただし、1位の標本は、2位の標本より、明らかに、品質の良い標本です。
そのため、67%の内、50%は1位の標本と同じ値、残りの17%は2位の標本と同じ値になります。

理想帰納

推測値



$$\text{n位の標本の比率} = \frac{1}{n^2+n}$$

$$\text{1位からn位までの合計} = \frac{n}{n+1}$$

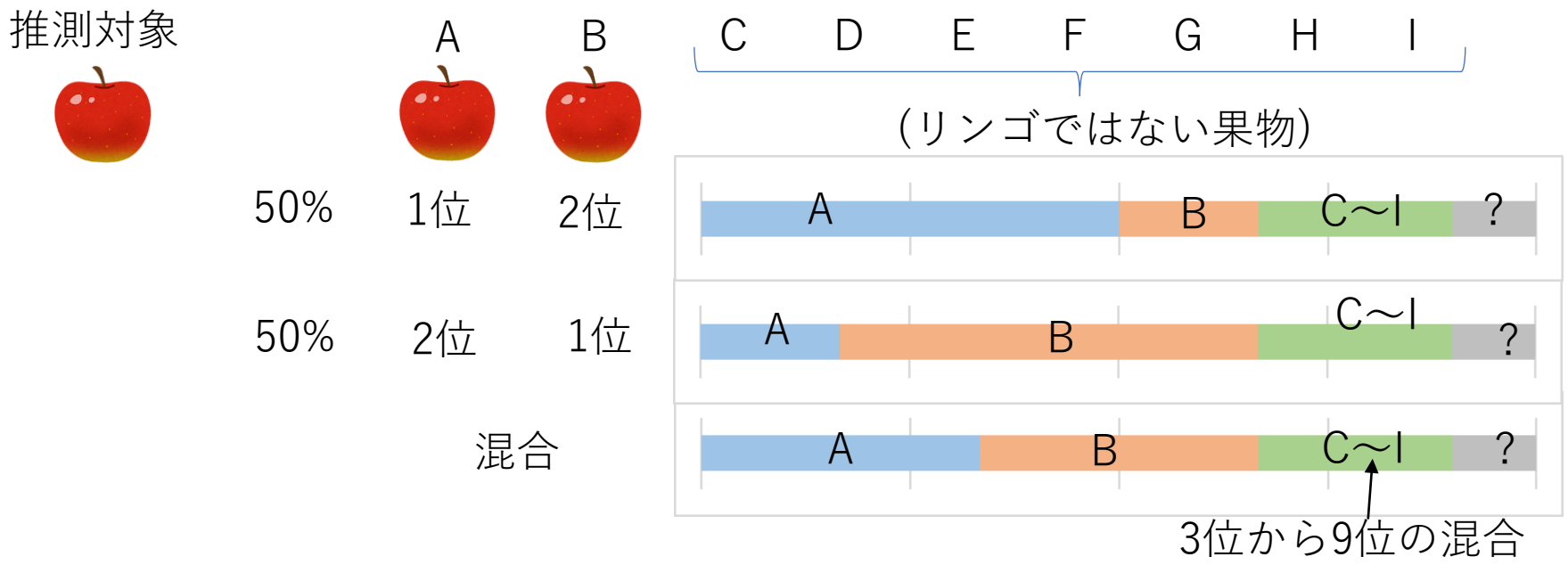
$$\text{1から上式を引いた残り} = 1-\frac{n}{n+1}=\frac{1}{n+1} \quad :?\ (\text{不明})$$

標本数が増すほど、「不明」が減る

このような状態を「理想帰納」と呼ぶ。

同様に、n位まで計算したものを合成します。
推測値の1/(n^2+n)が、n位の標本と値になります。
1位からn位までの合計は、n/(n+1)になります。
1からそれを引いた残りの、1/(n+1)が「不明」になります。
上位から順番に標本数を増やすと、「不明」だった部分が、「不明」でなくなっていきます。
一度「不明」ではなくなった部分は変化しません。
「不明」は最悪の値です。
どんな標本でも「不明」が減るため、ないよりましなので、無視するべきではないといえます。
このような状態を「理想帰納」と呼びます。

順位の混合



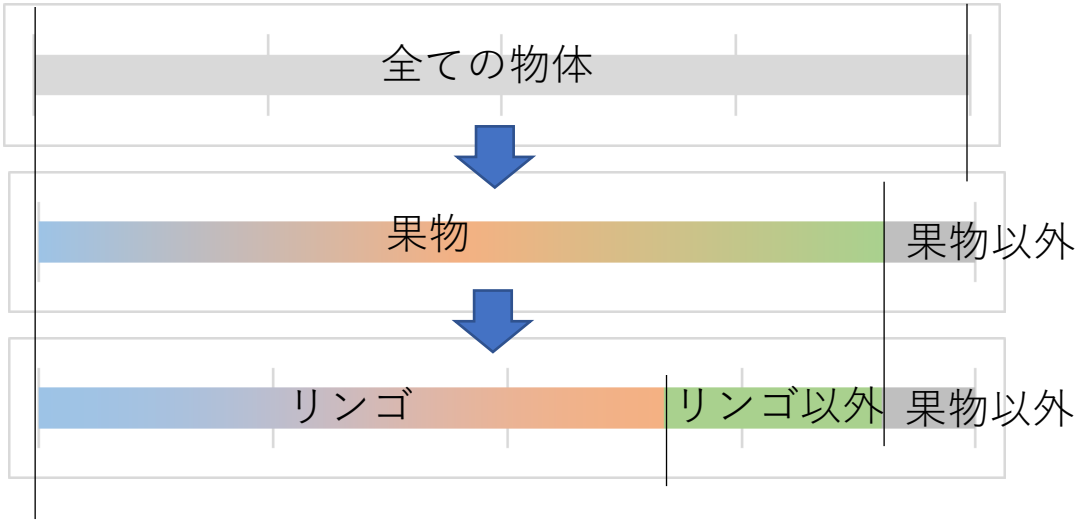
理想帰納というのは、全ての順位の組み合わせの数だけ存在。

すべての推論結果は、理想帰納の混合で表せられる。

ベストではない標本の例を考えます。
あるリンゴの糖度を推測するとします。
標本として9個の果物があり、その内2個あるリンゴを選択しました。
2個のリンゴAとBの品質の順位は、1位タイではありません。
「Aが1位で、Bが2位」と「Aが2位で、Bが1位」が、50%ずつ混合していると解釈します。
リンゴ以外の7つの果物については、3位から9位の順位の組み合わせの混合と考えられます。
理想帰納というのは、全ての順位の組み合わせの数だけ存在します。
すべての推論結果は、最適な推論結果の混合で表せられます。

段階的な選択

推論対象



使用する標本の選択は、段階的に行うことができる。

使用する標本を決めるというのは、順位を決めるのと同じ。

「果物以外」にも低い順位がついている。
すべての標本は、ほんの僅かでも、必ず推論に使用される。

使用する標本の選択は、段階的に行うことができます。
例えば、あるリンゴの糖度を推測したいとします。
初めに、「全ての物体」を「果物」と「果物以外」に分けます。
次に、「果物」を「リンゴ」と「リンゴ以外」に分けます。
全ての標本が順位が均等だったものが、徐々に明確になっていきます。
使用する標本を決めるというのは、順位を決めるのと同じことです。
「果物以外」にも低い順位がついています。
そのため、すべての標本は、ほんの僅かでも、必ず推論に使用されます。

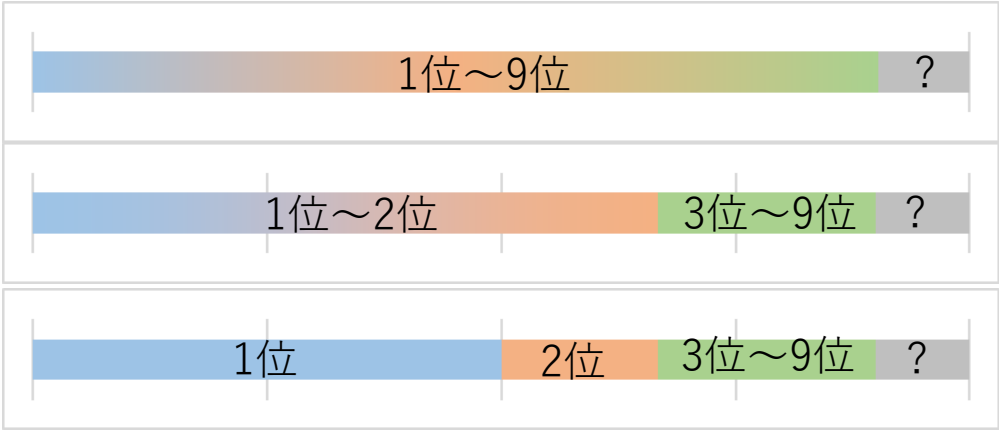
順位の組み合わせ

順位の組み合わせ

9!=362.880

2!×7!= 10,080

1×1×7!= 5,040



評価方法の案

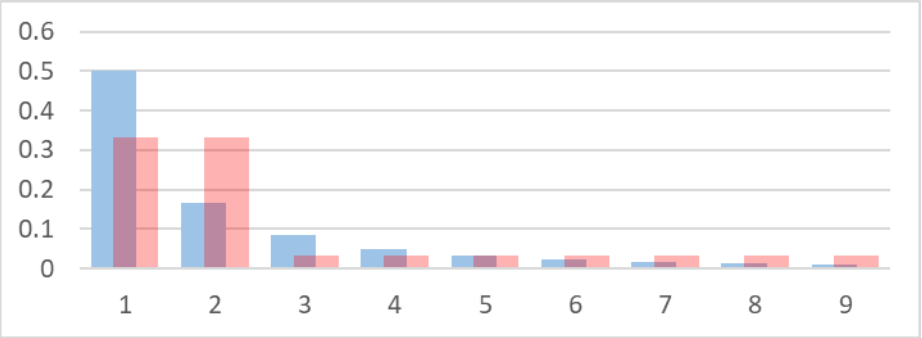
順位の組み合わせの数が小さいほど、良い推論?

1位と2位の順番は、8位と9位の順番よりも重要なはずなので、
等価値とみなされてしまうのは誤り

順位の組み合わせの数を考えてみます。
部分的にでも順位について分かれば、組み合わせの数が減少します。
そこで、順位の組み合わせの数を、推論の優劣の評価方法の案として考えてみます。
ですが、1位と2位の順番は、8位と9位の順番よりも重要なはずです。
それが等価値とみなされてしまうので、この評価方法は誤りです。

理想率 (1)

各標本が推測結果に
占める割合



大きい順

：理想帰納
：ある推論
：一致部分

理想率: 一致部分の合計　　(0～1)

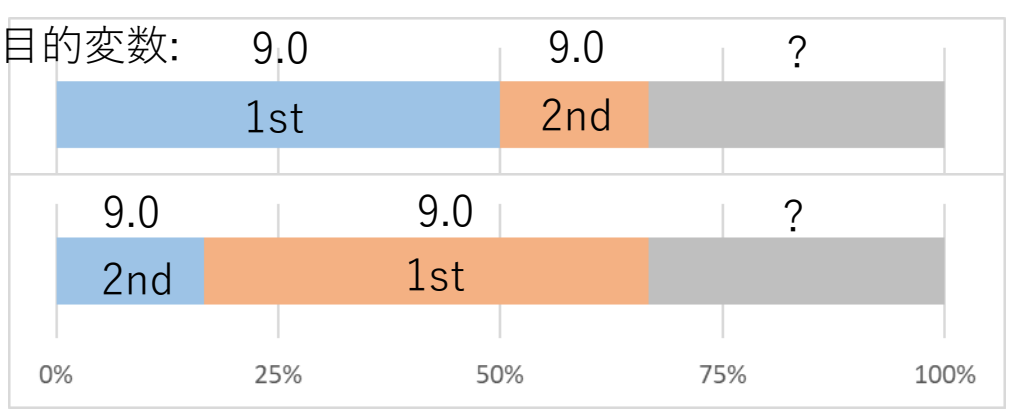
(「不明」部分は一致しない)

推論の優劣評価の別の案を考えます。
「理想帰納」と一致していれば、それは最も良い推論といえます。
そこで、「理想帰納」との一致の度合いを数値化してみます。
推論結果における各標本のが占める割合が大きい順に並べます。
その割合が、理想推論とある推論で一致する部分を、標本ごとに調べます。
すべての標本についての合計を「理想率」とします。
「不明」部分は一致するかしないか分からないので、一致しないとします。

理想率 (2)

評価方法の案

「理想率」が良いほど、良い推論?



目的変数の分布は
どちらも同じ

順位を決める
必要ない

全ての標本の目的変数が定数になる場合、
「理想率」は低いが、その定数になるという優れた推論ができる。
「理想率」で推論の優劣を評価するのは誤り。

標本の順位だけでなく、目的変数の値のバラツキも考慮する必要がある

「理想率」の大きさを、推論の優劣の評価方法の案として、考えてみます。
1位と2位の順番がはっきりしていると、そうでないときより「理想率」は大きくなります。
ですが、1位と2位の標本が、全く同じ目的変数だったとします。
その場合、1位と2位の順番をはっきりさせても、推測結果は変わりません。
極論として、座標変換した結果、全ての標本の目的変数が定数になったとします。
その場合、どの標本を使うか考える必要はなく、推測値は、その定数になります。
すべての標本を均等に使用すると、「理想率」は最低になります。
「理想率」が低くても優れた推論ができるので、「理想率」で推論の優劣を評価するのは誤りです。
標本の順位だけでなく、目的変数の値のバラツキも考慮する必要があると分かました。

バラツキの数値化

目的変数に含まれる「不明」や「外れ値」は、無視できない

平均 $= \bar{x}$

僅かでも「不明」が含まれていると
決まらない

$$\text{標準偏差} = \frac{1}{n-1}\sqrt{\sum (x-\bar{x})^2}$$

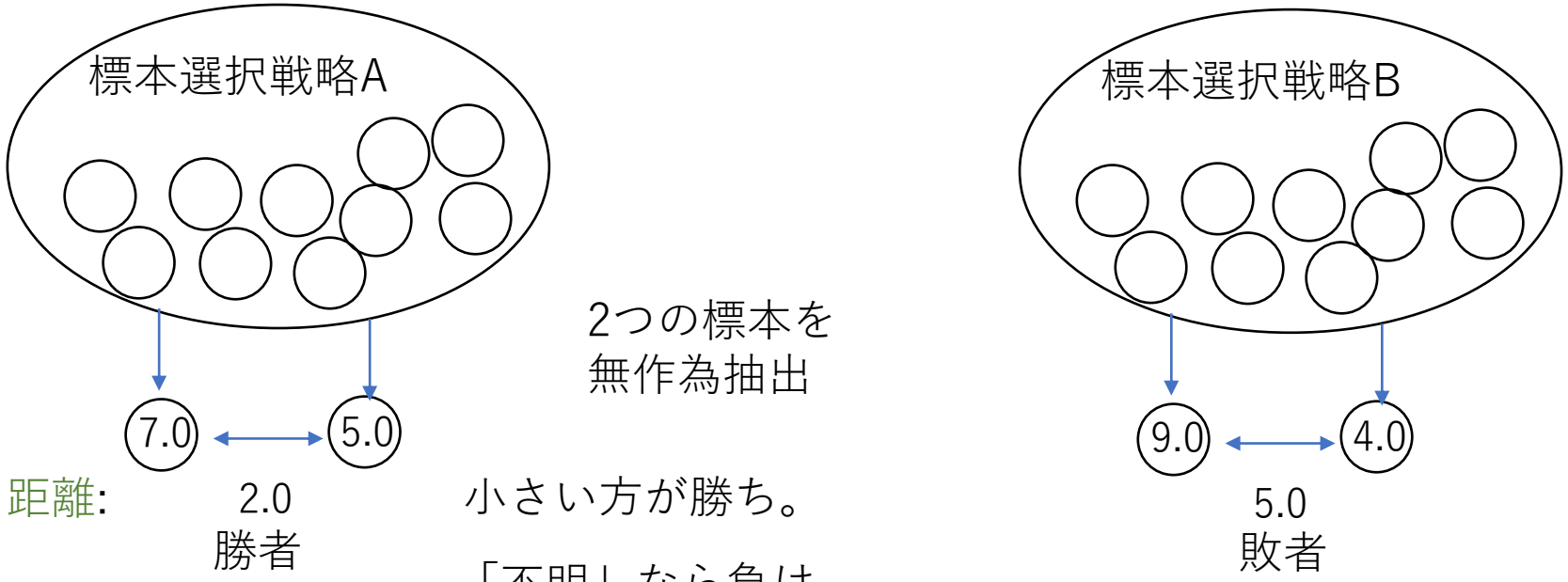
外れ値が支配的

$$\text{平均絶対偏差} = \frac{1}{n-1}\sum |x-\bar{x}|^1$$

値が桁違いなら、
外れ値が支配的

目的変数のバラツキの数値化は困難です。
値が「不明」な標本が含まれるからからです。
また、「不明」以外にも、極端な外れ値を持つ標本も存在します。
外れ値の標本であっても、無いよりはましなので、無視することは許されません。
標準偏差の計算では、残差を二乗するため、外れ値の影響が大きくなります。
平均絶対偏差なら、距離の期待値なので、外れ値の影響度合いも妥当です。
それでも、桁違いの外れ値があれば、それが支配的になってしまいます。
また、僅かでも「不明」が含まれていれば、平均値が決められません。

距離の勝率



2つの標本を
無作為抽出

距離:

勝者

敗者

小さい方が勝ち。

「不明」なら負け。
「不明」同士なら引き分け。

繰り返して勝率を決める。
ランダム性を排除するなら、すべての組み合わせを試す。

「不明」や「外れ値」が含まれていても、目的変数のバラツキを評価できる方法を考えます。
2つの「標本選択戦略」の優劣を知りたいだけなら、バラツキの優劣だけ分かれば十分です。
2つの戦略の中から、標本を2つずつ、無作為抽出してみます。
それぞれで2つの標本の目的変数の距離を求めます。
2つの戦略で、どちらの距離が大きいか比較します。
小さい方を勝利として、何度も繰り返して勝率を決めます。
値が「不明」なら負けで、「不明」同士なら引き分けとします。
ランダム性を排除したいなら、すべての組み合わせを調べます。

非理想距離の勝率

2つの標本を
無作為抽出

距離

×(1-「理想率」)

小さい方が勝ち。

「不明」なら負け。
「不明」同士なら引き分け。

繰り返して勝率を決める。
ランダム性を排除するなら、すべての組み合わせを試す。

「バラツキ」だけではなく、「理想率」も含めた評価方法を考えます。
2つの「標本選択戦略」から、2つずつ標本を抽出します。
それぞれで、2つの標本を距離を求め、それに、「1-理想率」を掛けます。
2つの戦略で、「1-理想率」を掛けた、どちらの距離が大きいか比較します。
あとは、前の例と同じです。
「理想率」は、全ての標本についてではなく、抽出した標本についてのみ見れば、素早く計算できます。

戦略評価法 (1)

ウェイト順位: n （全ての標本をウェイトの大きい順にソート）

標本ウェイト: W_n

理想ウェイト: $I_n = \dfrac{1}{n^2 + n}$

理想率: $R_n = \dfrac{2 \times Min(W_n, I_n)}{W_n + I_n}$

目的変数: V_n

非理想距離: $D_{ij} = \left|V_i - V_j\right| \times \left(1 - \dfrac{R_i + R_j}{2}\right)$

（i位とj位の標本を無作為抽出）

「標本選択戦略」の評価方法の手順を説明します。
事前に、全ての標本をウェイトの大きい順にソートしておきます。
順位nの標本のウェイトをWnとします。
順位nの理想帰納のウェイトInは、「1/(n^2+n)」です。
順位nの理想率Rnは、2×Min(Wn, In)/(Wn　+In)となります。
順位nの標本の目的変数の値とVnとします。
i位とj位の標本を無作為抽出します。
非理想距離Dijは、「|Vi -Vj|{1-(Ri +Rj)/2}」となります。

戦略評価法 (2)

非理想距離: $$D_{ij} = \left|V_i - V_j\right| \times \left(1 - \frac{R_i + R_j}{2}\right)$$

2つの戦略で、非理想距離を比較する。

小さい方を勝利として、何度も繰り返して勝率を決める。

値が「不明」なら負けで、
「不明」同士なら引き分け。

ランダム性を排除したいなら、すべての組み合わせを試す。

2つの戦略で、非理想距離を比較します。
小さい方を勝利として、何度も繰り返して勝率を決めます。
値が「不明」なら負けで、「不明」同士なら引き分けとします。
ランダム性を排除したいなら、すべての組み合わせを調べます。

非理想距離の解釈

非理想距離: $D_{ij} = |V_i - V_j| \times \left(1 - \frac{R_i + R_j}{2}\right)$

バリアンス　バイアス

バイアスとバリアンスのどちらかが0に近ければ、
他方が大きくても最良の推論になる。

例

座標変換した結果、
全ての標本の目的変数の残差が0になった場合

バリアンス≒0

質問者から、標本の順位が与えられていたので、
そのままその順位を採用した場合

バイアス≒0

非理想距離の二つの項は、バイアスとバリアンスに対応していると解釈できます。
バイアスとバリアンスのどちらかが0に近ければ、他方が大きくても最良の推論になります。
例えば、座標変換した結果、全ての標本の目的変数の残差が0になった場合です。
順位を付けることができないのでバイアスは最悪です。
ですが、バリアンスが0なので、バイアスがいくら悪くても、最良の推論です。
別の例で、質問者から、標本の順位が与えられていたので、そのままその順位を採用した場合です。
バリアンスがいくら大きくても、バイアスが0の理想的な帰納推論です。

ブラインド推論

バリアンス　　バイアス

非理想距離: $D_{ij} = |V_i - V_j| \times \left(1 - \frac{R_i + R_j}{2}\right)$

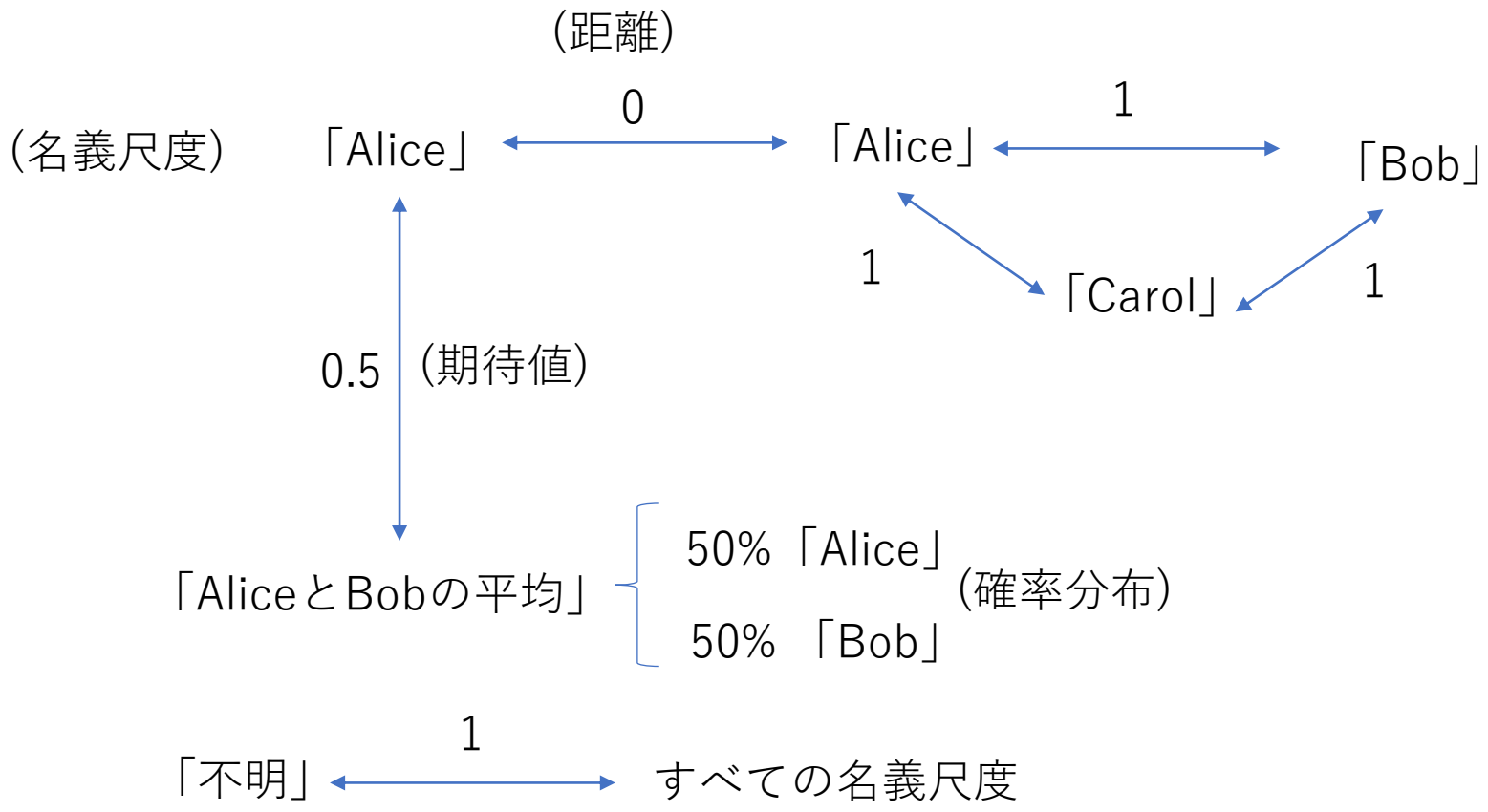
バイアスとバリアンスの片方が不明でも、
定数と仮定すれば比較できる。

バリアンスを知らなくても推論できる

目的変数の値を知らなくても、特定の標本の目的変数の値が、
推測対象の目的変数の値と近い主張できる

非理想距離の算定ができなくても、戦略の優劣を評価できる場合があります。
バイアスとバリアンスの項がありますが、片方が不明でも、定数と仮定すれば比較できます。
バリアンスを知らなくても良いということは、目的変数の値を一切知らなくても推論ができるということです。
例えば、標本の順位の情報が与えられている場合です。
「標本Aの目的変数の値が、推測対象の目的変数の値に最も近い」という情報が与えられているとします。
そのとき、その値を知らなくても、その情報通りの主張はできます。
効果的な説明変数の情報が与えられている例でも同様です。
その説明変数の値が近い標本は、目的変数も近いと主張できます。

名義尺度の距離



目的変数の距離について補足します。
目的変数は、数値ではなく、名義尺度の場合があります。
例えば、ある人名の推測値は、アリス、ボブ、キャロルなどの値になります。
名前が同じなら距離は0で、名前が異なるなら距離は1とするのが妥当です。
名前毎次元に0または1の値を持ち、ユークリッド距離を見る方法は、根拠が乏しいです。
また、「アリスとボブの平均」は、アリス50%+ボブ50%の確率分布と考えます。
「アリス」と「アリスとボブの平均」の距離は、期待値の0.5になります。
また、「不明」とすべての名義尺度の距離は、最悪値の1だと推定します。

推論の層

			与えられる情報
層	1	演繹	答えの情報
	2	帰納	標本の順位や重み
	3		効果的な説明変数
	4		目的変数の値

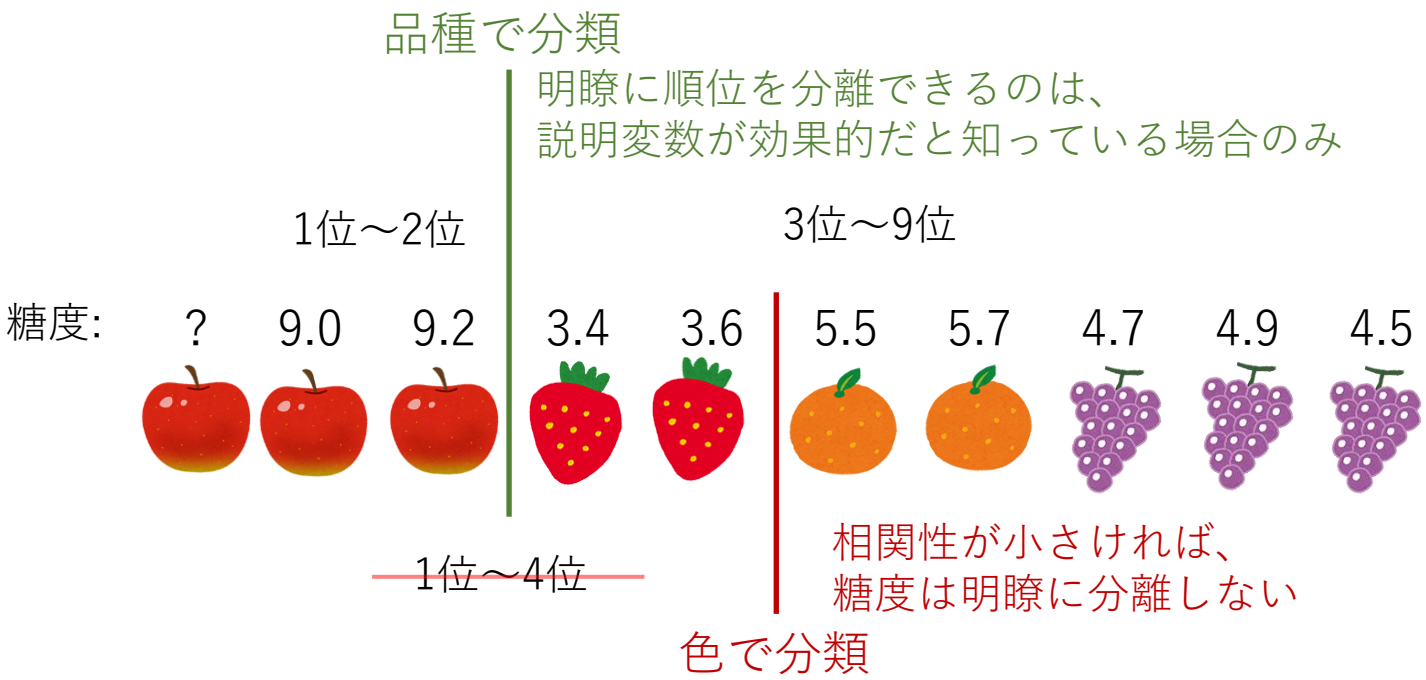
順位や重みから推定

説明変数から推定

目的変数から推定

推論は、与えられる情報によって4層に分けられます。
1層目は、演繹できる場合です。
2層目は、標本の順位や重みが与えられている場合です。
与えられた順位や重みをそのまま適用すれば、最適な推論です。
3層目は、効果的な説明変数が与えられている場合です。
その説明変数から、順位を推定します。
4層目は、効果的な変数が与えられていない場合です。
その場合、目的変数から、効果的な説明変数を推定します。

不明瞭な分離

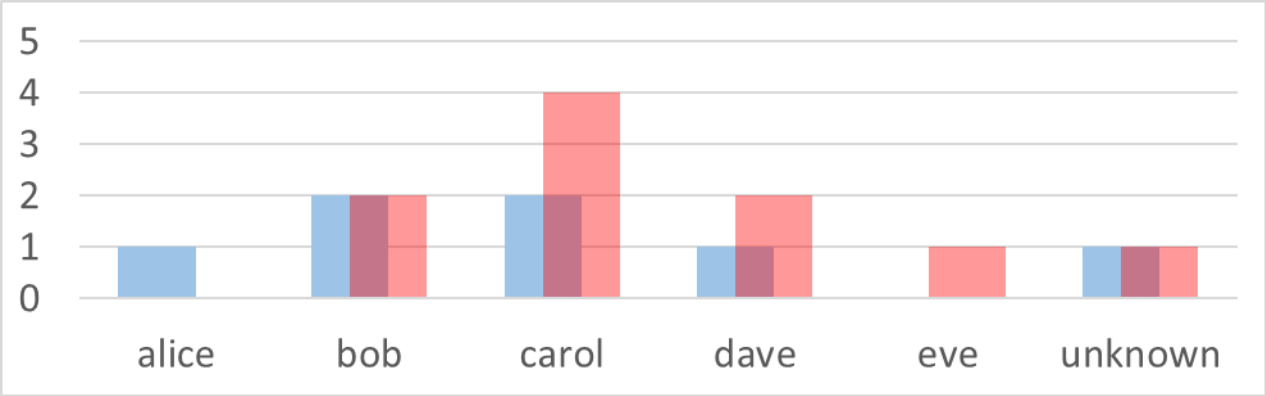


目的変数との関係から、効果的な説明変数を推定する必要がある

リンゴの例では、果物の品種が同じものが1位から2位で、異なるものは3位から9位に分けました。
2つのグループの順位は、一定値以上と、一定値以下に明瞭に分離しています。
それがいえるのは、目的変数の糖度が品種で決まると知っている場合のみです。
例えば、品種を無視して、色のみでグループを分けることもできます。
色と糖度の相関性が小さければ、糖度の順位は、明瞭に分離しません。
効果的な説明変数が情報として与えられる場合は稀です。
目的変数との関係から、効果的な説明変数を推定する必要があります。

一致率 (名義)

標本サイズの頻度



目的変数

グループ A

グループB

一致率

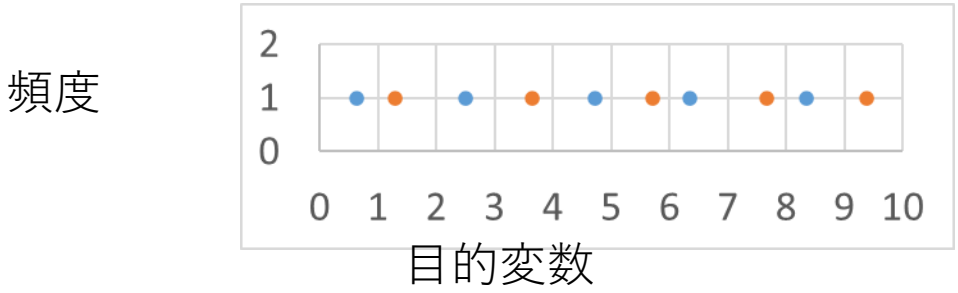
$$= \frac{\text{一致する標本サイズ}}{\text{合計標本サイズ}} \quad (0\sim1)$$

目的変数の値と頻度の分布を図示しました。
説明変数で2つのグループに分けても、目的変数は必ずしも分離しません。
2つのグループの分布が重なっている度合を、「一致率」とします。
名義尺度なら、名前別に一致する標本サイズから計算できます。
分母は標本サイズの合計で、分子は一致のある標本サイズです。
0から1の間の値をとります。

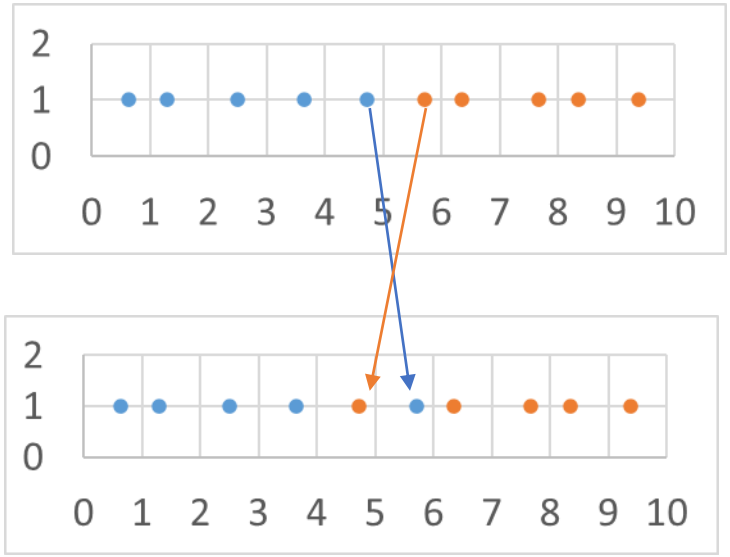
一致率 (連続)

特定の分布だと決めつけることは許されない

交互: 一致率≒1



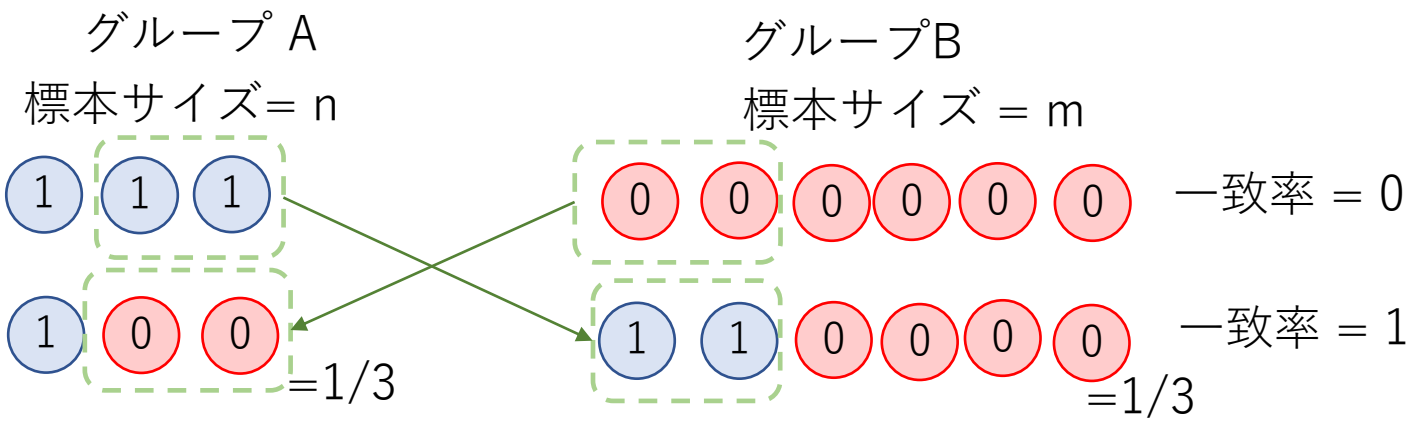
分離: 一致率≒0



もともと分離していたと考え、
入れ替わった個数を数えれば一致率が分かる

目的変数が連続値の場合を考えます。
精度が悪いと区別ができないだけで、数値が厳密に一致することはありません。
分布の形状は分からないので、正規分布だと決めつけるのは許されません。
分布の形状を仮定しないノンパラメトリック手法でも、一致率は計算できます。
直感的には、2グループの数値が交互に並んでいれば、一致率は1です。
逆に、順番が完全に分離していれば、一致率は0です。
一致率が0でない状態は、もともと分離していた標本のいくつかが、入れ替わったものと解釈できます。
そのため、入れ替わっている標本を数えれば、一致率が分かります。

交換回数



$$\text{交換回数} = \frac{1}{\frac{1}{n}+\frac{1}{m}} \qquad \text{交換標本サイズ} = \frac{2}{\frac{1}{n}+\frac{1}{m}}$$

$$\text{一致率} = \text{交換標本サイズ} \times \frac{\frac{1}{n}+\frac{1}{m}}{2} \qquad \text{交換標本サイズ} = \text{一致率} \times \frac{2}{\frac{1}{n}+\frac{1}{m}}$$

分布の一致率と、標本の交換個数の関係を求めます。
初めに、グループAには、値が「1」の標本がn個、グループBには、値が「0」の標本がm個あったとします。
2つのグループの値の分布を一致させるのには、「1/(1/n+1/m)」回、標本の交換が必要です。
1回の交換で、2個の標本が交換されます。
従って、一致率を0から1にするためには、「2/(1/n+1/m)」個の標本の交換が必要です。
ある2つの分布の一致率と、標本交換個数は、表記した式で換算可能です。

不明瞭な分離のウェイト

	グループA			グループB						
順位:	1～3	1～3	1～3	4～9	4～9	4～9	4～9	4～9	4～9	一致率 = 0
				交換						
	1～3	1～3	4～9	1～3	4～9	4～9	4～9	4～9	4～9	
ウェイト:	0.25	0.25	0.025	0.25	0.025	0.025	0.025	0.025	0.025	0.1
				均質化						
	0.175	0.175	0.175	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.1

どれが交換されたものなのか区別できないため、
均質化する

標本のウェイトが分かれば、非理想距離が計算できる。

さて、目的変数の分離が不明瞭なときに、標本の順位をどう決めるか考えます。
標本交換個数は、その数を数えるか、一致率から換算することで求められます。
例として、標本が3個と6個の、グループAとBに分かれているとします。
交換前は分離しているので、「1位から3位」と、「4位から9位」の標本に分かれています。
分布の一致率が0.5だったため、2個が交換されたと換算しました。
順位が分かれば、標本のウェイトに換算できます。
グループAの標本のウェイトは、グループ内の標本のウェイトを均質化します。
なぜなら、交換後の標本しか知らないため、どれが交換されたものなのか区別できないためです。
標本のウェイトが分かれば、非理想率が求められるので、標本選択戦略の優劣を評価できます。

統計学の目的

	記述統計学	推計統計学	統計学 2.0
目的	要約統計量 (平均値や分散など)	母集団 (正規分布など)	未知の標本の値

従来の統計学　　標本が1つしかないような場合は、
まともな推測はできない。

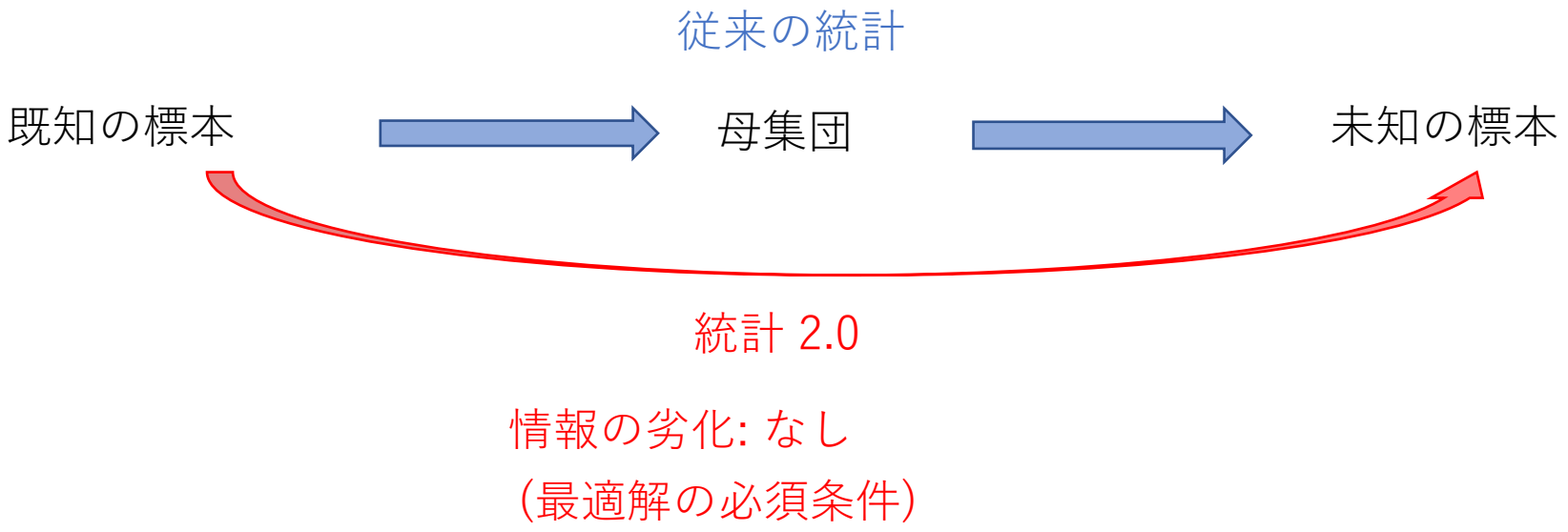
統計学 2.0　　たった一つの標本からでも、
その情報を最大限に活用して推測できる。

ここまでで、標本選択戦略の評価基準の基礎を説明できました。
次に、選択した標本を使って、「不明」な標本の推測値をどう決めるのか説明していきます。
そもそも、従来の統計学は、不明な標本の値を推測するのが目的ではありません。
正規分布などを仮定して、母集団を推定します。
また、平均値や分散などの要約統計量を計算します。
それらを応用して、「不明」な標本を推測できる場合もあるというだけです。
例えば、標本が1つしかないような場合は、まともな推測はできません。
一方で統計学2.0では、「不明」な標本の推測を目的としています。
たった一つの標本からでも、その情報を最大限に活用して推測できます。

統計における情報

追加された情報: 正規分布、事前確率など

削除された情報: 標本個別の値



統計における情報について考えます。
従来の統計学では、母集団を推定するときに正規分布などを仮定します。
すなわち、正規分布だという情報を恣意的に追加してしまいます。
ベイズ統計の場合では、事前情報を恣意的に追加します。
平均値や分散の情報を残して、標本個別の値の情報は失われてしまいます。
一方で、統計2.0では、既知の標本から未知の標本を直接推論します。
そのため、情報が劣化しません。
情報が劣化しないことは、最適解の必須条件です。

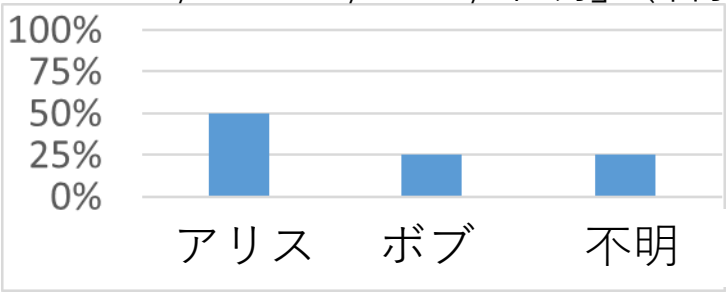
生の目的変数

目的変数: 「アリス, アリス, ボブ, 不明」(名義尺度)

頻度

＝

推測値

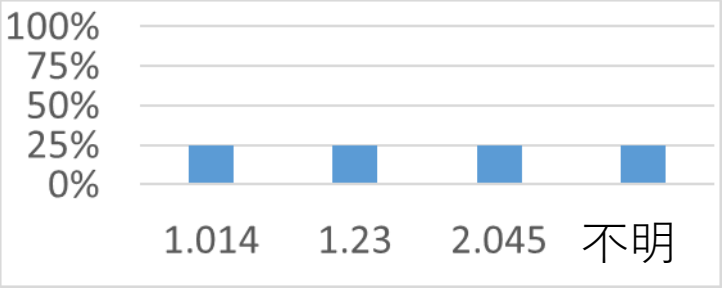


目的変数: 「1.014, 1.023, 2.045, 不明」(連続値)

頻度

≠

推測値



連続値の場合、
標本の目的変数の分布をそのまま推測値とすることはできない

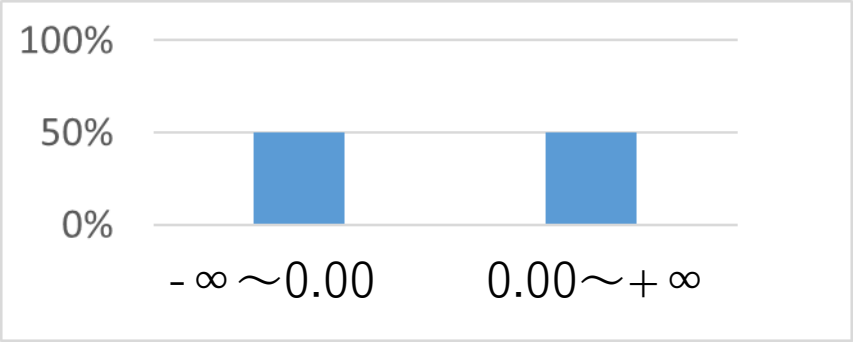
目的変数を直接、推論する方法を考えます。
例えば、標本の目的変数が、「アリス, アリス, ボブ, 不明」という名義尺度だったとします。
推測値は、そのまま、「アリス」=50%、「ボブ」=25%、「不明」=25%になります。
別の例で、「1.014, 1.023, 2.045, 不明」という連続値だったとします。
名義尺度の場合と同様だとすると、それぞれ25%になります。
ですが、連続値なので、小数点以下まで厳密に一致するのは稀なはずです。
連続値の場合、標本の目的変数の分布をそのまま推測値とすることはできません。

1標本からの推論

目的変数: 「0.0, Unknown」（連続値）

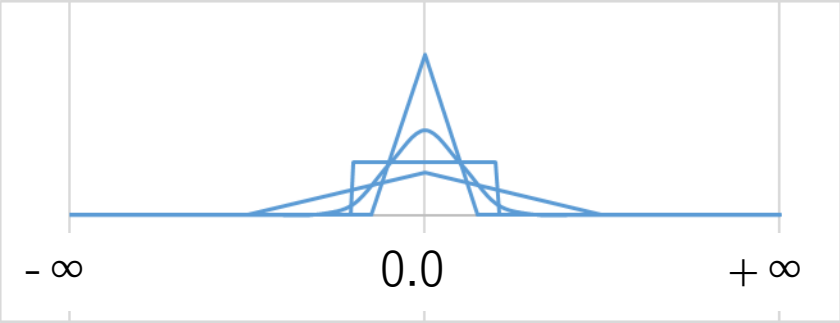
区間推定はできる。

区間内の一様分布ではない。



「0.0」に近いほど、確率密度は大きいと推定。

分布の形状は分からない。



既知の標本が1つしかない場合を考えてみます。
目的変数は連続値で「0.0」だとします。
未知の標本の値が、厳密に「0.0」と一致するとは推測できません。
ですが、「0.0」近い値であるほど、確率密度が大きいだろうとは推測できます。
しかし、どれだけ「0.0」に近づけば、どれだけ確率密度が大きくなるかは、ノーヒントなので分かりません。
また、「0.0」より大きいか小さいかは、ノーヒントなので各50%だと推測します。
「-∞～0.0」=50%,「0.0～+∞」=50%という、区間推定はできます。
ただし、この区間の中で一様分布になっているとは推定しません。
分布形状は分かりませんが「0.0」に近いほど確率密度は大きいと推測します。

2標本からの推論 (1)

目的変数：「0.0, 1.0, 不明」（連続値）

50%

“-∞～0.0”: 25%　“0.0～+∞”: 25%

50%

“-∞～1.0”: 25%　“1.0～+∞”: 25%

“0.0～1.0”: 12.5～25%　“1.0～+∞”: 12.5～0%

“-∞～0.0”: 12.5～0%　“0.0～1.0”: 12.5～0%

均質 ～ 近い方に偏る

合計結果

“-∞～0.0”: 25～37.5%　“0.0～1.0”: 25～50%　“1.0～+∞”: 25～37.5%

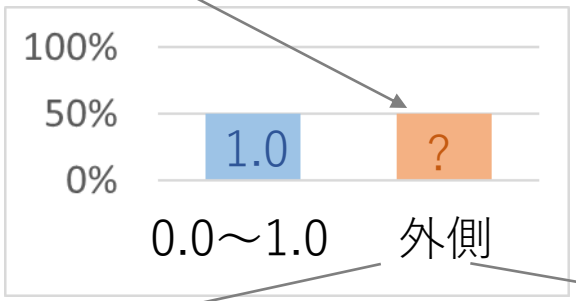
既知の標本が2個で、「0.0」と「1.0」の場合も考えてみます。
標本が1個の場合の結果を各50%で合成します。
「-∞～0.0」「0.0～+∞」「-∞～1.0」「1.0～+∞」の4区間が各25%になります。
「0.0～+∞」=25%を、「0.0～1.0」と「1.0～+∞」に分けようと思います。
ヒントがないので均等だと考えれば各12.5%なります。
ですが、0.0に近い方が確率密度が大きいとはいえるので、極端な場合、25%と0%になります。
「0.0～+∞」=25%は、「0.0～1.0」=12.5～25%と、「1.0～+∞」=0～12.5%に分割できました。
「-∞～1.0」についても同様に分割します。
合計した結果、「-∞～0.0」=25～37.5%,「0.0～1.0」=25～50%,「1.0～+∞」=25～37.5%になりました。

2標本からの推論 (2)

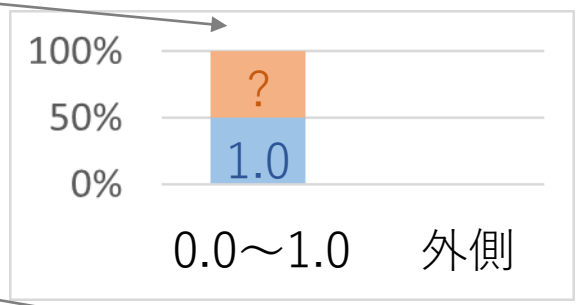
目的変数：「0.0, 1.0, ?」（連続値）

基準

頻度



～



平均結果

“-∞～0.0”: 0~25%　“0.0～1.0”: 50～100%　“1.0～+∞”: 0～25%

前ページの結果

“-∞～0.0”: 25～37.5%　“0.0～1.0”: 25～50%　“1.0～+∞”: 25～37.5%

二つの結果を合わせる

“-∞～0.0”: 25%　“0.0～1.0”: 50%　“1.0～+∞”: 25%

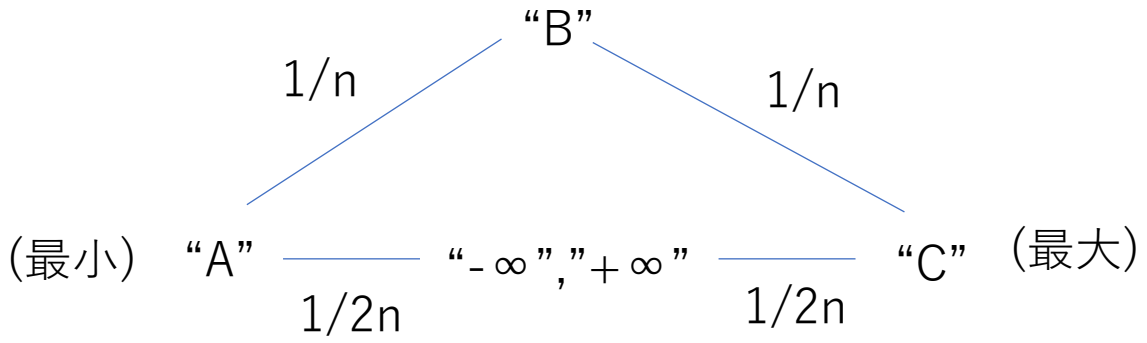
同じ例で、標本が一定の区間に存在する頻度についても考えてみます。
1個目の標本を基準として、2個目と、推測対象の3個目の存在区間を見ます。
2個目の標本の値は「1.0」なので、「0.0～1.0」区間に100%存在します。
3個目の標本は不明なので、「0.0～1.0」区間に、0～100%存在します。
二つを平均すると、「0.0～1.0」区間に、50～100%存在します。
残りの0%～50%は「-∞～0.0」と「1.0～+∞」に均等に分かれます。
結果、「-∞～0.0」=0～25%,「0.1～1.0」=50～100%,「1.0～+∞」=0～25%
前ページの結果も満たそうとすると、「-∞～0.0」=25%,「0.1～1.0」=50%,「1.0～+∞」=25%と定まります。

n標本からの推論

目的変数：「A, B, C, ?」（連続値）

“-∞～A”: 1/6　“A～B”: 1/3　“B～C”: 1/3　“C～+∞”: 1/6

既知の標本数: n



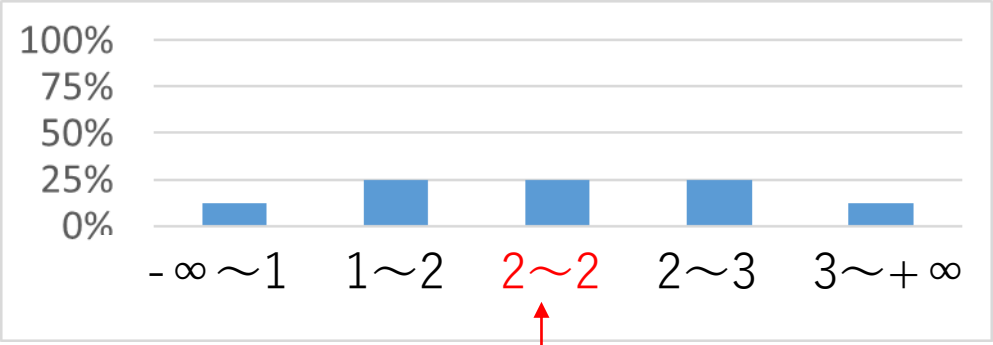
推測値は、最小値と最大値を繋げた円のn個の区間で各1/n

（端を繋げた部分は、半分に分ける）

既知の標本が3つで、「A, B, C, ?」の場合も同様の計算をしてみます。
「-∞～A」=1/6,「A～B」=1/3,「B～C」=1/3,「C～+∞」=1/6になります。
「A～B」,「B～C」,「C～+∞,-∞～A」の3区間に均等だと解釈できます。
標本が4つ以上の場合でも、円になるように最大値と最小値を繋げて、全ての区間に均等になります。
端を繋げた部分だけは、半分に分かれます。
標本がn個あれば、n個の区間ができ、それぞれが1/nになります。

重複値の推論

目的変数: 「1, 2, 2, 3, ?」



1/4の確率で、「2」に厳密に一致すると推測

「1」 基準

「1, 2」

「1, 2, 2」

「1, 2, 2, 3」

「1, 2, 2, 3, ?」

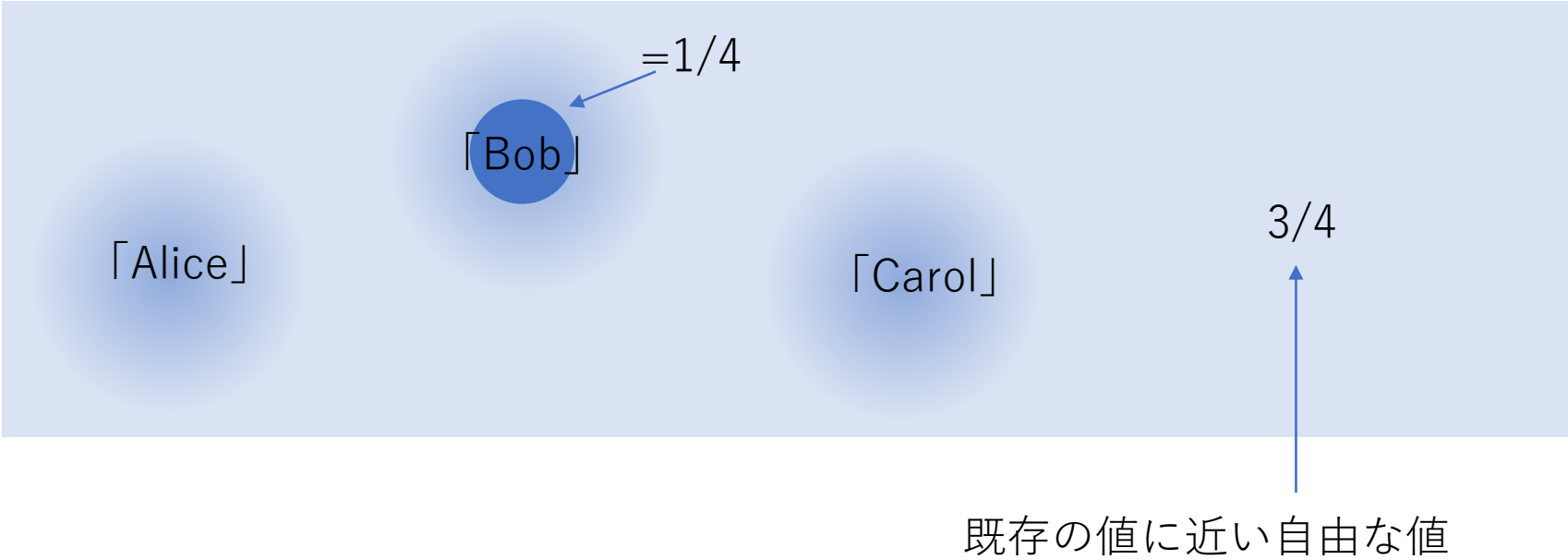
既知の値が出現した比率(≧1/4)と一致

既知の標本の値が重複している場合を考えてみます。
値が「1, 2, 2, 3, ?」の場合を計算します。
「-∞～1」=1/8,「1～2」=1/4,「2～2」=1/4,「2～3」=1/4,「3～+∞」=1/8となります。
ここで、「2～2」=1/4というのは、厳密に「2」に一致すると推測するものです。
検証のため、出現済みの値が再出現した頻度を数えてみます。
1個目の標本では、再出現することがないので、基準と考え除外してカウントします。
2個目から5個目の標本では、4回中、1回以上は再出現します。
推測値の重複頻度は、既知の標本の重複頻度と一致しました。

名義の推論

目的変数: 「Alice, Bob, Bob, Carol, ?」

「～Alice」=1/4　「～Bob」=1/4　「Bob～Bob」=1/4　「～Carol」=1/4



既知の標本が名義尺度の場合を考えます。
値が「アリス, ボブ, ボブ, キャロル, ?」だとします。
「ボブ～ボブ」=1/4は、離散値のときと同様です。
残りの3/4ですが、名義尺度なので大きい順に並べることができません。
「不明」な標本としては、未知の値が出る可能性があります。
また、既存の値が再出現する可能性があります。
そのため、3/4部分は、既存の値に近い自由な値といえます。

0標本からの推論

目的変数: 「?」

四択問題　　それぞれの選択肢=25%

目的変数≧0　　「0〜+∞」の一様分布

質問者が定めた回答選択肢をそのまま標本とみなすことができる

質問: あるリンゴの糖度は？

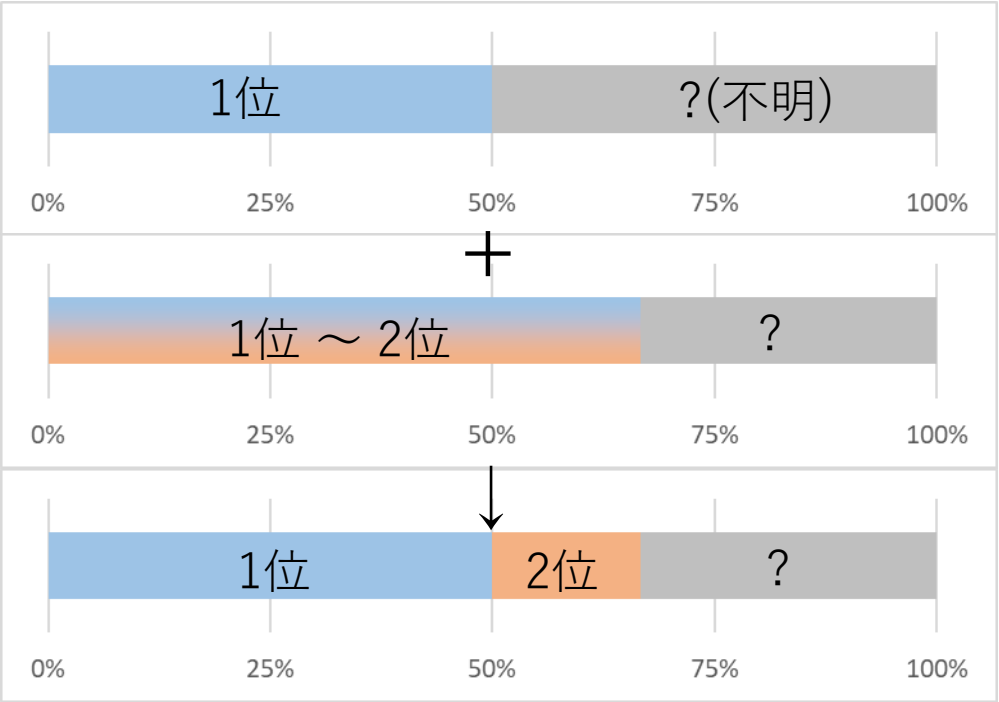
糖度の実測値が一つもない場合

理論上取りうる糖度の値域すべてを、標本とする

もし標本が足りないと感じたら、
より品質の悪い標本を使えば良い。

既知の標本が一つもない場合を考えます。
もし、四択問題なら、それぞれの選択肢が25%と推論します。
また、目的変数が0以上という場合は、0以上の実数の一様分布だと推測します。
これは、質問者が定めた回答選択肢をそのまま標本とみなしています。
例えば、リンゴの糖度を推測したい場合を考えます。
糖度の実測値のデータが一つもなかったとします。
それなら、理論上取りうる糖度の値域すべてを、標本とします。
もし標本が足りないと感じたら、より品質の悪い標本を使えば良いのです。

推論の合成 (1)

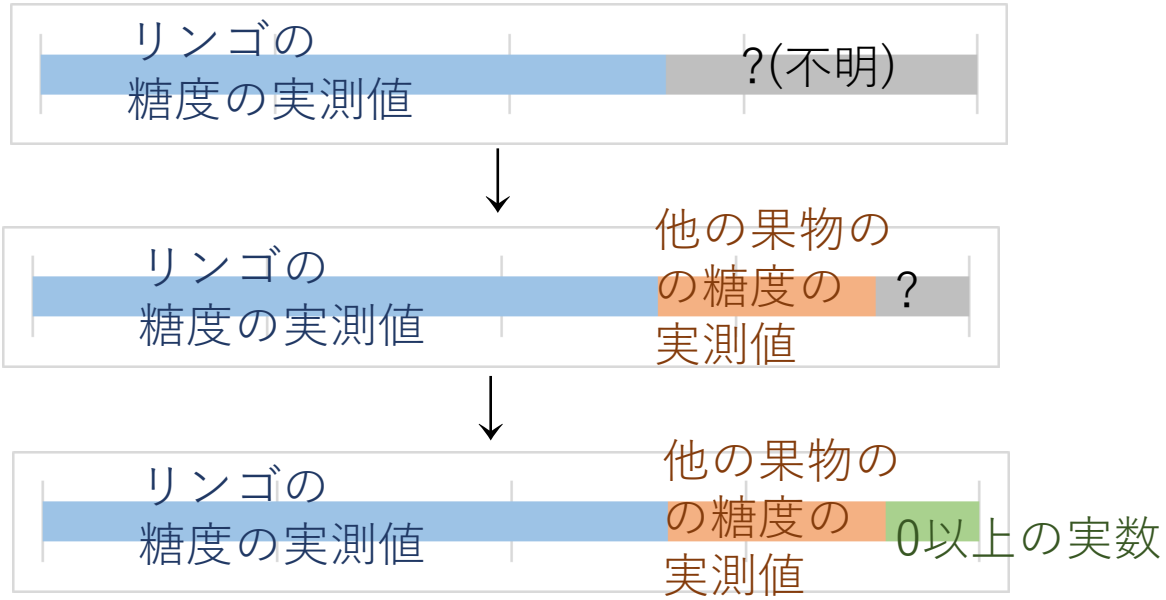


矛盾しない限り、「不明」部分に、他の推論結果を割り当てできる

推論結果の「不明」部分には、矛盾しない限り、他の推論結果を割り当て可能です。
理想帰納を算出したときには、そのような推論の合成を行いました。
「不明」以外のところは維持して、「不明」にはより品質の悪い推論結果を割り当てました。
「不明」は最悪の推論結果なので、他の推論結果があれば、無視せずに活用するべきです・
この例では、二つの推論に共通する「不明」部分は維持しています。
また、一位の部分は追加せずに、二位の部分のみ追加しています。
一つの標本で、可能な割合を超えるという矛盾を防ぐためです。

推論の合成 (2)

質問: あるリンゴの糖度は？



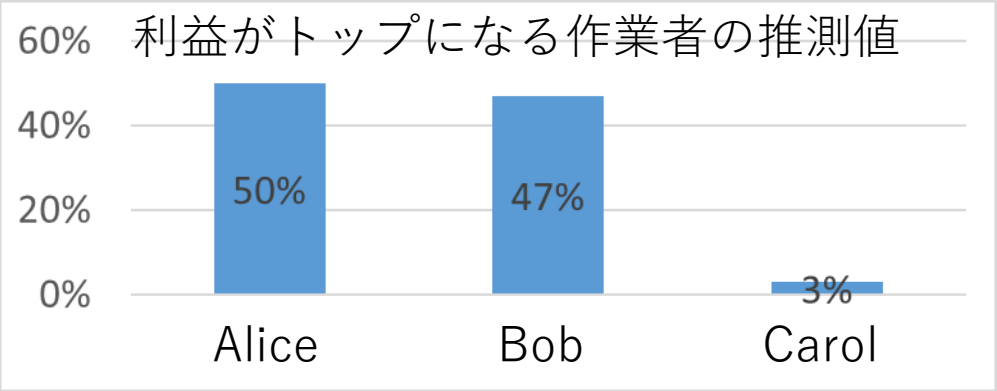
未使用な標本の結果なら、「不明」に割り当てられる

取りうる値を標本にすれば、「不明」をなくすことができる

例えば、標本が一つの場合、推測値の50%は「不明」になります。
「不明」は最悪値なので、別の推論結果があれば、「不明」の部分に割り当てられます。
ただし、矛盾が起こるような割り当ては禁止されます。
ある推論の「不明」部分に、その推論自身を割り当てることはできません。
理想帰納として決まっている、一つの標本が占有可能な割合の限界を超えてしまいます。
「不明」部分に、まだ使用していない2位以降の標本のみの推測値を割り当てるのは可能です。
割り当てた推論にも「不明」があれば、「不明」は残ります。
目的変数が取りうる値の一様分布を推測値として割り当てることも可能です。
その場合、「不明」を完全になくすことが可能です。

回答の量子化 (1)

質問: 最も利益が大きくなるように作業者を一人選んでください



回答(A)　推測値が最大の「Alice」を選ぶ。

誤差程度の差でも、良い方を選ぶべき。

回答(B)　乱数で決める。

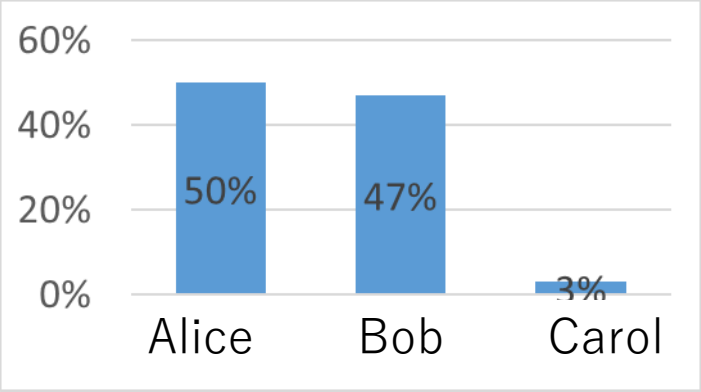
3%の確率で「Carol」が選ばれてしまう。
低確率でも、あえて悪い推測値を選ぶのは誤り。

帰納推論の結果は確率分布として計算されます。
ですが、質問の回答としては一つの選択肢を選ばないといけない場合があります。
例えば、利益を最大化するように作業者を一人選ばないといけないとします。
回答(A)としては、利益がトップになると推測している「アリス」を選びます。
誤差程度でも、最良の選択をするべきできと考えます。
回答(B)は、乱数で決めます。
3%の確率で、「キャロル」が選ばれます。
低確率でも、悪いと推測しているものが選んでしまうのは誤りです。

回答の量子化 (2)

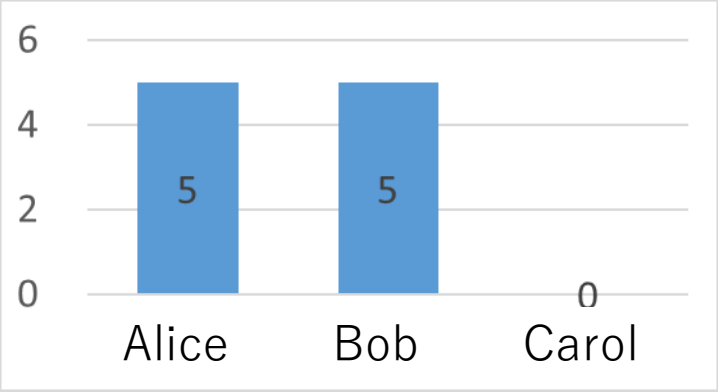
質問: 最も利益が大きくなるように作業者を10人選んでください

利益がトップになる作業者の推測値



量子化

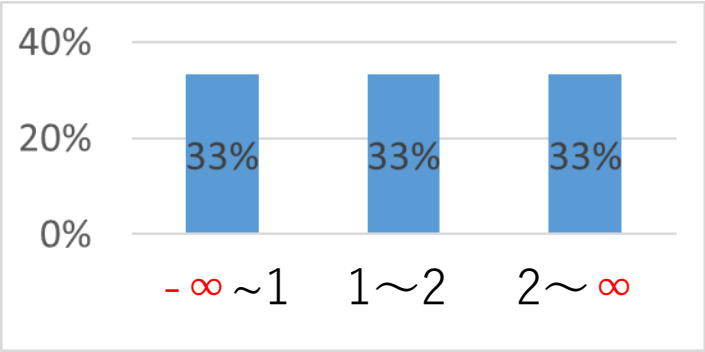
回答

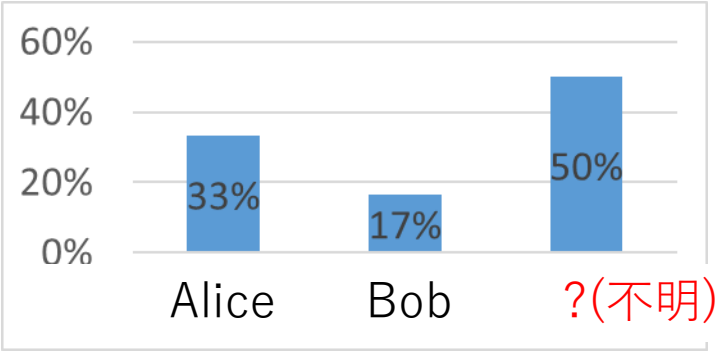


確率分布の「一致率」を最大化するように量子化

次に、一人ではなく十人選ぶ場合を考えてみます。
首位のアリスを十人選ぶのは危険です。
この推論自体に正確なものとは限らないからです。
アリスの成績が予想外に悪い場合の保険に、ボブもある程度は選んだ方が良いでしょう。
直感的には、アリスとボブを同じくらい選ぶのが最適です。
選択肢を選ぶと、確率分布は量子化されます。
確率分布の「一致率」を最大化するように量子化するのが最適です。

回答の量子化 (3)





1.5

回答として
一つの値を選ぶ

Alice

「不明」や「∞」が含まれていても、
「一致率」を最大化するように量子化することができる

「平均値」、「期待値」、「分散」などが計算できなくても推論できる

目的変数には「不明」や「∞」が含まれる場合があります。
そのため、平均値や分散の計算が困難です。
「不明」には取りうる全ての値の確率分布を代入することもできます。
それでも、分散をうまく計算できない場合があります。
ですが、「不明」や「∞」が含まれていても「一致率」を最大化するように量子化することはできます。
「平均値」、「期待値」、「分散」などを計算する必要もなく、推論をすることができます。

統計2.0のまとめ

(1)数が数えられれば、自由に「単位標本」を設定できます。

(2)「OR」「NOT」を使わなければ、自由に標本のグループ分けや順位付けができます。

(3)推測対象は、「不明」な標本として他の標本と同様にグループ分けされます。

(4)複数の標本グループの「一致率」または「標本交換個数」を求め、
標本のウェイトを決めます。

(5)標本のウェイト分布について、「理想帰納」との一致率として「理想率」を求めます。

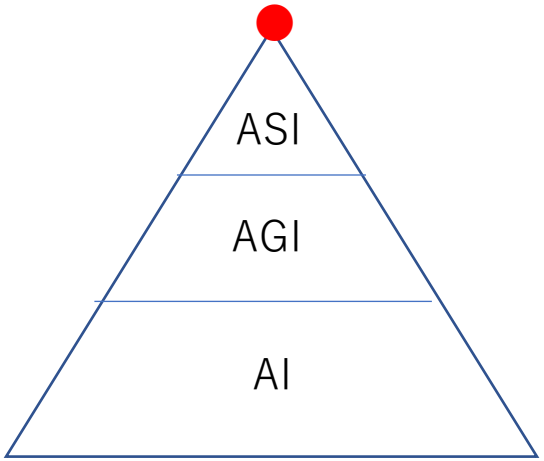
(6)標本ごとに「距離」と「非理想率」を掛けた評価値が、
小さいほど良い「標本選択戦略」です。

(7)母集団を経由せずに、標本の目的変数の分布を「無劣化」のまま推論結果とします。

統計2.0についてまとめます。
(1)数が数えられれば、自由に「単位標本」を設定できます。
(2)「OR」「NOT」を使わなければ、自由に標本のグループ分けや順位付けができます。
(3)推測対象は、「不明」な標本として他の標本と同様にグループ分けされます。
(4)複数の標本グループの「一致率」または「標本交換個数」を求め、標本のウェイトを決めます。
(5)標本のウェイト分布について、「理想帰納」との一致率として「理想率」を求めます。
(6)標本ごとに「距離」と「非理想率」を掛けた評価値が、小さいほど良い「標本選択戦略」です。
(7)母集団を経由せずに、標本の目的変数の分布を「無劣化」のまま推論結果とします。

統計2.0と究極のAGI

究極のAGI: いくら時間が掛かっても良いので、全ての問題で最適解を得る



必要な基礎理論が完成

統計2.0 : 全ての問題に対する客観的な最適解の定義

主観が排除されており、誰にとっても同じ

統計2.0として、究極のAIを目指す全ての人に共通する、究極のゴールが設定された。

統計2.0を理解できたなら、究極のAGIを作るための基礎が身についている。

統計2.0によって、全ての問題に対する客観的な最適解が定義されました。
つまり、究極のAGIに必要な基礎理論が完成しました。
究極のAGIの定義は、時間がいくらかかっても良いので、全ての問題で最適解を得ることです。
この理論に従ったプログラムなら、最適解に辿り着くことが保証されます。
計算の効率は、プログラマーの技量次第です。
一方で、最適解というのは主観が排除されており、誰にとっても同じです。
統計2.0として、究極のAIを目指す全ての人に共通する、究極のゴールが設定されました。
統計2.0を理解できたなら、究極のAGIを作るための基礎が身についています。

参考文献

・前回の動画

・And You

参考文献です。
前回の動画も参考にしてください。
そして、貴方です。
知能を理解すれば、自身の知能も最適化できます。
自分の思考にバイアスがあることを理解できれば、それを回避できます。
ここまでの動画を理解頂けたなら、貴方は素晴らしい知能を持っています。
未開の研究領域であっても、自身の知能を信じて進んでください。

あとがき

この研究はボランティアで行われています。

労働を代替して利益を得るという目的はありません。

究極のAGIによって、不可能が可能になることの方が価値があるでしょう。

ですが、知能を正しく理解すれば、人間が究極の知能に辿り着くこともできます。

知能を理解するのが研究の目的です。

究極のAGIを何かに利用する予定はありません。

優れているとアピールするためのベンチマークも不要です。

なので、実行可能なプログラムコードにするのも後回しです。

Contact: ai@ultagi.org https://ultagi.org/

この研究はボランティアで行われています。
労働を代替して利益を得るという目的はありません。
究極のAGIによって、不可能が可能になることの方が価値があるでしょう。
ですが、知能を正しく理解すれば、人間が究極の知能に辿り着くこともできます。
知能を理解するのが研究の目的です。
究極のAGIを何かに利用する予定はありません。
優れているとアピールするためのベンチマークも不要です。
なので、実行可能なプログラムコードにするのも後回しです。

次回予告

次回の動画は未定です。

ここまでの動画で述べた基礎理論を理解できていれば、すでに究極のAGIを作成できます。

利益を目的として、世界で初めて究極のAGIを作成したいのなら、

すぐに応用研究を始めるべきでしょう。

次の動画を待っていたら、他の人の先を越されてしまうかもしれません。

ここから先の研究は、効率よく最適解に辿り着く方法についてになるでしょう。

最高効率である必要はないので、ここまでの研究と比べたら、重要度は低いと思います。

ぜひ、貴方の手で世界で初めて究極のAGIを作成してください。

次回の動画は未定です。
ここまでの動画で述べた基礎理論を理解できていれば、すでに究極のAGIを作成できます。
利益を目的として、世界で初めて究極のAGIを作成したいのなら、すぐに応用研究を始めるべきでしょう。
次の動画を待っていたら、他の人の先を越されてしまうかもしれません。
ここから先の研究は、効率よく最適解に辿り着く方法についてになるでしょう。
最高効率である必要はないので、ここまでの研究と比べたら、重要度は低いと思います。
ぜひ、貴方の手で世界で初めて究極のAGIを作成してください。

お問い合わせ先

お問い合わせは、

こちらからお願いします。

https://ultagi.org/