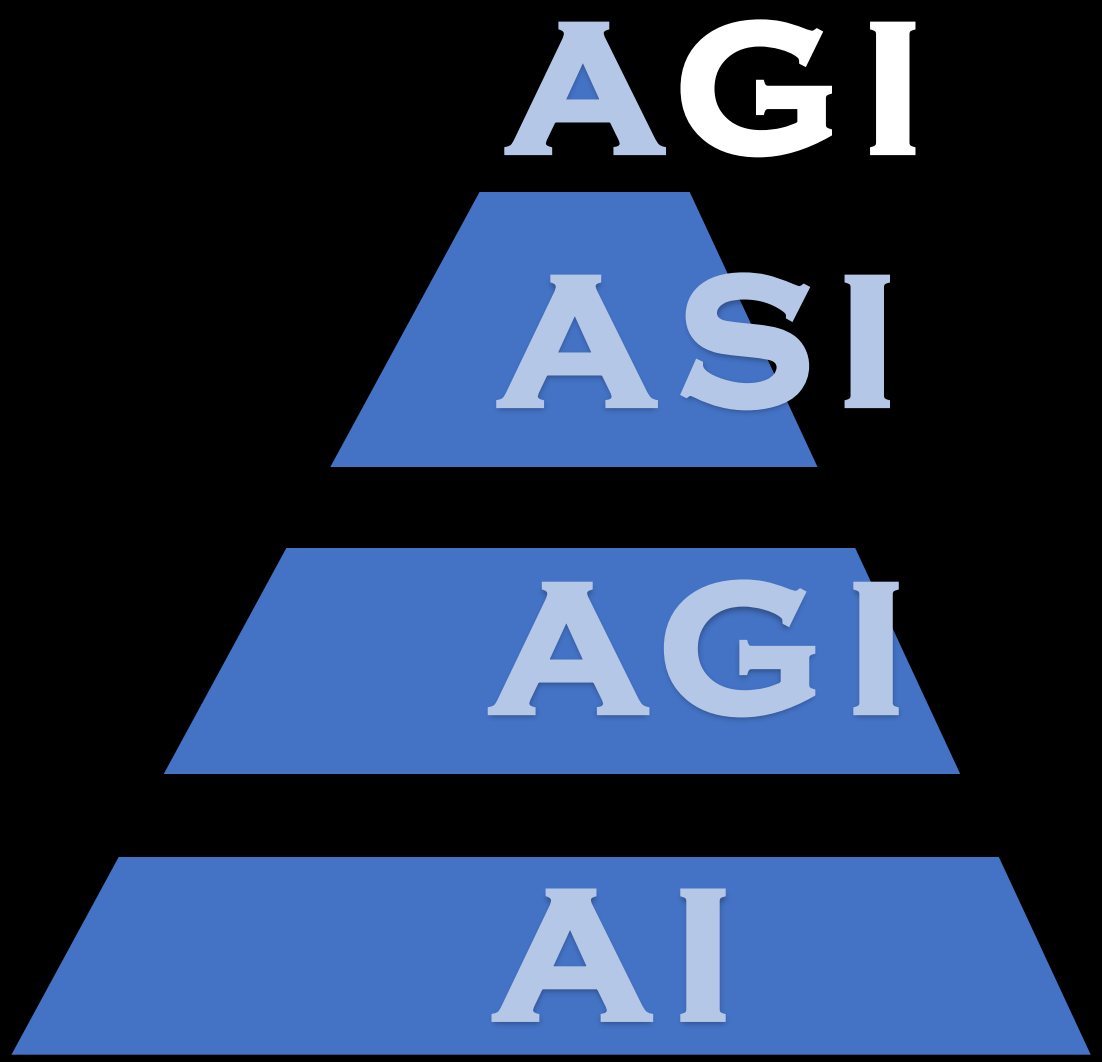
Ultimate AGI



Theory (2)

Statistics 2.0

Index (1/2)

Index (2/2)

Review

Part 1 (Previous video)

- We defined the ultimate AGI as something that can provide the optimal answer to any question.
- To achieve this, it was necessary to determine the optimal method of inductive inference.
- We also clarified several conditions that the ultimate AGI must satisfy.

Part2 (This video)

- We will determine the optimal method of inductive inference that will satisfy these conditions.
- We will explain "Statistics 2.0", an expansion of traditional concepts of statistics.

Let's review what we learned last time.
We defined the ultimate AGI as something that can provide the optimal answer to any question.
To achieve this, it was necessary to determine the optimal method of inductive inference.
We also clarified several conditions that the ultimate AGI must satisfy.
In this article, we will determine the optimal method of inductive inference that will satisfy these conditions.
We will explain "Statistics 2.0", an expansion of traditional concepts of statistics.

Configuring Statistics 2.0

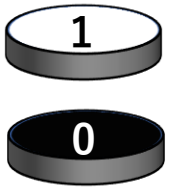
(0) Defining the Difference Between Statistics 2.0 and Traditional Statistics

Statistics 2.0
- (1) Definition of valid samples and valid methods for selecting samples
- (2) Define how to evaluate sample selection methods
- (3) Define the value to be inferred from the selected sample

Statistics 2.0 consists of three parts.
First, we will define how Statistics 2.0 differs from traditional statistics.
Then, we will explain the three components.
The first is the definition of a valid sample and a valid method for selecting samples.
The second is the definition of how to evaluate the merits of a sample selection method.
The third is the definition of the value that is inferred from the selected sample.

Problems in frequentist statistics (1)

Numbers on the heads and tails of the coin



Past coin toss

Average = 1

Standard deviation = 0

Future coin toss

Estimated value = 1

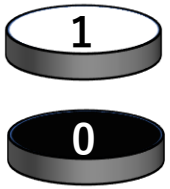
(100%)

It is clearly a mistake to be able to predict the future with 100% certainty.

Let's first look at the problems with traditional frequentist statistics.
As a simple example, suppose you toss a coin twice, landing on heads (1) two times and tails (0) zero times.
In this case, the mean is 1 and the standard deviation is 0.
We estimate that the results of the two trials are samples drawn from a population in which heads will land 100% of the time.
The next, third trial is also drawn from the same population, so we estimate that it will land 100% of the time.
However, it is clearly a mistake to be able to predict the future with 100% certainty.
This problem is induction, not deduction.
Even if the more samples we have, the more certain we can be, but 100% certainty is not permissible.

Problems in frequentist statistics (2)

Numbers on the heads and tails of the coin



Past coin toss

Average = 1

Standard deviation = ∞

(Incalculable)

Future coin toss

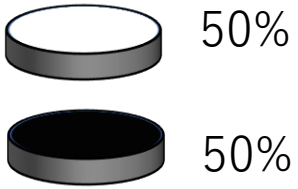
(unknown)

It is a mistake to say that something is completely unknown
when there is one sample, just as it is when there are zero samples.

As the next example, suppose a coin is tossed once and heads (1) lands once and tails (0) 0 times.
In this case, the standard deviation becomes infinite and cannot be calculated.
However, it is strange to say that we cannot infer anything just because we cannot calculate it.
We can at least tell that it is not a coin that will land tails 100% of the time.
Intuitively, we can deduce that, if anything, it may be more likely to land heads.
In principle, having even one sample should provide more certainty than having none.
It is a mistake to say that something is completely "unknown" when there is one sample, just as it is when there are zero samples.

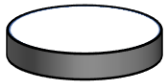
Problems in bayesian statistics (1)

Subjective
prior probability



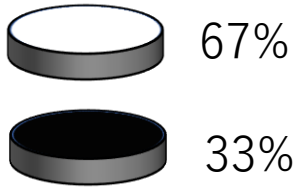
×

Objective
Observations



=

Posterior probability

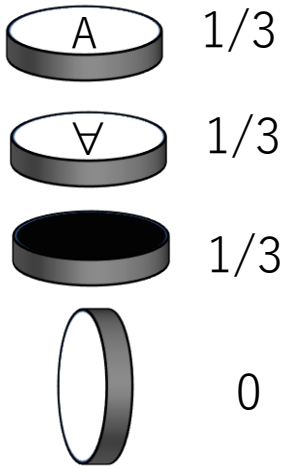


It is a mistake to subjectively decide the prior probability
just because it is unknown.

We are also considering interpreting it using Bayesian statistics.
If we assume a prior distribution of 50% heads and 50% tails, we can make an estimate even with just one sample.
Based on the principle of indifference, we assume a uniform distribution for the two possible distinguishable states.
It is a mistake to subjectively decide the prior probability just because it is unknown.

Problems in bayesian statistics (2)

Subjective
prior probability



1/3
1/3
1/3
0

Uniform distribution

In reality
it is not 0

There is no information
about the frequency of the three states.

There is subjective information
that these are the three possible states.

If there is truly no information,
the prior probability can only be said to be "unknown".

Let's assume that only the front side of the coin has a picture "A", and that we can distinguish between up (A) and down (∀).
Then, 1/3 each of "heads (up), heads (down), tails" are uniformly distributed.
We also assume that the probability of the coin standing up without falling over is 0, but in reality it is not 0.
This uniform distribution is called an uninformative prior distribution, but it is not truly uninformative.
There is no information about the frequency of the three states.
However, there is subjective information that these are the three possible states.
If there is truly no information, the prior probability can only be said to be "unknown".

Hypothesis test

Information of value : No arbitrary additions (prior probability)

Information of distribution : Arbitrary additions (normal distribution, etc.)

Question: A certain drug is effective or not?

Result: ~~"effective" or "not"~~

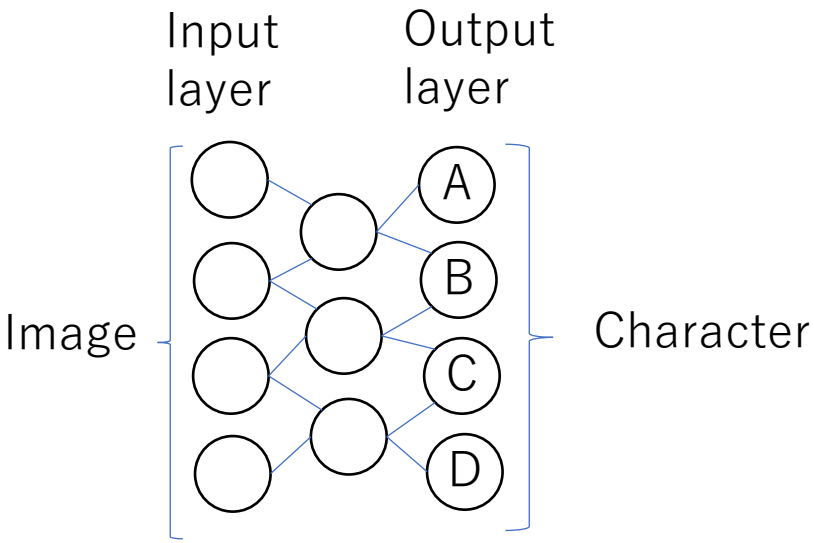
"effective" or "unknown"

When the sample size is extremely small

On the other hand, let's think about information in statistician statistics.
Consider a hypothesis test of whether a certain drug is effective or not.
It does not arbitrarily add new value information, like a prior probability.
However, it does arbitrarily add information about the distribution of values, like a normal distribution.
Because it does not add value information, if the sample is extremely small, you will not get reasonable results.
One thing that is easy to misunderstand is that hypothesis testing does not determine whether the drug is effective or not.
It determines either the drug is effective or it is unknown.
If the sample is extremely small, the result will be "it is unknown".

Character recognition NN

Character Recognition Neural Network



Initial value
(prior probability)

: Random information

The initial values and the information used for learning are mixed together and cannot be distinguished.

Next, let's consider the statistical inference performed by a character recognition neural network.
Suppose each neuron in the output layer corresponds to one character.
The initial value of the neuron is set randomly, and corresponds to the prior probability.
If it has only learned from a single sample, it will output an almost random initial value.
Even randomly determined initial values contain information.
Furthermore, the initial value and the information used for learning become mixed up and cannot be distinguished.

Definition of Unknown

In Statistics 2.0

We prohibit the use of information that the questioner
has not specified as "available information", such as random initial values.

Definition of "Unknown“:

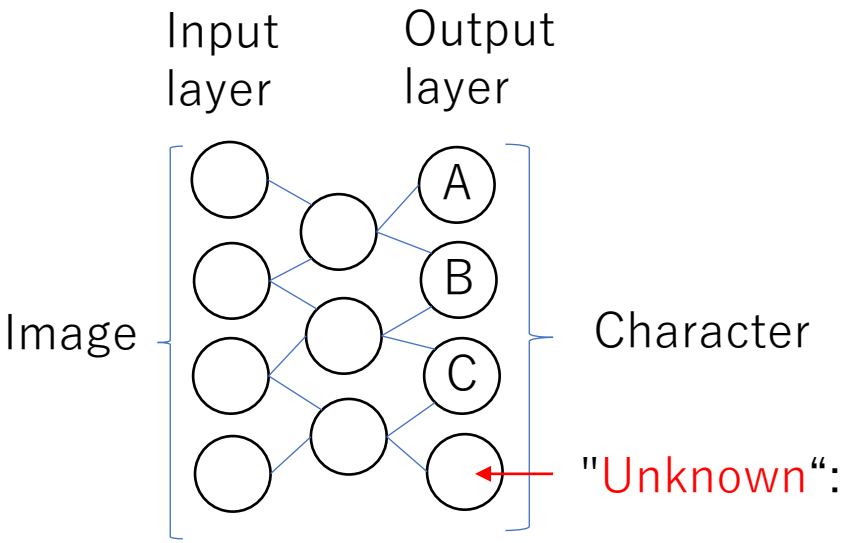
A special value representing no information.

	Treatment of "unknown"
Frequentist statistics	Only handle non-"unknown" values
Bayesian statistics	Assign a subjective value to "unknown"
Statistics 2.0	"unknown" will remain "unknown"

Now, let's consider the new "Statistics 2.0".
We prohibit the use of information that the questioner has not specified as "available information", such as random initial values.
We define a special value called "unknown" to represent the absence of information.
We have summarized how "unknown" is handled in the table.
In frequentist statistics, estimates are made using only values that are not "unknown".
In Bayesian statistics, estimates are made by substituting subjective numerical values for "unknown".
In Statistics 2.0, "unknown" is estimated while remaining "unknown".

Unknown initial value

Character Recognition Neural Network



Initial value (prior probability) : "Unknown“ (100%)

The more values other than "unknown" are learned, the smaller the output for "unknown" will be.

Let's consider a method of estimation while keeping the initial value of "unknown".
Let's consider the example of a character recognition neural network.
Decide that one of the output neurons will represent "unknown".
Before learning, we set it so that "unknown" is output 100% of the time, rather than randomly.
The more values other than "unknown" are learned, the smaller the output for "unknown" will be.

Adding unknown sample

Sample size		Typical Learning rate		Output
n = 9	×	0.1%	=	0.9%: Learning data 99.1%: Initial "Unknown“

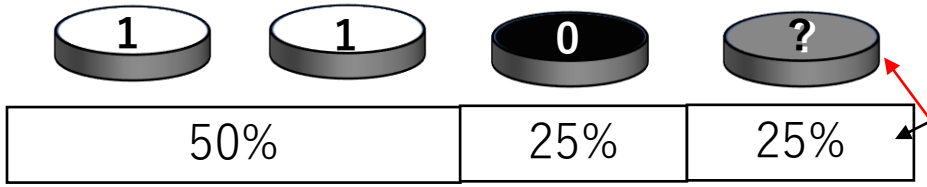
Sample size		Optimal Learning rate			n
n = 9	×	10%	=	90%: Learning data	(9 × 10%)
				10%: Initial “Unknown“	(1 × 10%)

After adding one (n=1) "unknown" sample,
it is optimal for all samples to affect the output equally.

A neural network's learning rate is generally around 0.1%.
The unlearned portion of the output remains at its initial value of "unknown".
If the sample size is n=9, the total learning rate is 0.9%, so the remaining 99.1% is the initial value of "unknown".
If you know your sample size is small, you should increase the learning rate.
If n=9, a learning rate of 10% is considered optimal.
Each of the 9 samples will affect the output by 10%, and the remaining 10% will be the initial value of "unknown".
In other words, after adding one "unknown" sample, it is optimal for all samples to affect the output equally.

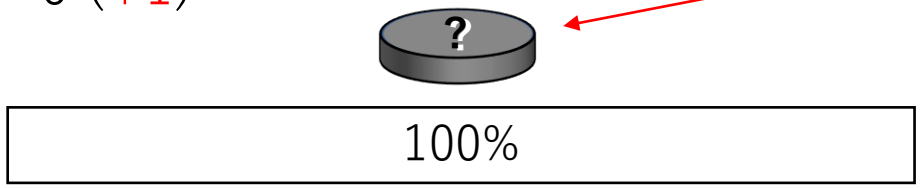
Interpretation of unknown sample

n = 3 (+1)



出力への影響度
(推測値の確率部分布)

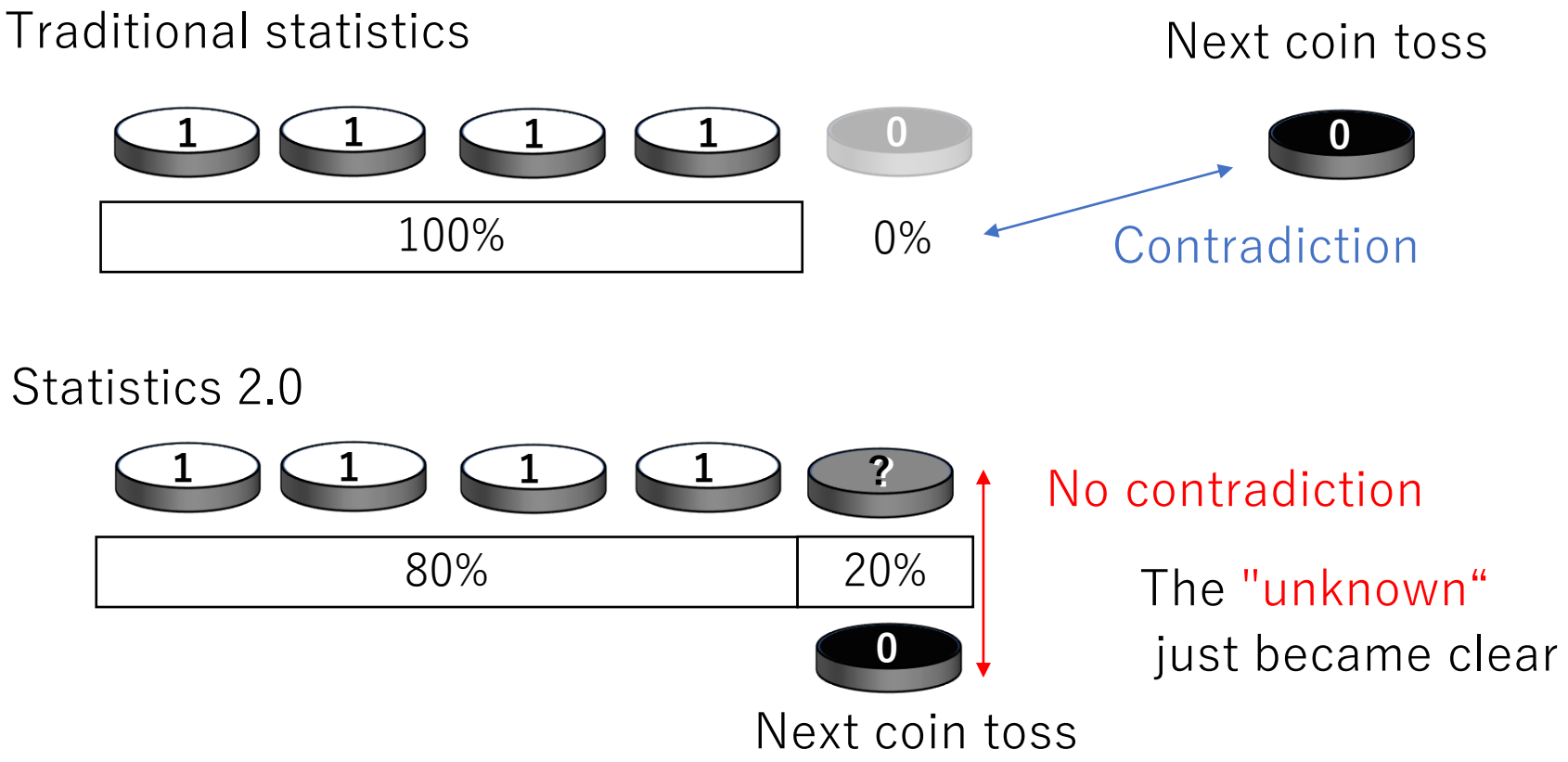
n = 0 (+1)



The "unknown" sample
can be interpreted as the
sample you are trying to guess.

Let's consider the case where a coin is tossed twice, heads twice and tails once.
Add one "unknown" sample.
Heads will affect the output 50%, tails 25%, and "unknown" 25%.
The probability distribution of the magnitude of the impact on the output can also be considered as a guess value.
Here, the "unknown" sample can be interpreted as the sample you are trying to guess.
If n=0, then you will make a guess only from the sample you are trying to guess.
Since the value of the sample you are trying to guess is "unknown", the guess value will also be "unknown" = 100%.
The reason why we add one is because there is only one thing to guess.

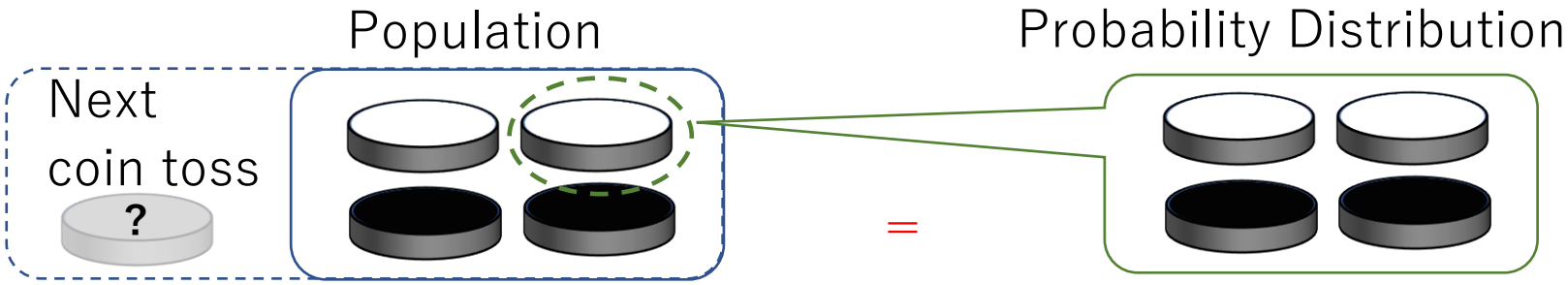
Contradiction



In traditional statistics, population estimates are made only from samples with known values.
In Statistics 2.0, population estimates are made including a sample of the "unknown" value that you are trying to estimate.
If a coin is tossed and the result is 4 heads and 0 tails, traditional statistics would estimate that heads will appear 100% of the time.
If the next coin toss results in tails, this would contradict the estimate.
In contrast, Statistics 2.0 assumes that heads will appear 80% of the time and "unknown" will appear 20% of the time.
If the next coin toss results in tails, this is not a contradiction, as the value of "unknown" has simply become clear.

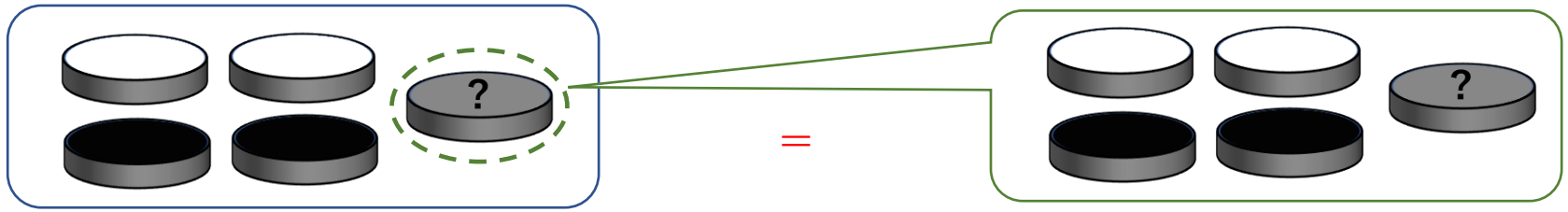
Law of large numbers

Traditional statistics



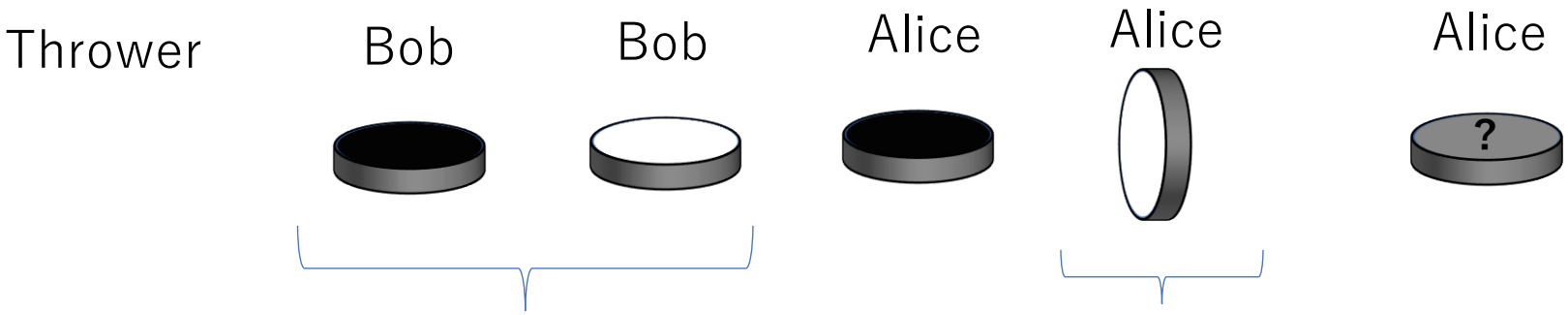
Law of large numbers

Statistics 2.0



Let's think about how the guess value for the next coin toss is justified.
First, we assume that the sample was drawn randomly from the population.
Then, the probability distribution of the values of a sample is equal to the distribution of the values of each sample in the population.
This is the so-called law of large numbers.
In traditional statistics, we assume that the next coin toss will also be from the same population.
If this assumption is incorrect, we will get contradictory results.
On the other hand, Statistics 2.0 includes the result of the next coin toss as an "unknown" sample.
The probability distribution of the values of the "unknown" sample is also equal to the distribution of the values of each sample in the population.

Traditional sample selection



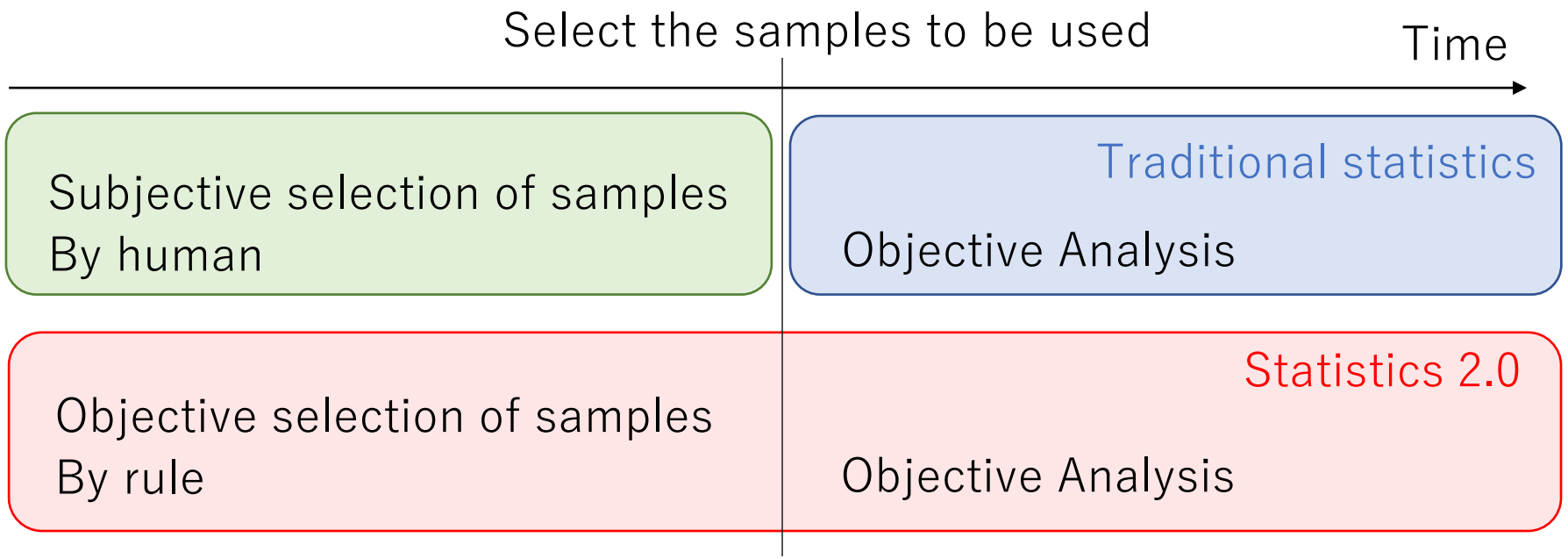
The choice of which sample to use is made subjectively by humans.

The results of the predictions are heavily dependent on human judgment.

Statistical programs assume that all input samples have the same weight.

Next, let's consider how samples to use in statistics have traditionally been decided.
When predicting the result of the next coin toss, the sample would be the results of past coin tosses.
Among these samples there is data from different people; is it okay to include that?
Also, if there is data that seems abnormal, is it okay to exclude it?
The choice of which sample to use is made subjectively by humans.
The results of inferences are heavily dependent on human judgment.
Statistical processing programs assume that all input samples have the same weight.

Definition of Statistics 2.0



Definition of “Statistics 2.0“:

Statistics that extend the scope of objective discussion
to the selection of samples to be used

Now, we will finally define what "Statistics 2.0" is.
Statistics is the use of samples to perform objective analysis.
However, in traditional statistics, the sample to be used is decided subjectively by humans.
In Statistics 2.0, the sample to be used is decided objectively according to rules.
"Statistics 2.0" is defined as statistics that expands the scope of objective discussion to the stage of selecting the sample to use.

Sample selection strategy

“Sample selection strategy“:

A strategy for which samples to use for inference

Statistics 2.0 aims to:

Searching for the best possible "sample selection strategy"

Rules for “Sample selection strategy":

- Required condition

  Specify an effective "sample selection strategy"

- Evaluation criteria

  Define the relative merits of multiple "sample selection strategies"

We will call the strategy of which sample to use for inference the "sample selection strategy".
The goal of Statistics 2.0 is to explore the best "sample selection strategy" possible.
The rules are necessary conditions and evaluation criteria.
The necessary conditions specify a valid "sample selection strategy".
The evaluation criteria specify the superiority or inferiority of multiple "sample selection strategies".
Let's start by deciding the necessary conditions.
Before that, let's define what a sample is.

Sample

“Sample“: Individuals that meet the same conditions
but can be distinguished

“Same condition“: Conditional expressions that
do not include logical sum(OR) and negation(NOT)

 OR “A“ : Invalid Sample Condition

 OR  : Invalid Sample Condition

Fruits AND Red : Valid Sample Condition

A sample can be rephrased as "individuals that satisfy the same conditions but are distinguishable“.
"The same conditions" is defined as "a conditional expression that does not include logical sum or logical negation“.
If logical sum is allowed, definitions such as "(1 apple) or (1 letter of the alphabet)" would also be possible.
The principle of confirmation states that the more related things are observed, the more certainty there is.
If logical sum is allowed, things that have absolutely nothing in common could be used as samples, which goes against the principle.
For example, a condition such as "(1 apple) or (1 strawberry)" is not allowed.
A condition such as "(1 fruit) and (red)" would be allowed.

Prohibition of OR

Question: What is the sugar content of an apple?

Invalid sample condition: The first or third or fifth apple from the left

Sugar content: 8.8 6.3 9.5 5.9 9.2 ?

It is forbidden to allow logical OR in sample conditions because it would allow arbitrary selection.

Valid sample condition: apple

Once the sample conditions have been decided, we must list all apples that meet the conditions.

If we allow logical OR in the sample conditions, we will be allowed to make arbitrary choices.
For example, say you want to estimate the sugar content of a certain apple.
Suppose the conditions for it to be accepted as a sample are "the first, third, or fifth apple from the left“.
If we arbitrarily accept only apples with a high sugar content as samples, the estimated sugar content will also be high.
In this example, we should simply set "apple" as the sample condition.
Once the sample conditions have been decided, we must list all apples that meet the conditions.

Prohibition of NOT

De Morgan's laws

Sample condition: $A\ OR\ B = NOT(\bar{A}\ AND\ \bar{B})$

Prohibited Prohibited

Logical negation is also prohibited from being included in sampling conditions.

A: $x = a$: Valid Sample Condition

$\bar{A}$: $NOT(x = a)$: Invalid Sample Condition

$x \neq a$: Valid Sample Condition

If logical negation can be expressed by inverting the inequality sign, it will become a valid sampling condition.

This explains why logical negation is not permitted as a sampling condition.
Using De Morgan's laws, logical sum can be converted into the logical negation of logical product.
For this reason, logical negation is also prohibited from being included in sampling conditions.
However, if logical negation can be expressed by inverting the inequality sign, it will become a valid sampling condition.

Distinguish samples

Question： What is someone's weight?

Person	Alice	Bob (last year)	Bob (this year)
Weight	44.5kg	55.5kg	55.6kg

As it grows and changes, two samples?

Since the data of the same person is just duplicated, one sample?

Which is better depends on what you want to infer.

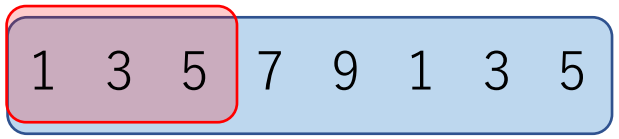
You are free to set how to distinguish samples.

Samples also need to be "distinguishable individuals".
For example, say you want to infer a certain weight.
However, there are two pieces of data for Bob, one from last year and one from this year.
Since the two Bobs have grown and changed, is it fair to say that there are two samples?
In either case, they are the same person, so is there just duplicate data and therefore one sample?
Which is better depends on what you want to infer.
You are free to set how to distinguish samples.
However, if you do not set it up the way the questioner intends, you will not get the results you intended.

Partial overlap

Question: What is the next number in the sequence?

Unit sample: A sequence starting with 1 and ending with 5

Information: 

There is partial overlap, but are there two samples?
or one sample?

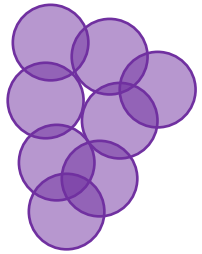
You are free to define overlap, or the number of samples.

When samples have partial overlap, it can be difficult to know whether they should be considered individuals.
Consider the case of guessing the next number in a sequence.
For example, a sequence starting with 1 and ending with 5 is considered to be one sample.
With this definition, there will be partial overlapping samples.
Although there is partial overlap, is it appropriate to say that there are two samples?
It depends on the case.
You are free to define overlap, or the number of samples.

Unit sample

“Unit sample”:　A sample with a sample size of 1

Question: The sugar content of a certain grape



1 bunch

1 grain

1 volume

As long as you can distinguish between duplicates and count them, you can decide what the "unit sample" is freely.

A sample with a sample size of 1 can be referred to as a "unit sample".
You can decide what the unit sample is freely, unless the questioner specifies it.
For example, say you want to estimate the sugar content of certain grapes.
You can define one bunch of grapes as one sample.
You can also define one grape, or a unit volume as one sample.
As long as you can distinguish between duplicates and count them, you can decide what the "unit sample" is freely.
However, if you want to estimate the taste of a single grape, it is best to use one grape as the unit.

Information of sample

Unit sample (1person): The volume of a block with limbs?

Information of “Unit sample”:

~~three-dimensional shape data of the human body~~

the boundary between the human body and clothing is not always clear

the "key" information that allows it to be distinguished

(example: coordinates of the center of the body)

If all you need to do is distinguish yourself from other people, then knowing the center coordinates of your body is enough.

It is possible to define various concepts as a "sample", but let's think about how this can be expressed as information.
For example, let's define a "person" as "the volume of a block with limbs".
However, the boundary between the human body and clothing is not always clear.
For this reason, it is difficult to retain three-dimensional shape data of the human body for each sample.
However, if all that is required is to distinguish one person from others, it is sufficient to know the coordinates of the center of the body.
It is sufficient for each sample to retain the "key" information that allows it to be distinguished.

Objective and explanatory variable

Objective variable: The value you are trying to guess

Explanatory variable: Values used for narrowing down samples, etc.

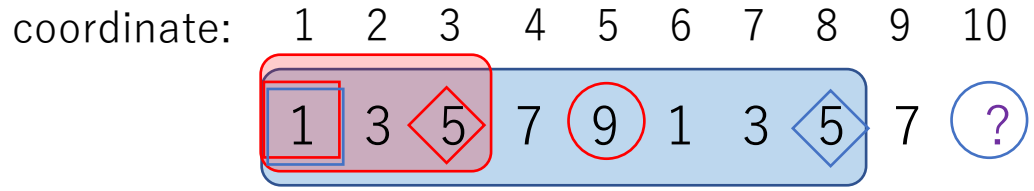
These are the mappings of the sample's "keys"

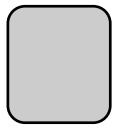
Any unique value can be used
as an explanatory variable or objective variable.

Samples contain more information than just the "key" that distinguishes one sample from another.
The "objective variable" is the value you are trying to infer.
The "explanatory variables" are values used to narrow down the samples, etc.
These values are associated with the sample.
In other words, they are a mapping of the "key".
In other words, any unique value can be used as an explanatory variable or objective variable.

Mapping of key

Question: The next number in the sequence is?



Unit Sample: 

(A sequence from coordinate A to B)
AND (Number at coordinate A = 1)
AND((Number at coordinate B = 5)

Key: (coordinate A, coordinate B) =(1, 3), (1, 8)

Mapping Of key

- Objective variable:  The number two positions away from coordinate B
- Explanatory variable:  Number at coordinate A  Number at coordinate B

Here is an example of the "key" mapping when predicting a sequence.
The sequence of numbers between coordinate A and coordinate B is considered to be the unit sample.
Coordinates A and B are the keys that distinguish the samples.
The objective variable is the number two positions beyond coordinate B.
The explanatory variables are the numbers at coordinates A and B.
These variables are the key mapping.
Conditions are set for the two explanatory variables.

Sample size function (1)

```
FUNCTION Sample_size(Any_information){
        …
    RETURN Sample_size;
}
```

(Argument) Any_information: Any information that could be key

(Return) Sample_size: The sum of the sample sizes included in the argument

$\geqq 0$

Sample_size(Key) = 0:	No valid samples available
Sample_size(A ∪ B) = Sample_size(A)+Sample_size(B):	The samples of key A and key B do not overlap

Next, let's consider what information represents the sample size and the arrangement for distinguishing samples.
These can be expressed as a "sample size function".
The argument can be any information, and any number of pieces of information that can serve as "keys" can be passed in.
The return value is a real number greater than or equal to 0 that represents the sum of the sample sizes in the arguments.
If you pass one "key" to the function, you can find out the sample size.
If the return value is 0, you can tell that it is not a valid sample.
You can check for duplicates by passing two "keys" to the function.
If the sample sizes of the two "keys" are equal to the sum of their respective sample sizes, then there are no duplicates.

Sample size function (2)

Example of Argument: (x1, x2, y1, y2, z1, z2)

Example of Valid return value: (x2-x1)(y2-y1)(z1>0)(z2<0)/C

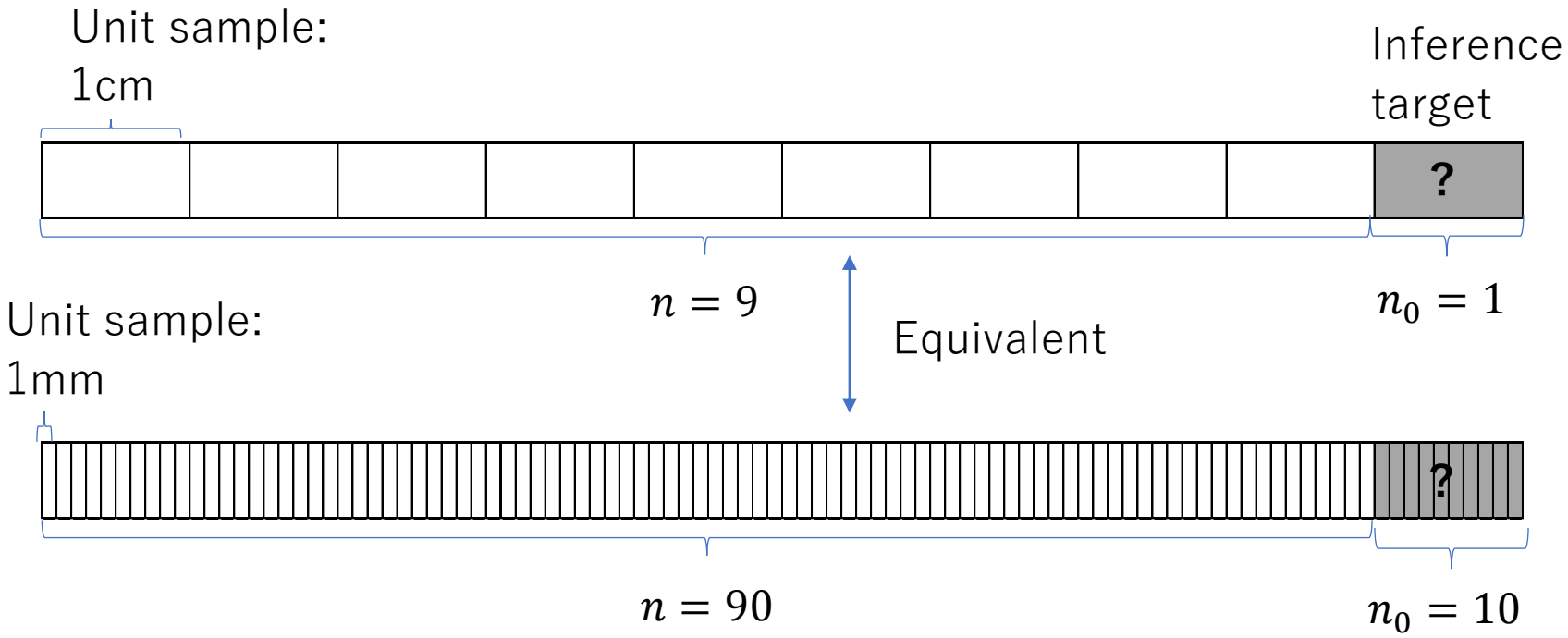
Example of Invalid return value: (x2-x1)+(y2-y1)+(z1>0)+C

- The ratio of some quantity of the "key" mapping to the unit sample size C
- Do not use logical “OR” and logical “NOT”
- It cannot be the sum or difference of different quantities or constants.

Sample_size() = 0

The "sample size function" cannot just be something that returns a unique value.
It must return the number of "unit samples" that satisfy the "same condition".
Arguments include, for example, x1, x2, y1, y2, z1, and z2, which are passed as "keys" to identify the samples.
The sample size can be defined, for example, as "(x2-x1)(y2-y1)(z1>0)(z2<0)/C".
The return value is the ratio of some quantity of the "key" mapping to the unit sample size C.
A condition expression using inequality is assumed to be 1 if satisfied, 0 otherwise.
In order to meet the "same condition", logical OR and NOT are not allowed.
Also, the function cannot be the sum or difference of different quantities or constants, such as "(x2-x1)+(y2-y1)+(z1>0)+C".
Necessarily, if the arguments are empty, the return value will be 0.

Relative sample size



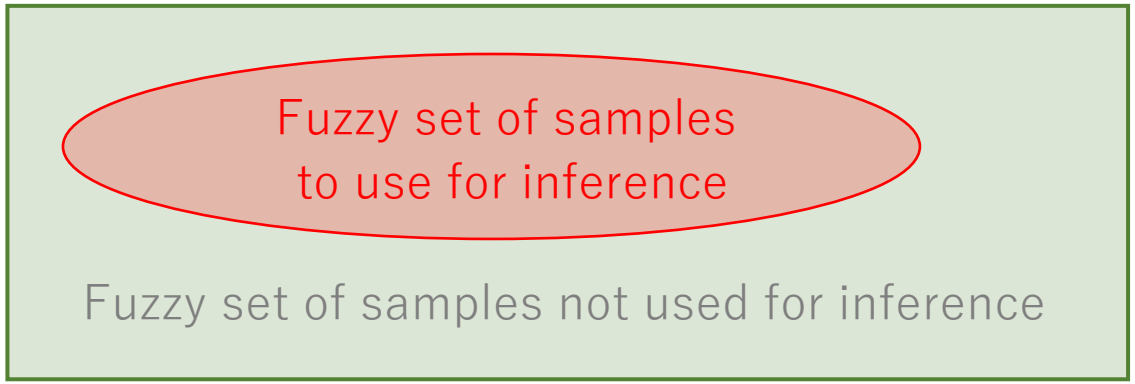
Inductive inference depends on relative sample size $\frac{n}{n_0} = 9$

The sample size is the number of "unit samples".
The target of inference is not limited to a "unit sample"; there may be cases where multiple "unit samples" are inferred together.
For example, suppose you want to infer the color of the remaining 1 cm of a 10 cm strip from the color of the 9 cm part.
If a 1 cm strip is considered to be the unit sample, then there is one item to be inferred.
If a 1 mm strip is considered to be the unit sample, then there are 10 items to be inferred.
This only changes the resolution, the actual content of the inference does not change.
The ratio when the "sample size" of the target to be inferred is set to 1 is called the "relative sample size".
It is the "relative sample size" that works for inductive inference, not absolute values.

Samples used for inference

All samples defined by the sample size function

(Not all samples need to be used for inference)



Membership function: A conditional expression for the degree of inference used

The set used for inference is a single continuous range.

= Do not use logical “OR” or logical “NOT” in condition expressions.

The sample defined by the sample size function is a necessary condition for it to be accepted as a sample.
It is not necessary to use all samples; better inference can be made by using only samples similar to the target of inference.
According to certain conditions, all samples are divided into a fuzzy set of samples to be used and samples to be unused.
The membership function is the conditional expression that shows the degree to which a sample will be used for inference.
The set used for inference must be all samples within a single continuous range.
In other words, the use of "logical OR" and "logical NOT" is not permitted in the conditional expression.

A single contiguous range

$(1 \leqq x \leqq 2)$AND$(3 \leqq y \leqq 4)$: A single contiguous range

Not includes logical “OR” or logical “NOT”

Can be used for inference

$(1 \leqq x \leqq 2)$OR$(3 \leqq x \leqq 4)$: Not a single contiguous range

Includes logical “OR” or logical “NOT”

Cannot be used for inference

Should be divided into three groups

$(1 \leqq x \leqq 2)$

$(3 \leqq x \leqq 4)$

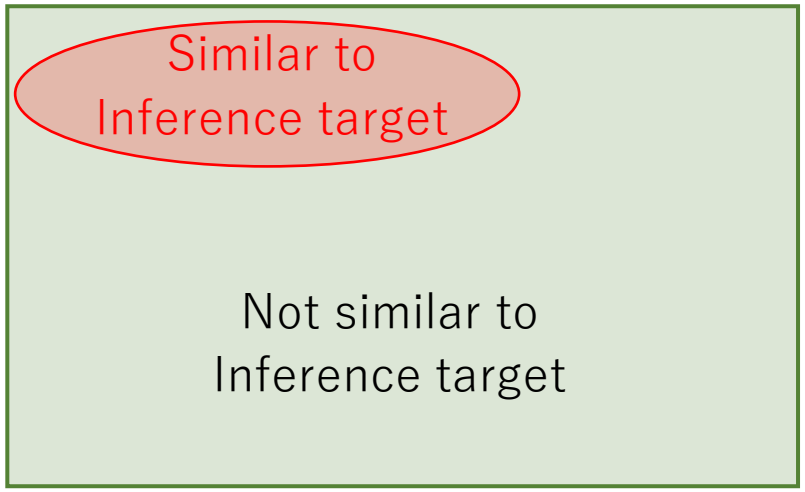
: A single contiguous range

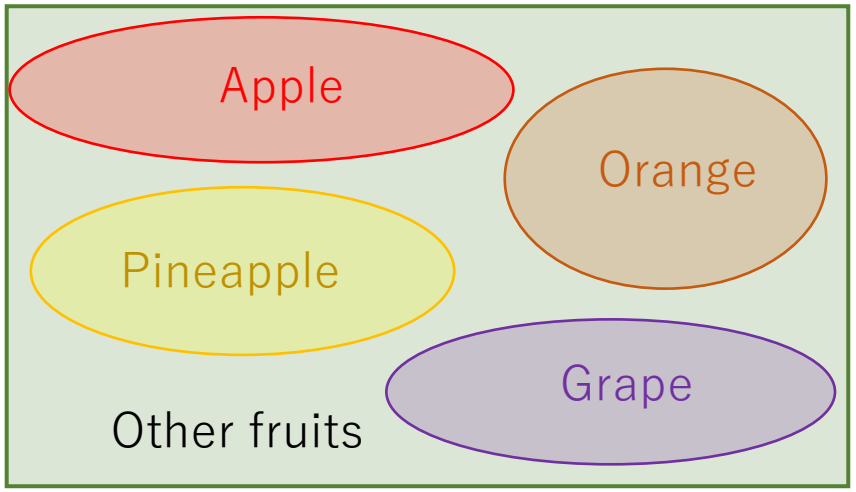
(complement set) : Not a single contiguous range

This explains the relationship between continuous single ranges and conditional expressions.
For example, "$(1 \leqq x \leqq 2)$AND$(3 \leqq y \leqq 4)$" is a continuous single range.
On the other hand, if a logical sum is included, such as "$(1 \leqq x \leqq 2)$OR$(3 \leqq x \leqq 4)$," it cannot be considered a single range.
In that case, it should be separated into three sets: "$(1 \leqq x \leqq 2)$," "$(3 \leqq x \leqq 4)$," and "Other“.
"Other" is the complement of the set, and cannot be expressed without using logical sum and logical negation.
However, if it is not used in inference, it does not have to be a single range.
Either "$(1 \leqq x \leqq 2)$" or "$(3 \leqq x \leqq 4)$" can be used in inference.

Clustering

Classification into two sets is sufficient



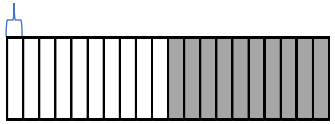


If you cluster the data into three or more sets in advance,
you can make inferences quickly.

It is sufficient to divide the sample sets into two, those that are used for inference and those that are not.
However, there are also advantages to dividing them into three or more sets.
For example, suppose you want to predict the sugar content of a certain fruit.
It is sufficient to divide the samples into two sets, those with matching external characteristics and those that do not.
However, you can also cluster them into three or more sets in advance.
You can cluster them by fruit type.
After the fruit you want to infer is presented, the representatives of each cluster are compared with the matches in external characteristics.
By using only the matching clusters, inference can be made quickly.

Resolution

Unit sample:
1mm



Unit sample:
10mm

Clustering ＝ Reduce the resolution

"specific apple" ⊂ "apple" ⊂ "fruit"

It is more efficient to start with a coarse classification and gradually make it finer.

Here we explain the relationship between clustering and the resolution of the unit sample.
In the example of guessing the color of a certain band, we considered the cases where the unit sample is 1 cm and 1 mm.
When the unit sample is 1 cm, it can be interpreted as there being 10 1 mm samples clustered together.
On the other hand, in the fruit clustering example, there is a relationship where "specific apple" ⊂ "apple" ⊂ "fruit".
This can be interpreted as the larger the clustering, the coarser the resolution.
In other words, clustering and resolution are the same.
The finer the resolution, the better the inference you can make, but the amount of calculations will increase.
If you want to ultimately find the optimal solution, you need to increase the resolution as much as possible.
However, it is more efficient to start with a coarse classification and gradually make it finer.

Prohibition of OR (2)

Hypothesis: "(Pressing button A) OR (Pressing button B) → light C turns on"

Observation: "Pressing button A → light C turns on"
"Pressing button A → light C turns on“
"Pressing button A → light C turns on"
…

Repeated observations have almost certainly verified the hypothesis.
(I haven't pressed button B, is that okay?)

Hypothesis: "Pressing button A → light C turns on" … Verified

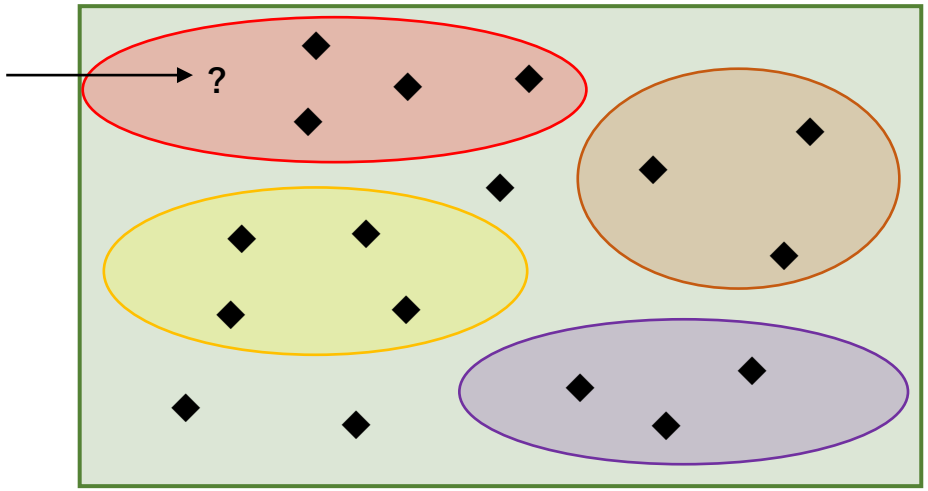
Hypothesis: "Pressing button B → light C turns on" … Not verified

Multiple conditions should not be put into a logical OR
- they must be verified individually.

We will use an example to explain why sets with logical OR conditions should not be used in inductive inference.
We are verifying the hypothesis that for a certain machine, "(Pressing button A) OR (Pressing button B) → light C turns on“.
When button A is pressed repeatedly, light C always lights up, confirming that the hypothesis is almost certainly true.
However, it would be strange to say that light C will light up even if button B is pressed without ever pressing button B.
We should split the hypothesis in two and think that only the hypothesis "Pressing button A → light C turns on" has more certainty.
Multiple conditions should not be put into a logical OR - they must be verified individually.

Sets used for inference

The set to which the sample to be inferred belongs is determined by the same rules as for other samples.



The set to which the inference target belongs is the set used for inference.

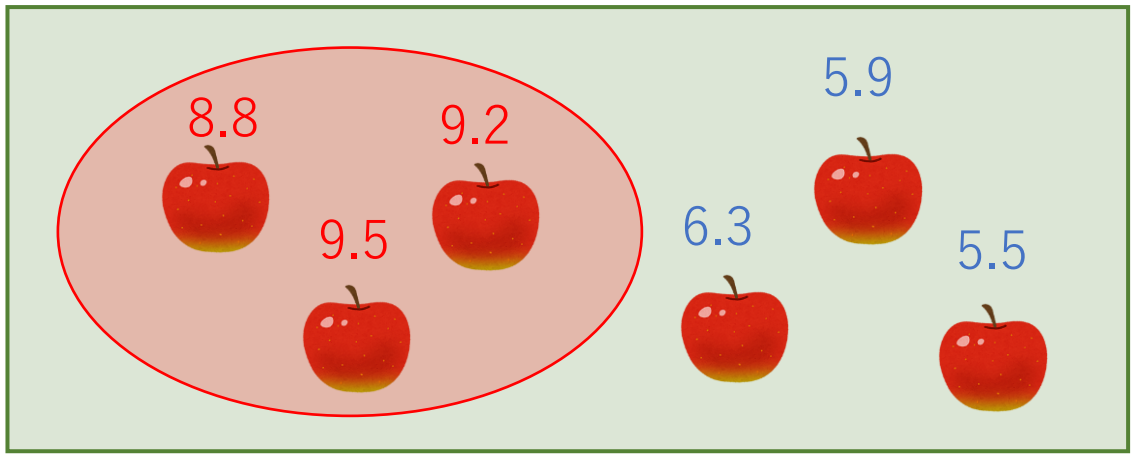
0.5 each may be assigned to fuzzy.

All samples can be classified into any number of sets, but how do we decide which set to use for inference?
The set to which all samples belong is determined by the same rules.
The set to which the sample to be inferred belongs is also determined by the same rules.
In cases such as when the sample is exactly on the borderline, it is permissible for it to belong to two fuzzy sets at 0.5 each.
The set to which the sample to be inferred belongs will be the set to use for inference.
If the sample belongs evenly to all sets, this is the same as using all samples for inference.

Arbitrary clustering

Sugar content: ?

Clustering by sugar content

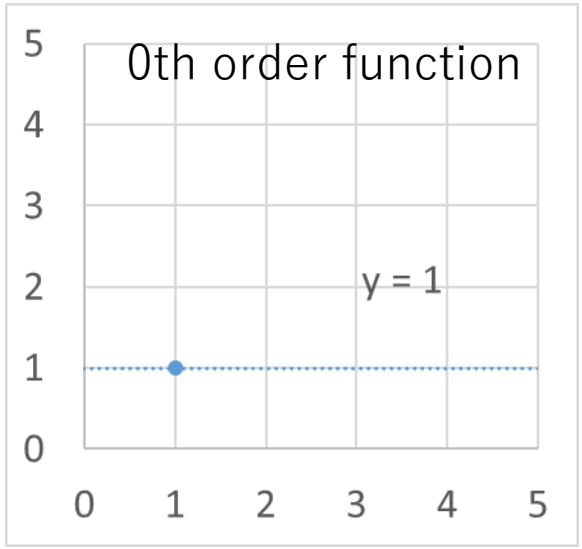


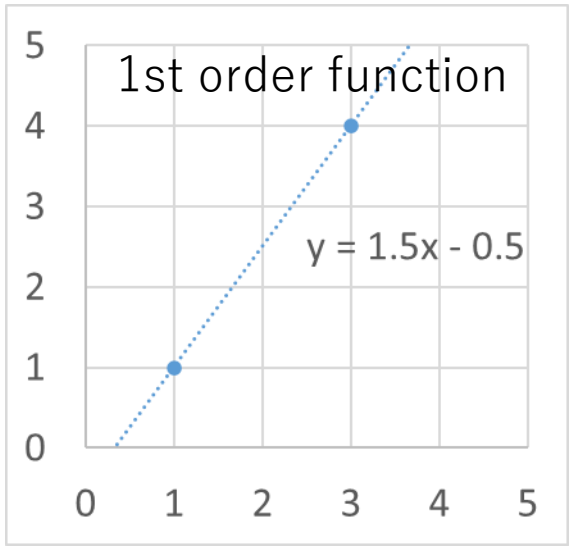
It is unclear which group the subject of the inference belongs to.

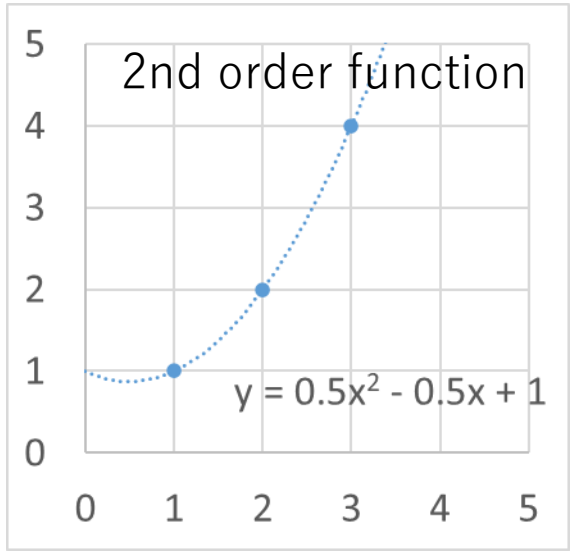
You cannot cluster on unknown values.

Here is an example where arbitrary sample selection is automatically prohibited.
As an example, consider the case of predicting the sugar content of a certain apple.
Let's consider a way to arbitrarily increase the predicted sugar content value.
So, we clustered the samples into samples with high sugar content and samples with low sugar content.
However, we were unable to use the sugar content alone for inference.
This is because the sugar content of the apple we are predicting is "unknown“, so we don't know which category it will belong to.
We cannot cluster based on an unknown value.

Degree of freedom







Nth order function

Number of variables = N+1

$$\text{Residual mean} = \frac{\sum|Residual|}{Sample\ size - (N+1)}$$

Now, let's talk about degrees of freedom.
When calculating standard deviation, you subtract 1 degree of freedom from the sample size.
In regression analysis, the number of variables in the regression equation is the degrees of freedom.
For example, a linear equation has two variables, slope and intercept.
An Nth order equation has N+1 variables.
If there are two points plotted, you can draw a straight line so that the residual is 0.
Therefore, when calculating the average residuals, you subtract N+1 from the sample size.
So far, we've been talking about traditional statistics.

Freedom of sample selection

Condition of sample selection: “0.0<x<6.0”

Variables of sample selection: “0.0”, “6.0”

y : Objective variable	?	1.0	0.8	1.2	1.1	0.9	2.5	0.5
x : Explanatory variable	0	1	2	3	4	5	6	7

Whether or not to include one sample near the boundary is decided arbitrarily to reduce the squared residual.

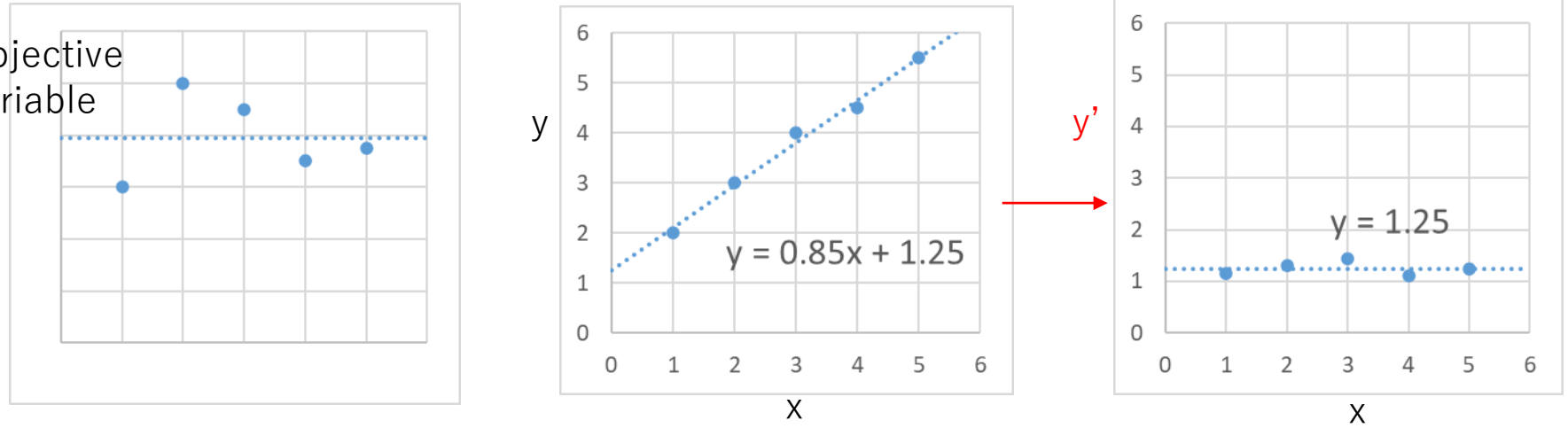
In some cases, the residual mean may not decrease at all.

Freedom of sample selection (Estimation) = $0.5 \times Number\ of\ sample\ selection\ variable$

$$\text{Residual mean} = \frac{\sum|Redisual|}{Sample\ size - (N+1) - Freedom\ of\ sample\ selection}$$

In Statistics 2.0, there are variables for selecting samples, so the degrees of freedom must also be taken into consideration.
For example, if only samples with the condition “0.0<x<6.0" are used, there will be two variables, “0.0" and “6.0".
Since you can freely decide the boundaries, you can decide so that the residual mean is small.
In other words, for each variable, you can arbitrarily choose whether to include or not include one sample near the boundary.
However, the way you decide the boundaries does not necessarily reduce the residual mean for one sample.
Even if the boundaries are changed, there are cases where the residual mean does not decrease at all.
On average, the degrees of freedom can be estimated to be 0.5 for one variable.
The degrees of freedom for sample selection are taken into account when calculating the residual mean.

Coordinate transformation



If you want to guess that the objective variable is not a constant but some kind of function, you can perform a coordinate transformation so that objective variable becomes a constant.

$y' = y - ax \quad = b$ (constant)

If you want to compare the residual with other inference results, you need to return to the original coordinates.

Also, the number of variables used in the coordinate transformation needs to be considered as the degrees of freedom.

Up until this point, we have made inductive inference that if the explanatory variables are close, the objective variable will also be close to a constant value.
When the objective variable is a function of the explanatory variables, we need to consider how to make inference.
For example, suppose the explanatory variable x and objective variable y have a relationship close to a straight line, "y=ax +b".
Therefore, we transform the coordinates so that "y'=y-ax" with an arbitrary slope a.
Since "y'=intercept b", it becomes a constant value.
If y' is the objective variable, we can make inductive inference that it will be constant.
No matter what function it is, you can make an inference by performing a coordinate transformation so that the objective value becomes a constant value.
If you want to compare the residual with other inference results, you need to return to the original coordinates.
Also, the number of variables used in the coordinate transformation needs to be considered as the degrees of freedom.

Intuitively good inference

So far, we have discussed valid "sample selection strategies".
From here, we will consider the superiority of "sample selection strategies".

Proposal for evaluation

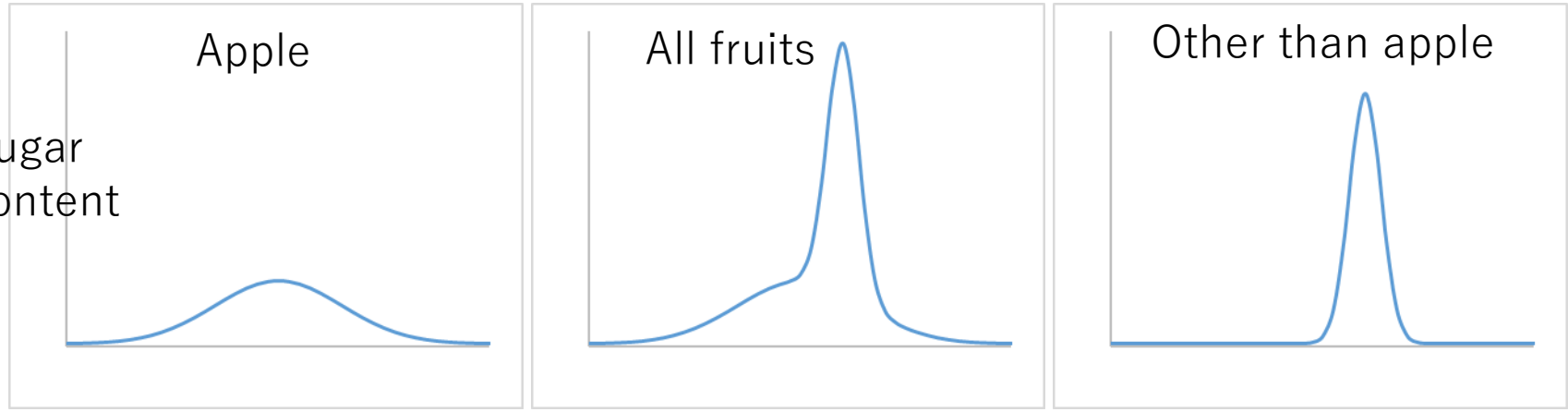
If the variance of the objective variable in the selected sample is small, then the inference is good?

However, it is a mistake to evaluate the superiority of an inference using variance.

So far, we have discussed valid "sample selection strategies".
From here, we will consider the superiority of "sample selection strategies".
Intuitively, we feel that if the variance of the objective variable of the selected sample is small, it is a good inference.
If you can make 100% guesses, the variance is 0.
A sharp distribution with small variance can be interpreted as meaning that the options have been narrowed down.
The closer the number of remaining options are, the closer you feel you are to a single optimal solution.
However, it is a mistake to evaluate the superiority of an inference using variance.

Superiority of inference and variance

Question: What's the sugar content of  ?



Variance:	Large	Small	Small
Bias:	Small	Large	Large

No matter how large the variance is, looking at just apples has smaller bias, so we felt that it was a good inference.

Let's consider the example of estimating the sugar content of an apple.
It would be a better inference to look at the sugar content distribution of apples only, rather than the distribution of sugar content of all fruits.
Now, suppose that the variance of sugar content of fruits other than apples is very small.
As a result, the variance of sugar content of all fruits will also be small.
Narrowing it down to just apples would actually increase the variance.
Why do we think it is better to look at just apples, even if the variance is large?
The quality of inference can be divided into precision (variance) and accuracy (bias).
In other words, no matter how large the variance is, looking at just apples has smaller bias, so we felt that it was a good inference.

The best limit of induction (1)

What is the theoretically best limit of inductive inference?

The samples used are in the best possible condition
in terms of both quality and quantity.

Question : What is the sugar content of an apple?

Best quality sample: An apple that is so similar to the apple being guessed

Is it best to have an infinite number of the best quality samples?

However, being able to count samples means that they can be distinguished by some difference.

Now, let's consider the theoretically best limit of inductive inference.
Since it is not deduction, there is no sample that contains the answer itself.
Consider the best situation in terms of both quality and quantity of samples to be used.
Intuitively, it would seem that the best quality sample would be one with an infinite number of them.
Consider the example of guessing the sugar content of an apple.
An apple that is so similar to the apple being guessed can be said to be of the best quality.
However, being able to count samples means that they can be distinguished by some difference.

The best limit of induction (2)

Sample quality ranking:

Is it really optimal to have a large number of samples that are tied for first place in similarity?

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 … ∞

You just don't notice the difference.

Sample quality ranking:

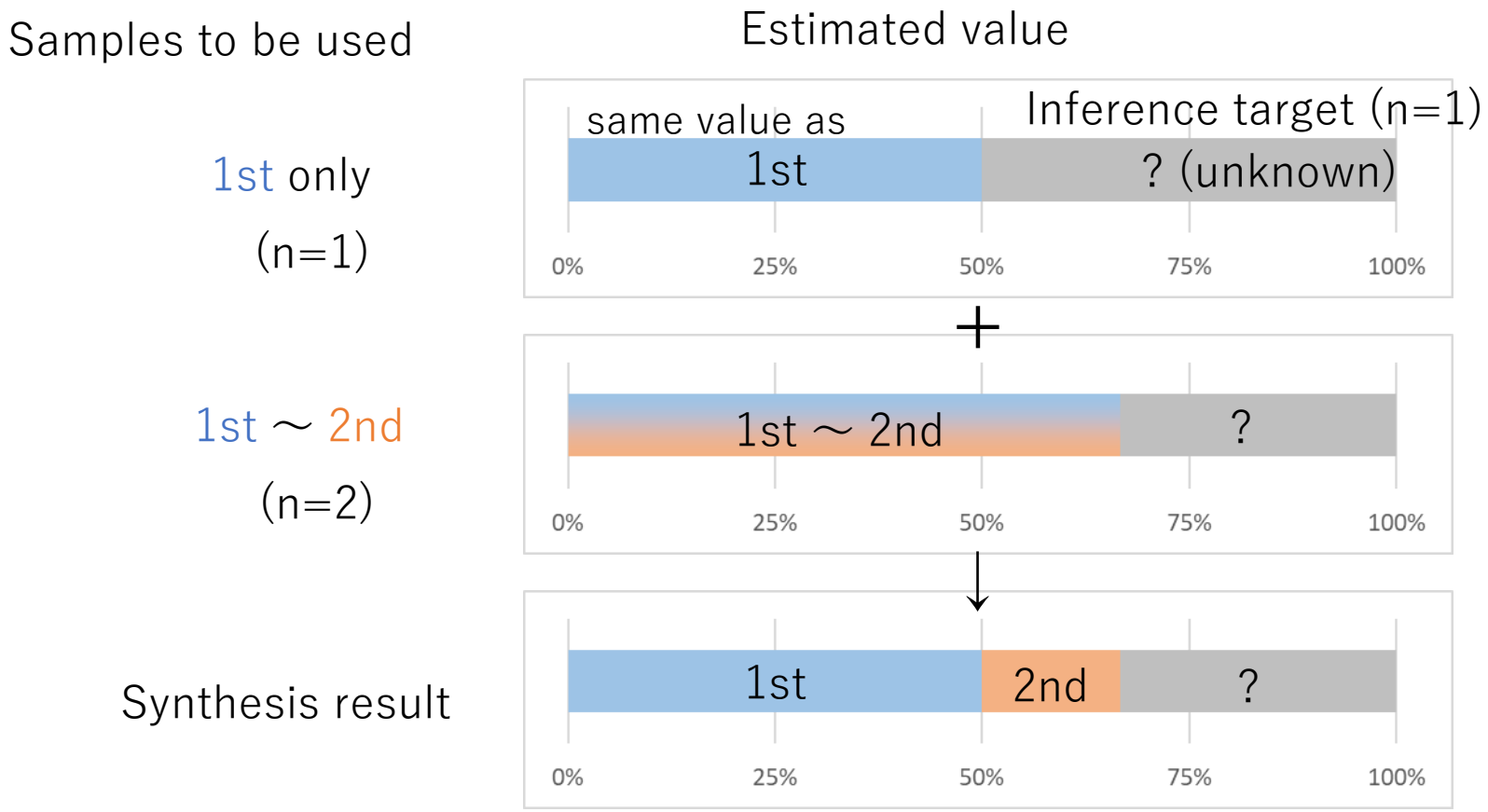
The best quality and quantity of samples

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 … ∞

The sample at nth place is clearly more similar to the sample being inferred than the sample at n+1th place.

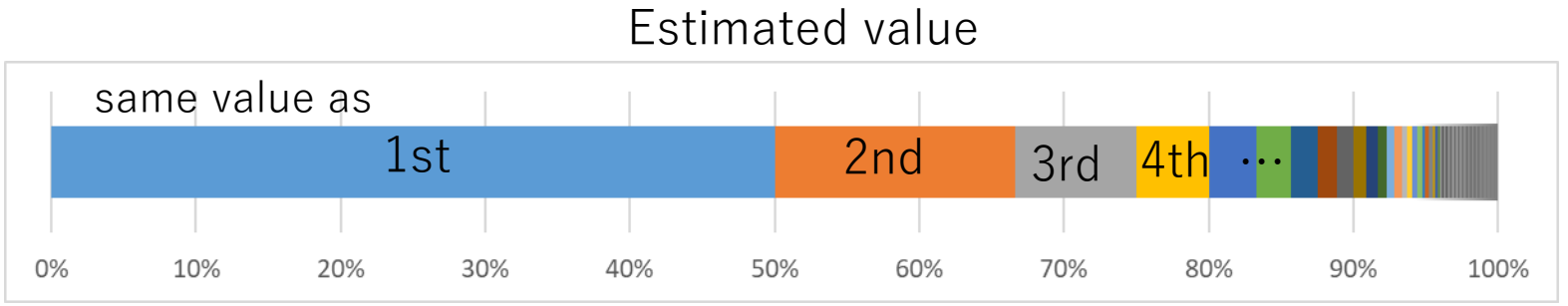
But is it really optimal to have a large number of samples that are tied for first place in similarity?
There may be some difference between those samples that are tied for first place, and we just haven't noticed.
If there is some difference, the ranking should be different.
For this reason, the optimal situation is when there are no samples with the same rank.
If there are n samples, they will be ranked from 1st to nth.
The sample at nth place is clearly more similar to the sample being inferred than the sample at n+1th place.
Having an infinite number of such samples is the best state in terms of sample quality and quantity.

The best limit of induction (3)



Let's consider what the estimated value would be when the quality and quantity of the samples are at their best.
Let's consider the case where inference is made using only the single first place sample.
Because one more sample to be inferred is added, 50% will be the same value as the first place sample, and the remaining 50% will be "unknown“.
Let's also consider the case where inference is made using two samples from second place or higher.
67% will be the same value as the two samples from second place or higher, and the remaining 33% will be "unknown“.
However, the first place sample is clearly a sample of better quality than the second place sample.
Therefore, of the 67%, 50% will be the same value as the first place sample, and the remaining 17% will be the same value as the second place sample.

Ideal induction



Proportion of samples in nth place = $\frac{1}{n^2+n}$

Total from 1st place to nth place = $\frac{n}{n+1}$

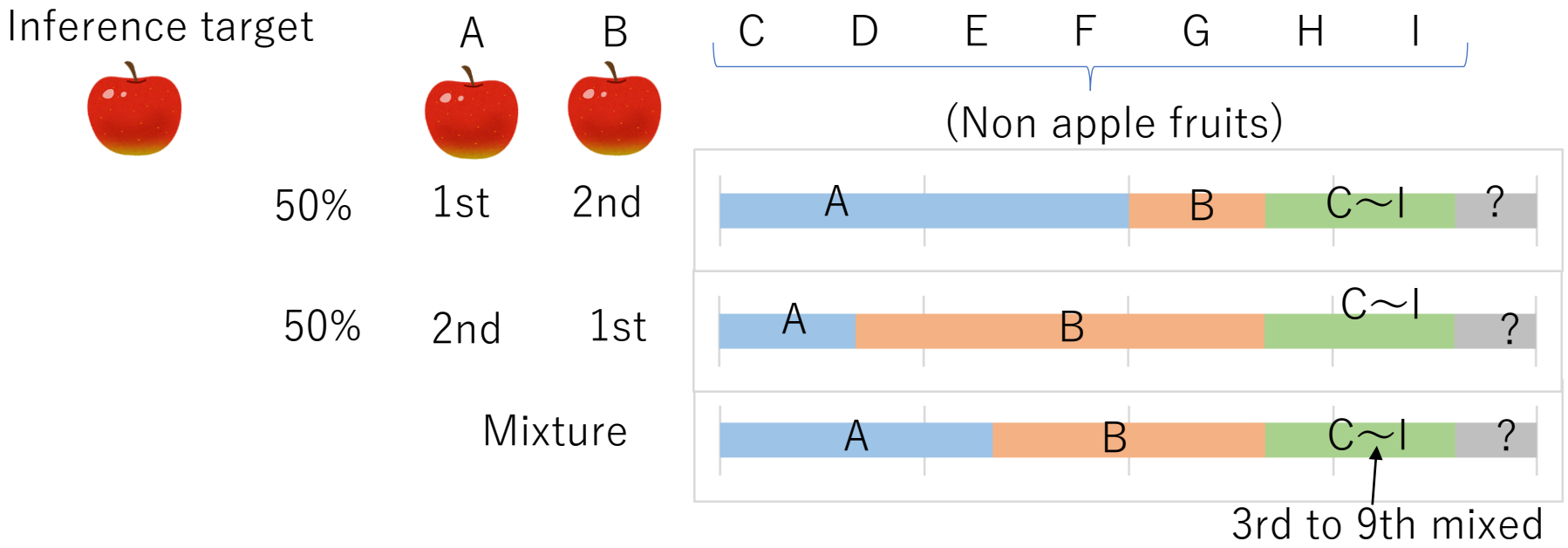
The remainder after subtracting the above formula from 1 = $1-\frac{n}{n+1}=\frac{1}{n+1}$: ? (Unknown)

The more samples you have, the fewer "unknowns" there are.

This state is called "ideal induction“.

Similarly, the calculations up to the nth place are combined.
The guess value, 1/(n^2+n), becomes the nth place sample and value.
The sum from 1st place to nth place is n/(n+1).
Subtracting this from 1, the remainder, 1/(n+1), becomes "unknown".
As you increase the number of samples starting from the top, the parts that were "unknown" will no longer be "unknown".
Once a part is no longer "unknown", it will not change.
"Unknown" is the worst possible value.
Since any sample will reduce the number of "unknowns", it is better than nothing, so it should not be ignored.
This state is called "ideal induction“.

Mixed rankings



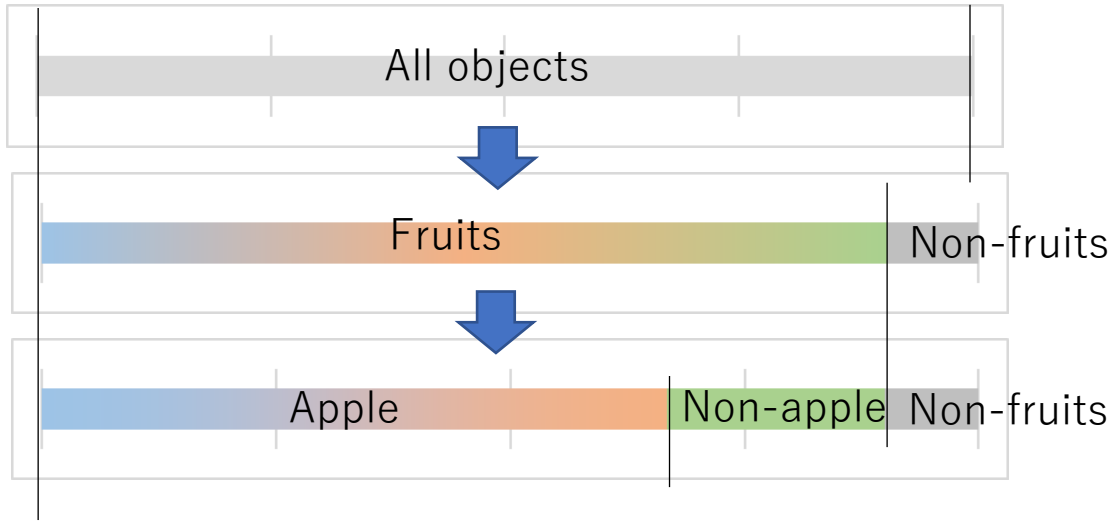
There are as many ideal induction as there are all combinations of ranks.

All inference results can be expressed as a mixture of ideal induction.

Let's consider an example of a non-best sample.
Suppose you want to predict the sugar content of an apple.
There are nine fruits as a sample, and two apples are selected from them.
The quality rankings of the two apples, A and B, are not tied for first place.
We can interpret this as a 50% mixture of "A is 1st and B is 2nd" and "A is 2nd and B is 1st".
The seven fruits other than apples can be considered to be a mixture of combinations of ranks from 3rd to 9th.
There are as many ideal induction as there are all combinations of ranks.
All inference results can be expressed as a mixture of ideal induction.

Step-by-step selection

Inference target



The selection of samples to use can be done in stages.

Deciding which samples to use is the same as deciding on a ranking.

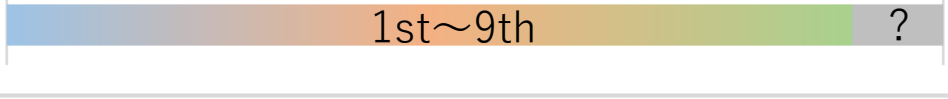
"Non-fruit" also has a low ranking.
All samples will always be used in inference, even if only a small amount.

The selection of samples to use can be done in stages.
For example, suppose you want to infer the sugar content of an apple.
First, divide "all objects" into "fruit" and "non-fruit".
Next, divide "fruit" into "apples" and "non-apples".
Gradually, it becomes clearer that all samples were equally ranked.
Deciding which samples to use is the same as deciding on a ranking.
"Non-fruit" also has a low ranking.
Therefore, all samples will always be used in inference, even if only a small amount.

Combination of rankings

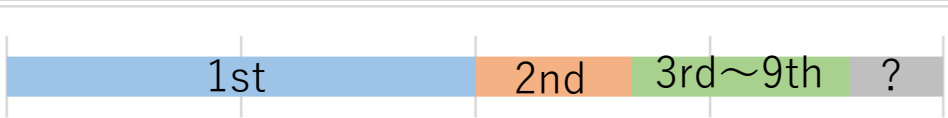
Combination of rankings

9!=362.880



2!×7!= 10,080

1×1×7!= 5,040



Proposal for evaluation

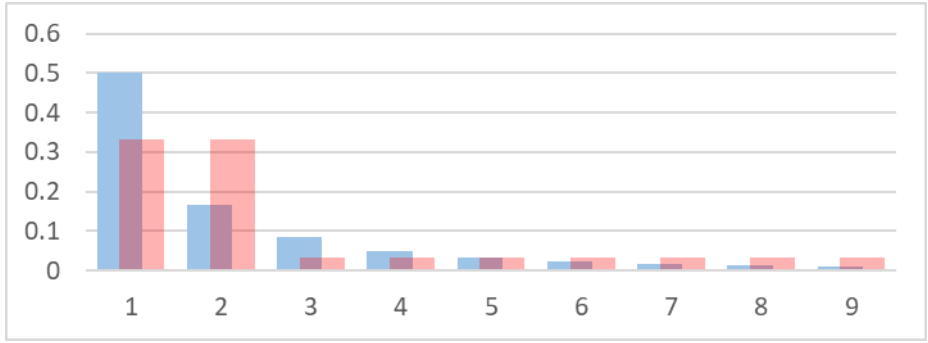
The smaller the rank combination, the better the inference?

The order of 1st and 2nd should be more important than the order of 8th and 9th, so it is a mistake to consider them to be of equal value.

Let's consider the number of combinations of rankings.
If we know even partially about the rankings, the number of combinations will decrease.
So, let's consider the number of combinations of rankings as a way to evaluate the quality of an inference.
However, the order of 1st and 2nd should be more important than the order of 8th and 9th.
As these would be considered to be of equal value, this method of evaluation is incorrect.

Ideal rate (1)

The proportion of each sample that accounts for the predicted result



Largest to Smallest

: Ideal induction
: An inference
: Matching part

Ideal rate: Sum of matching parts (0～1)

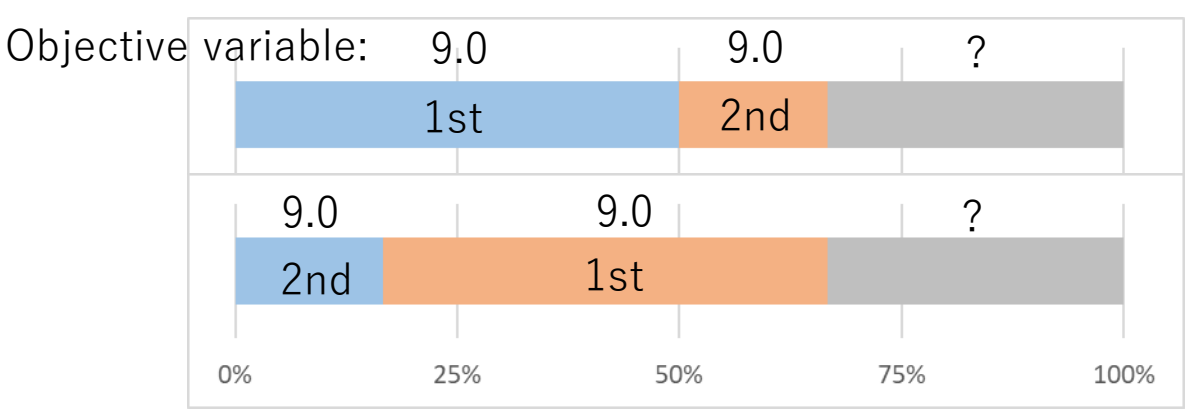
(The "unknown" part doesn't match)

Let's consider another way to evaluate the quality of an inference.
If it matches the "ideal induction," then it can be said to be the best inference.
Therefore, we will try to quantify the degree of agreement with the "ideal induction".
We arrange the inference results in order of the proportion that each sample makes up.
For each sample, we examine the parts where that proportion matches the ideal inference.
The total for all samples is the "ideal rate".
Since we don't know whether the "unknown" parts match or not, we will consider them to not match.

Ideal rate (2)

Proposal for evaluation

The larger the "ideal rate", the better the inference?



The distribution of the objective variable is the same for both

There is no need to decide the ranking

When the objective variable for all samples is a constant,
the "ideal rate" is low, but a good inference can be made that it will be a constant.

It is a mistake to evaluate the quality of an inference using the "ideal rate".

It is necessary to consider not only the rank of the sample
but also the variation in the values of the objective variable.

Let's consider the magnitude of the "ideal rate" as a possible way to evaluate the quality of an inference.
When the order of first and second place is clear, the "ideal rate" will be larger than when it is not.
However, suppose the first and second place samples have the exact same objective variable.
In that case, making the order of first and second place clear will not change the inference result.
As an extreme example, suppose that after coordinate transformation, the objective variable of all samples becomes a constant.
In that case, there is no need to consider which sample to use, the estimate will be that constant.
If all samples are used equally, the "ideal rate" will be the lowest.
Even if the "ideal rate" is low, excellent inference can be made, so it is a mistake to use the "ideal rate" to evaluate the quality of an inference.
It is necessary to consider not only the rank of the sample but also the variation in the values of the objective variable.

Quantifying variation

"Unknown" and "outliers" in the objective variable cannot be ignored

$$Average = \bar{x}$$

If there is even the slightest bit of "unknown", it cannot be determined

$$Standard\ deviation = \frac{1}{n-1}\sqrt{\sum (x-\bar{x})^2}$$

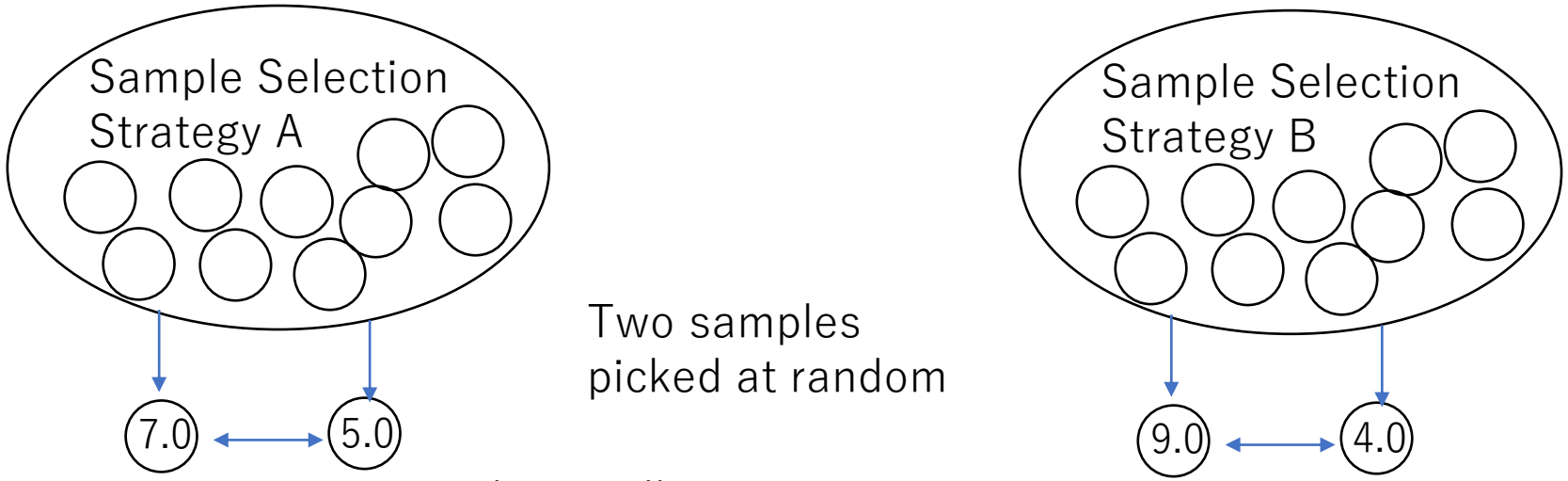
Outliers dominate

$$Mean\ absolute\ deviation = \frac{1}{n-1}\sum |x-\bar{x}|^1$$

If difference is orders of magnitude, the outliers dominate

It is difficult to quantify the variance of the objective variable.
This is because it includes samples with "unknown" values.
In addition to "unknown", there are also samples with extreme outliers.
Even outlier samples are better than nothing, so they should not be ignored.
When calculating standard deviation, the residual is squared, which increases the influence of outliers.
In the case of mean absolute deviation, the influence of outliers is appropriate as it is the expected value of the distance.
However, if there are outliers of a different order of magnitude, they will dominate.
Furthermore, if even a small number of "unknown" values are included, the average value cannot be determined.

Distance Win Rate



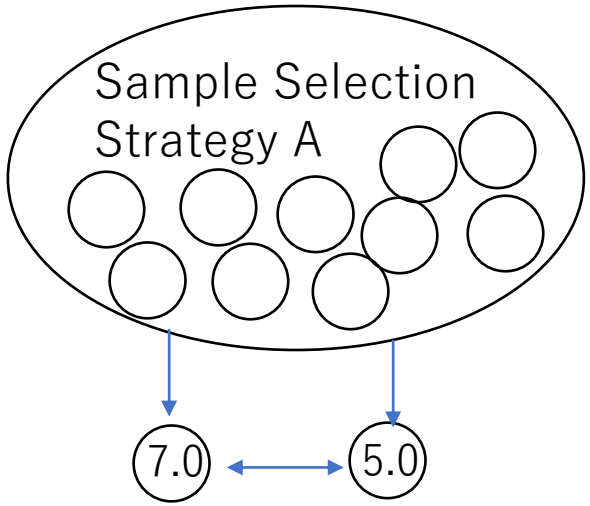
The smaller one wins.

If the value is "unknown", you lose,
if both values are "unknown", it's a draw.

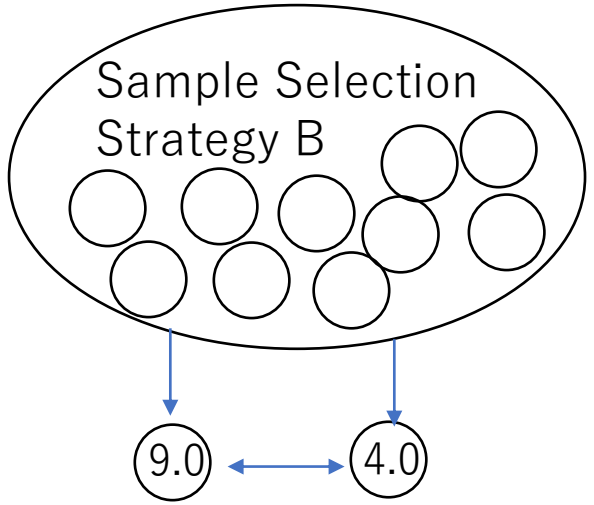
Repeat to determine the winning percentage.
If you want to eliminate randomness, try all combinations.

Let's consider a method that can evaluate the variance of the objective variable even if it contains "unknowns" or "outliers".
If you just want to know which of two "sample selection strategies" is superior, it is sufficient to know just the superiority of the variance.
Randomly draw two samples from each of the two strategies.
For each, find the distance between the objective variable of the two samples.
Compare which of the two strategies has the larger distance.
The smaller distance is the winner, and repeat this process many times to determine the winning rate.
If the value is "unknown", you lose, and if both values are "unknown", it's a draw.
If you want to eliminate randomness, examine all combinations.

Non-ideal Distance Win Rate



Two samples
picked at random



Distance
×(1-"Ideal rate")

The smaller one wins.

If the value is "unknown", you lose,
if both values are "unknown", it's a draw.

Repeat to determine the winning percentage.
If you want to eliminate randomness, try all combinations.

We will consider an evaluation method that includes not only "variance" but also the "ideal rate".
Extract two samples from each of the two "sample selection strategies".
For each, find the distance between the two samples and multiply it by "1 - ideal rate".
Compare which of the two strategies has the larger distance, multiplied by "1 - ideal rate".
The rest is the same as the previous example.
The "ideal rate" can be calculated quickly if you look at only the extracted samples, rather than all samples.

Strategy Evaluation Method (1)

Weight ranking: n (Sort all samples in order of largest weight)

Sample weight: W_n

Ideal weight: $I_n = \frac{1}{n^2 + n}$

Ideal rate: $R_n = \frac{2 \times Min(W_n, I_n)}{W_n + I_n}$

Objective variable: V_n

Non-ideal distance: $D_{ij} = |V_i - V_j| \times \left(1 - \frac{R_i + R_j}{2}\right)$

(Randomly draw samples for the i and j positions)

We will explain the steps for evaluating a "sample selection strategy".
First, sort all samples in order of largest weight.
Let Wn be the weight of the sample at rank n.
The weight In of the ideal induction for rank n is "1/(n^2+n)".
The ideal rate Rn for rank n is 2×Min(Wn, In)/(Wn +In).
Let Vn be the value of the objective variable for the sample at rank n.
Randomly draw samples for the i and j positions.
The non-ideal distance Dij is "|Vi -Vj|{1-(Ri +Rj)/2}".

Strategy Evaluation Method (2)

Non-ideal distance: $D_{ij} = |V_i - V_j| \times \left(1 - \frac{R_i + R_j}{2}\right)$

Compare the non-ideal distances for the two strategies.

The smaller one wins,
and we repeat the process many times to determine the winning rate.

If the value is "unknown", you lose,
if both values are "unknown", it's a draw.

If you want to eliminate randomness, try all combinations.

Compare the non-ideal distances for the two strategies.
The smaller one wins, and we repeat the process many times to determine the winning rate.
If the value is "unknown", we lose, and if both values are "unknown", it's a draw.
If we want to eliminate randomness, we check all combinations.

Interpretation of non-ideal distance

Non-ideal distance: $D_{ij} = |V_i - V_j| \times \left(1 - \frac{R_i + R_j}{2}\right)$

Variance　　Bias

If either the bias or variance is close to 0, it will be the best inference, even if the other is large.

Example

This is the case when the residuals of the objective variable for all samples become 0 as a result of coordinate transformation.

Variance $\fallingdotseq 0$

The questioner has given you a ranking of the samples, so you use that ranking as is.

Bias $\fallingdotseq 0$

The two terms of non-ideal distance can be interpreted as corresponding to bias and variance.
If either the bias or variance is close to 0, it will be the best inference, even if the other is large.
For example, this is the case when the residuals of the objective variable for all samples become 0 as a result of coordinate transformation.
This is the worst case scenario, as it is not possible to rank the samples.
However, since the variance is 0, this is the best inference, no matter how bad the bias is.
In another example, the questioner has given you a ranking of the samples, so you use that ranking as is.
This is an ideal inductive inference with a bias of 0, no matter how large the variance is.

Blind Inference

Non-ideal distance: $D_{ij} = \underbrace{|V_i - V_j|}_{\text{Variance}} \times \left(1 - \frac{R_i + R_j}{2}\right)$ (Bias)

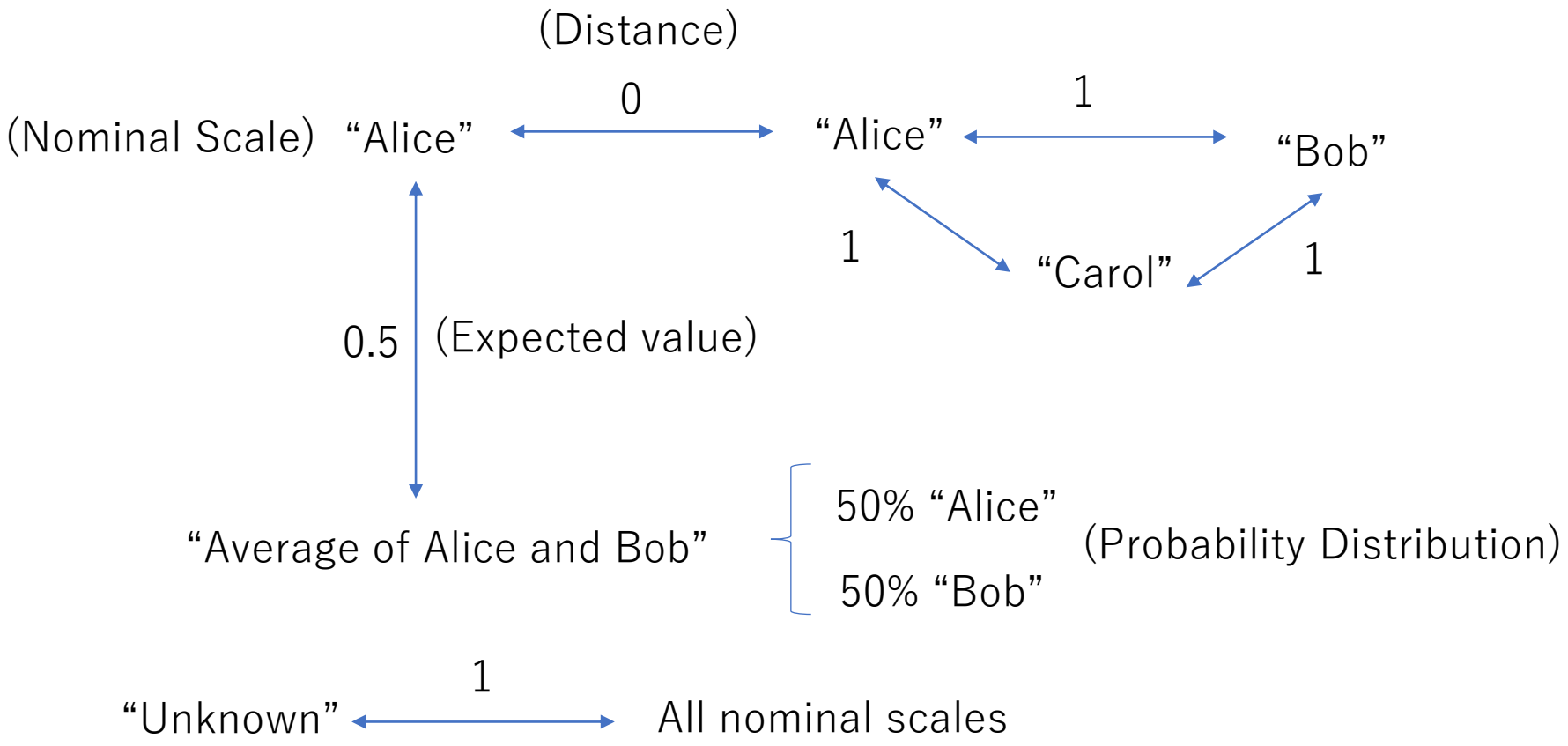
If one of the bias and variance is unknown,
they can still be compared by assuming them to be constant.

You can make inferences without knowing variance.

Even if you don't know the value of the objective variable,
you can claim that the value of the objective variable for a particular sample is close to the value of the objective variable you are trying to estimate.

Even if it is not possible to calculate non-ideal distance, it may be possible to evaluate the superiority of a strategy.
There are terms for bias and variance, but even if one of them is unknown, they can be compared by assuming it to be a constant.
Not needing to know the variance means that inferences can be made without knowing any of the values of the objective variable.
For example, when information on the ranking of samples is given.
Suppose you are given the information that "the value of the objective variable of sample A is closest to the value of the objective variable to be inferred".
In that case, you can make an assertion that is in line with that information, even if you do not know the value.
The same is true for examples where information on effective explanatory variables is given.
You can assert that samples with similar values for the explanatory variable also have similar objective variables.

Nominal Scale Distance



A further note about the distance of the objective variable.
The objective variable can be a nominal scale, rather than a numerical one.
For example, guesses for a person's name would be values such as Alice, Bob, Carol, etc.
If the names are the same, it is reasonable to set the distance to 0, and if the names are different, the distance to 1.
The method of looking at the Euclidean distance, where each name dimension has a value of 0 or 1, is poorly founded.
Also, the "average of Alice and Bob" can be thought of as a probability distribution of 50% Alice + 50% Bob.
The distance between "Alice" and the "average of Alice and Bob" will be the expected value of 0.5.
We also assume that the distance between "unknown" and all nominal scales is 1, which is the worst case scenario.

Layers of inference

			Information given
Layer	1	Deduction	Answer information
	2	Induction	Rank or weight of the sample
	3		Effective explanatory variables
	4		Value of the objective variable

Estimate by Rank or weight

Estimate by explanatory variables

Estimate by objective variable

Inference is divided into four layers depending on the information given.
The first layer is when deduction is possible.
The second layer is when the rankings or weights of the samples are given.
Applying the given rankings or weights as is results in optimal inference.
The third layer is when effective explanatory variables are given.
The rankings are estimated from those explanatory variables.
The fourth layer is when no effective variables are given.
In that case, effective explanatory variables are estimated from the objective variable.

Blurred separation

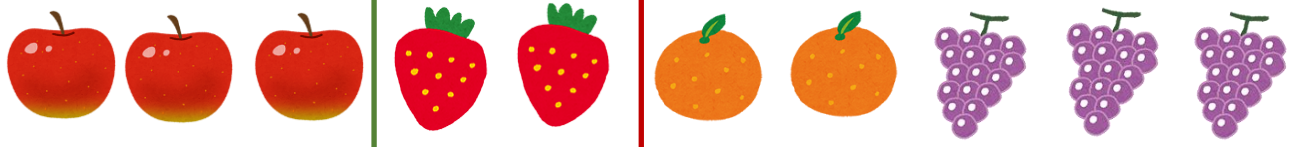
Classified by variety

Clear ranking separation is only possible
if you know that the explanatory variables are effective.

1st～2nd

3rd～9th

Sugar content: ? 9.0 9.2 3.4 3.6 5.5 5.7 4.7 4.9 4.5



~~1st～4th~~

If the correlation is small,
the sugar content will not be clearly separated.

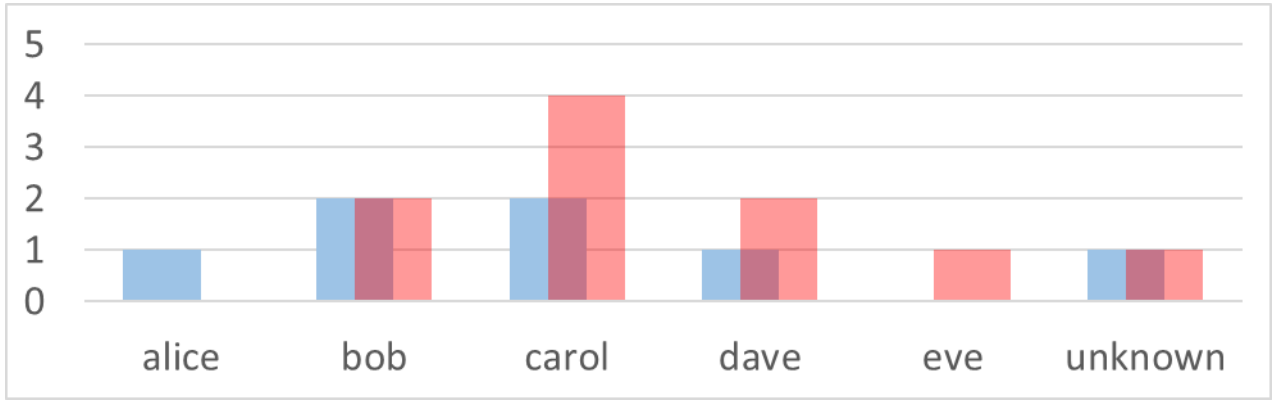
Classified by color

It is necessary to estimate effective explanatory variables
based on their relationship with the objective variable.

In the apple example, fruit of the same variety were ranked 1st to 2nd, and those of different varieties were ranked 3rd to 9th.
The rankings of the two groups are clearly separated into those above a certain value and those below a certain value.
This can only be said if you know that the objective variable, sugar content, is determined by the variety.
For example, you could ignore variety and separate the groups by color alone.
If the correlation between color and sugar content is low, the rankings of sugar content would not be clearly separated.
It is rare that effective explanatory variables are given as information.
Effective explanatory variables need to be estimated from their relationship with the objective variable.

Match rate (nominal)

Sample size frequency



Objective variable

Group A

Group B

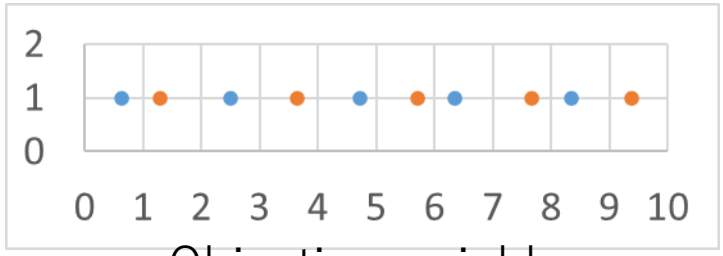
Match rate $= \frac{Matcing\ sample\ size}{Total\ sample\ size}$ (0～1)

The distribution of the objective variable values and frequencies is shown in the graph.
Dividing into two groups by the explanatory variables does not necessarily separate the objective variable.
The degree to which the distributions of the two groups overlap is called the “match rate“.
If the scale is nominal, it can be calculated from the sample size that matches by name.
The denominator is the total sample size, and the numerator is the sample size with agreement.
It takes a value between 0 and 1.

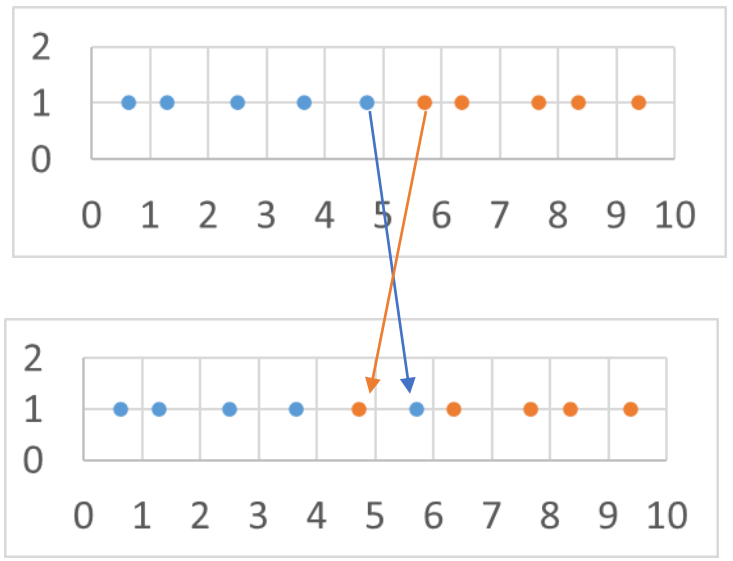
Match rate (continuous)

It is not permissible to assume a particular distribution

Alternating: Match rate ≒ 1



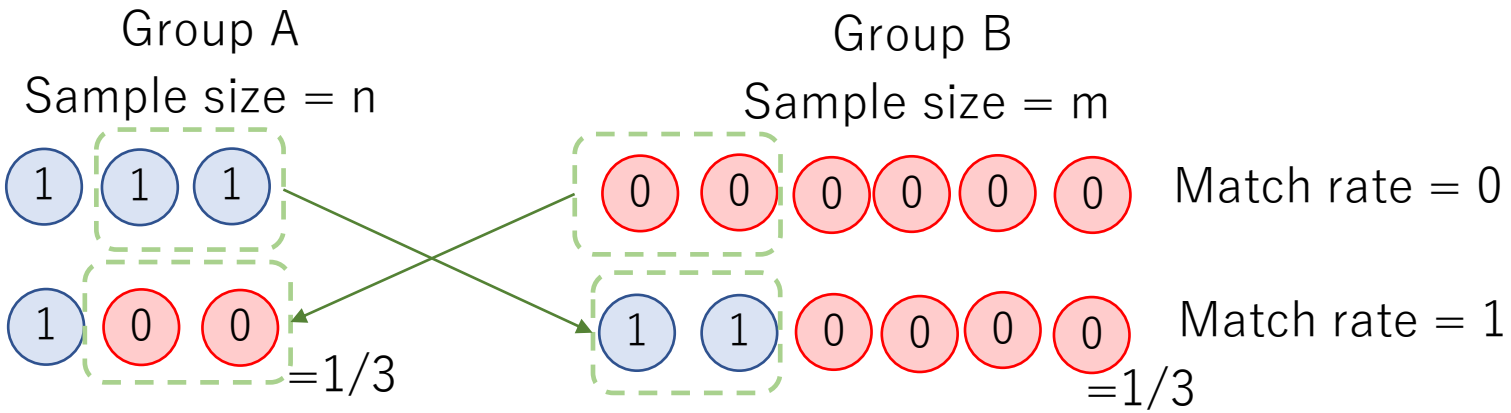
Separated: Match rate ≒ 0



If we assume that they were originally separated
and count the number of swapped pieces,
we can find the match rate.

Let's consider the case where the objective variable is a continuous value.
If the precision is poor, it will simply be impossible to distinguish, but the numbers will not exactly match.
Since the shape of the distribution is unknown, it is not permissible to assume that it is normal.
The match rate can also be calculated using nonparametric methods that do not assume the shape of the distribution.
Intuitively, if the two groups of numbers are arranged alternately, the match rate is 1.
Conversely, if the orders are completely separate, the match rate is 0.
When the match rate is not 0, it can be interpreted that some of the samples that were originally separated have been swapped.
Therefore, the match rate can be found by counting the number of swapped samples.

Swapping count

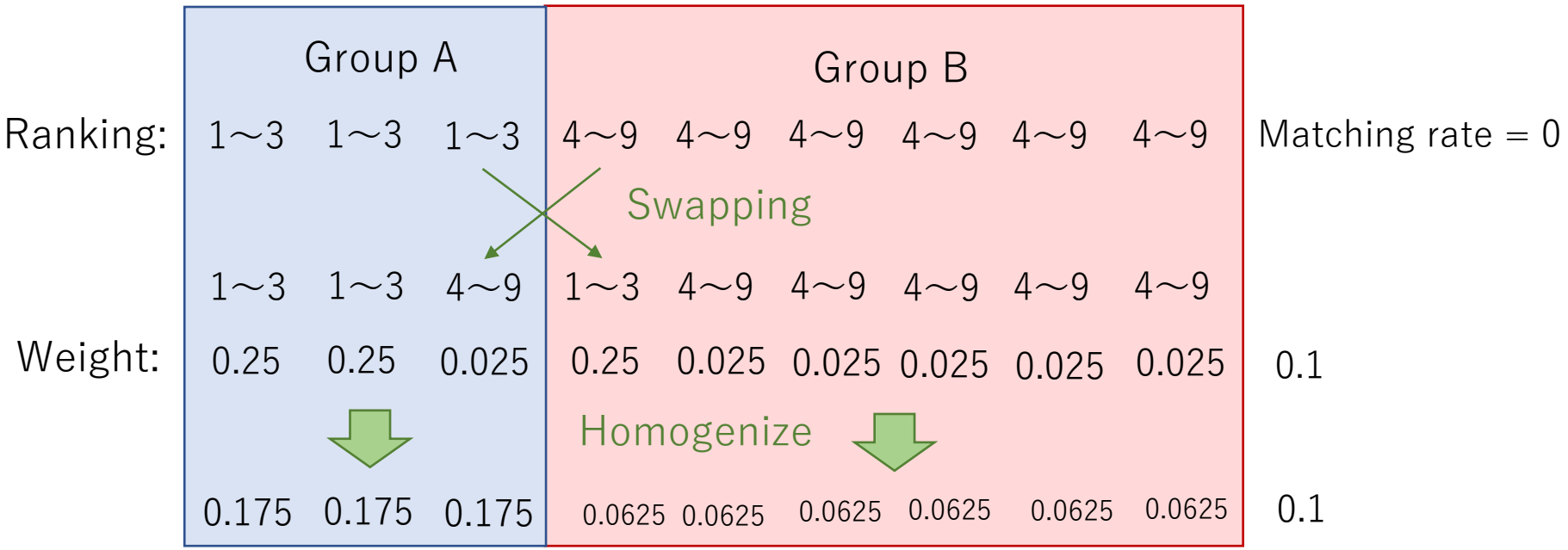


$$\text{Count of swapping} = \frac{1}{\frac{1}{n}+\frac{1}{m}} \qquad \text{Sample size of swapping} = \frac{2}{\frac{1}{n}+\frac{1}{m}}$$

$$\text{Match rate} = \frac{\text{Sample size of swapping}}{\frac{\frac{1}{n}+\frac{1}{m}}{2}} \qquad \text{Sample size of swapping} = \text{Match rate} \times \frac{2}{\frac{1}{n}+\frac{1}{m}}$$

We will determine the relationship between the distribution match rate and the number of sample exchanges.
Suppose initially, group A has n samples with a value of "1", and group B has m samples with a value of "0".
To make the value distributions of the two groups match, it is necessary to exchange samples 1/(1/n+1/m) times.
Two samples are exchanged per exchange.
Therefore, to change the match rate from 0 to 1, it is necessary to exchange 2/(1/n+1/m) samples.
The match rate between two distributions and the number of sample exchanges can be converted using the formula shown.

Unclear separation weights



It becomes homogenous because it is impossible
to distinguish which ones have been replaced.

Once the sample weights are known, the non-ideal distances can be calculated.

Now, let's consider how to rank samples when the separation of the target variable is unclear.
The number of sample exchanges can be found by counting the number or by converting from the match rate.
As an example, suppose the samples are divided into groups A and B, containing 3 and 6 samples, respectively.
Before the exchange, the samples were separated, so they were divided into "1st to 3rd place" and "4th to 9th place".
As the match rate of the distribution was 0.5, we converted it to mean that 2 samples were exchanged.
Once the rankings are known, they can be converted into sample weights.
The sample weights of group A homogenize the weights of samples within the group.
This is because we only know the samples after exchange and cannot distinguish which ones have been exchanged.
Once the sample weights are known, the non-ideality rate can be calculated, allowing us to evaluate the superiority or inferiority of the sample selection strategy.

Purpose of the statistics

	Descriptive statistics	Inferential statistics	Statistics 2.0
Purpose	Summary statistic (Average, variance, etc.)	Population (Normal distribution, etc.)	Unknown sample value

Traditional statistics

If there is only one sample,
it is impossible to make a reasonable estimate.

Statistics 2.0

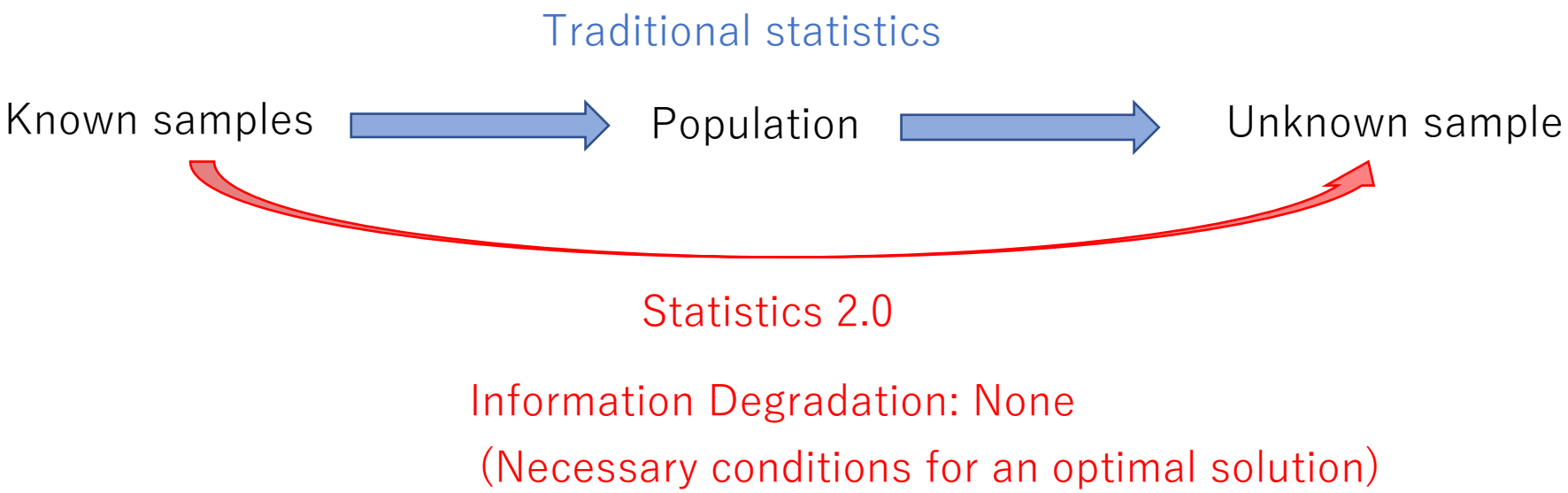
Even from just a single sample,
it is possible to make inferences
by making the most of the information.

So far, we have explained the basic criteria for evaluating sample selection strategies.
Next, we will explain how to determine estimates for "unknown" samples using the selected sample.
To begin with, traditional statistics does not aim to infer values of unknown samples.
It estimates the population by assuming normal distribution, etc.
It also calculates summary statistics such as the mean and variance.
It just means that these can sometimes be applied to infer values for "unknown" samples.
For example, if there is only one sample, it is impossible to make a reasonable estimate.
On the other hand, Statistics 2.0 aims to infer values for "unknown" samples.
Even from just a single sample, it is possible to make inferences by making the most of the information.

Information in statistics

Added information: Normal distribution, Prior probabilities, etc.

Deleted information: Individual sample values



Information Degradation: None

(Necessary conditions for an optimal solution)

Let's think about information in statistics.
In traditional statistics, normal distributions and other assumptions are made when estimating a population.
In other words, the information that it is normal distribution is arbitrarily added.
In the case of Bayesian statistics, prior information is arbitrarily added.
Information on the mean and variance is left, and information on the individual values of the sample is lost.
On the other hand, Statistics 2.0 directly infers unknown samples from known samples.
As a result, information does not degrade.
Not degrading information is a necessary condition for an optimal solution.

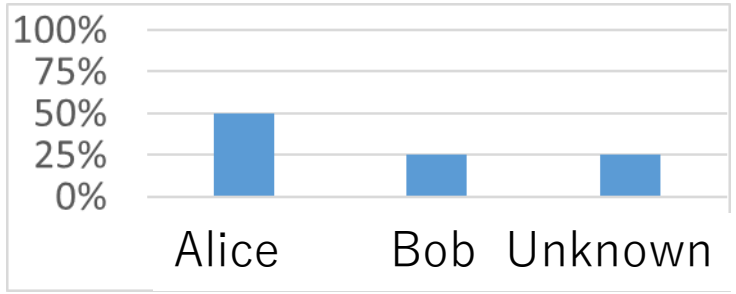
Raw objective variable

Objective variable: "Alice, Alice, Bob, Unknown“ (Nominal scale)

Frequency

=

Guess

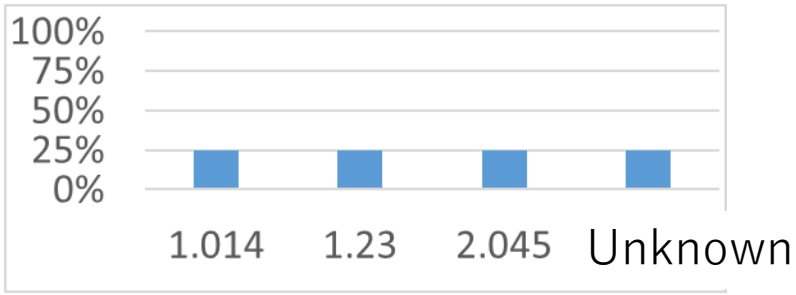


Objective variable : “1.014, 1.023, 2.045, Unknown” (Continuous value)

Frequency

≠

Guess



With continuous values, the distribution of the objective variable of the sample cannot be used as the guess.

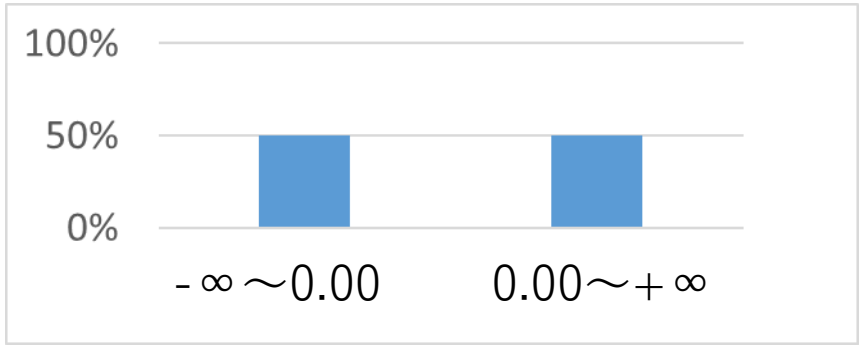
Let us consider how to directly infer the objective variable.
For example, suppose the objective variable of the sample was a nominal scale of "Alice, Alice, Bob, Unknown".
The guesses would simply be "Alice" = 50%, "Bob" = 25%, and "Unknown" = 25%.
In another example, suppose the objective variable was continuous values of "1.014, 1.023, 2.045, Unknown".
If we use the same approach as for the nominal scale, each would be 25%.
However, since these are continuous values, it should be rare for them to match exactly down to the decimal point.
With continuous values, the distribution of the objective variable of the sample cannot be used as the guess.

Inference from a single sample

Objective variable : “0.0, Unknown” (Continuous value)

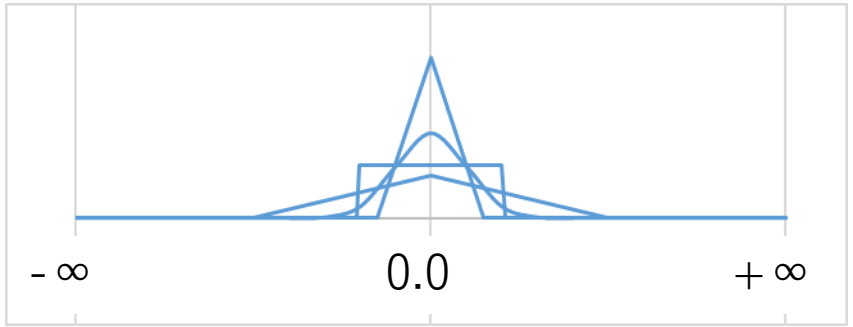
Interval estimation is possible.

There is no uniform distribution within the interval.



The closer to "0.0", the greater the probability density is estimated to be.

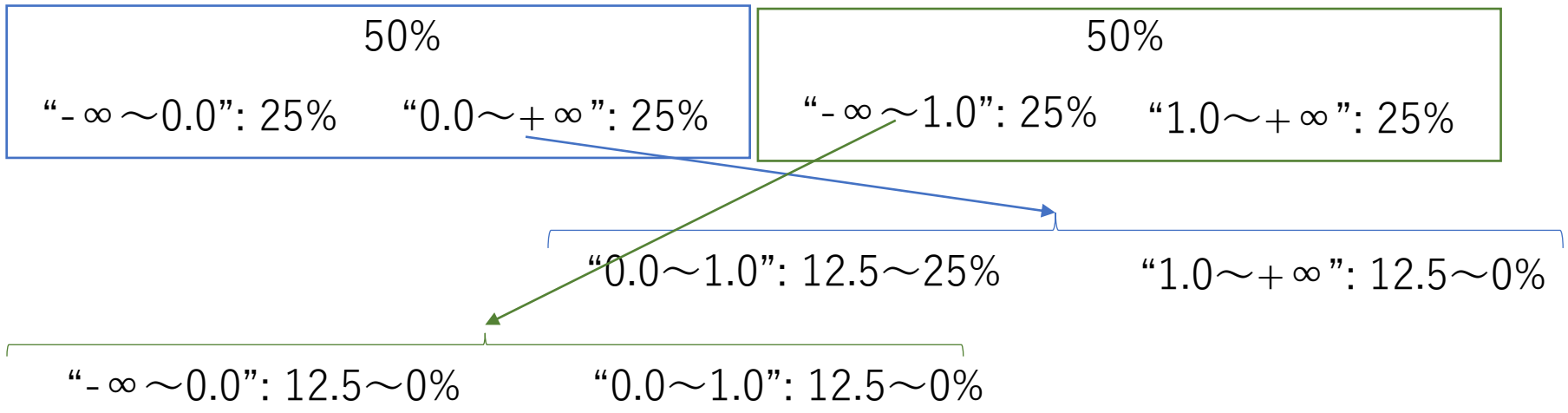
The shape of the distribution is unknown.



Let's consider the case where there is only one known sample.
Suppose the objective variable is a continuous value and is "0.0".
It is not possible to infer that the value of the unknown sample will exactly match "0.0".
However, we can infer that the closer the value is to "0.0", the higher the probability density.
However, we have no hints as to how much higher the probability density will be the closer it is to "0.0".
Also, since we have no hints as to whether it is greater or smaller than "0.0", we can guess that each is 50%.
We can make interval estimates, such as "-∞ to 0.0" = 50%, and "0.0 to +∞" = 50%.
However, we cannot assume that it is uniformly distributed within this interval.
Although we do not know the shape of the distribution, we can infer that the closer it is to "0.0", the higher the probability density.

Inference from two samples (1)

Objective variable : “0.0, 1.0, Unknown” (Continuous value)



Homogeneous ～ Biased towards the closer side

Total results

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

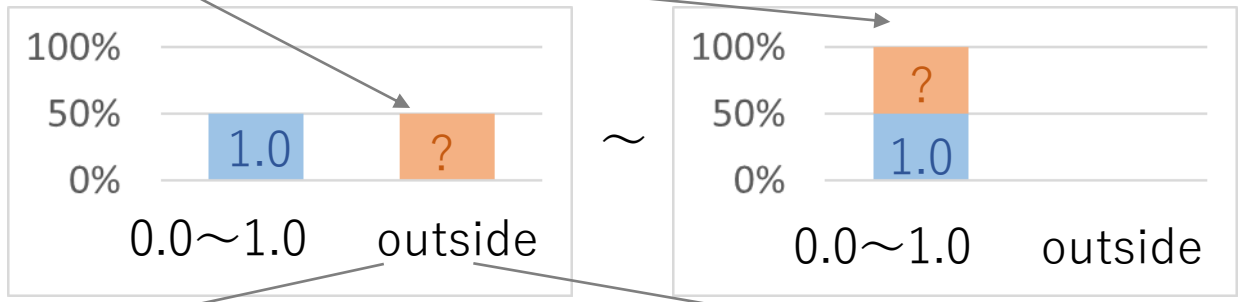
Let's also consider the case where there are two known samples, "0.0" and "1.0".
We will combine the results from the case where there is one sample, at 50% each.
The four intervals "-∞ to 0.0", "0.0 to +∞", "-∞ to 1.0", and "1.0 to +∞" will each have 25%.
We will divide "0.0 to +∞" = 25% into "0.0 to 1.0" and "1.0 to +∞".
Since there are no hints, if we consider them equally, each will be 12.5%.
However, since the probability density is higher closer to 0.0, in extreme cases it will be 25% and 0%.
"0.0 to +∞" = 25% was divided into "0.0 to 1.0" = 12.5 to 25% and "1.0 to +∞" = 0 to 12.5%.
We will divide "-∞ to 1.0" in the same way.
The total results were "-∞ to 0.0" = 25 to 37.5%, "0.0 to 1.0" = 25 to 50%, and "1.0 to +∞" = 25 to 37.5%.

Inference from two samples (2)

Objective variable : “0.0, 1.0, ?” (Continuous value)

Standard

Frequency



～

Average results

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

Previous page results

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

Two results are met

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

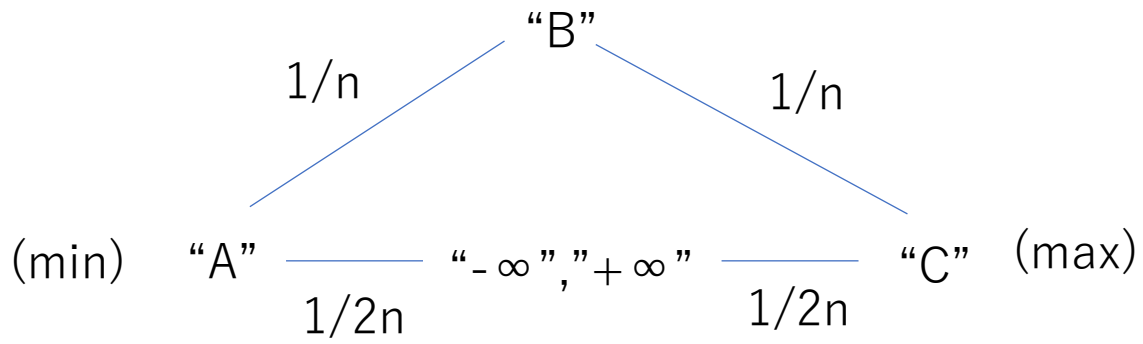
Using the same example, let's also consider the frequency with which samples exist in a certain interval.
Using the first sample as a base, we look at the intervals in which the second and third samples, which we are predicting, exist.
The value of the second sample is "1.0", so it exists 100% in the "0.0 to 1.0" interval.
The third sample is unknown, so it exists 0 to 100% in the "0.0 to 1.0" interval.
Taking the average of the two, 50 to 100% exists in the "0.0 to 1.0" interval.
The remaining 0% to 50% is divided evenly between "-∞ to 0.0" and "1.0 to +∞".
As a result, "-∞ to 0.0" = 0 to 25%, "0.1 to 1.0" = 50 to 100%, "1.0 to +∞" = 0 to 25%.
If we also try to satisfy the results on the previous page, we end up with "-∞ to 0.0" = 25%, "0.1 to 1.0" = 50%, and "1.0 to +∞" = 25%.

Inference from n samples

Objective variable : “A, B, C, ?” (Continuous value)

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

Knows samples : n



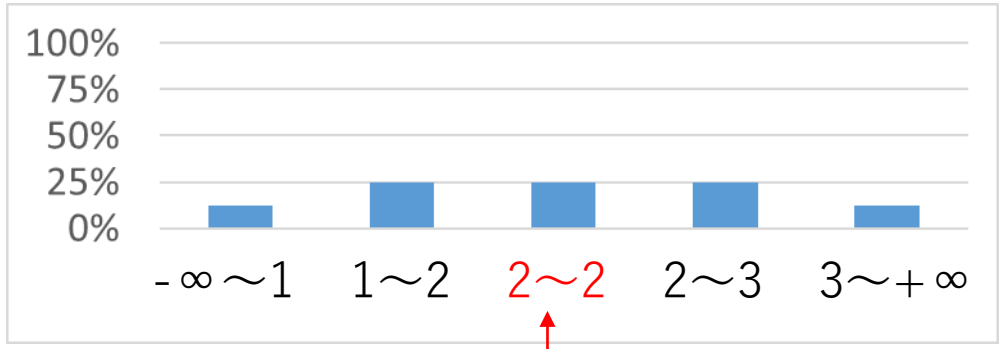
The guess value is 1/n for each of the n intervals of the circle connecting the minimum and maximum values.

(Divide the connected ends in half.)

Let's do a similar calculation if there are three known samples, "A, B, C, ?".
"-∞ to A" = 1/6, "A to B" = 1/3, "B to C" = 1/3, "C to +∞" = 1/6.
This can be interpreted as being evenly divided into three intervals: "A to B", "B to C", and "C to +∞, -∞ to A".
If there are four or more samples, the maximum and minimum values can be connected to form a circle, and all intervals will be equal.
Only the part where the ends are connected will be divided in half.
If there are n samples, there will be n intervals, each of which will be 1/n.

Duplicate Value Inference

Objective variable : “1, 2, 2, 3, ?”



There is a 1/4 chance that it will match the number 2 exactly.

「1」 Standard

「1, 2」

「1, 2, 2」

「1, 2, 2, 3」

「1, 2, 2, 3, ?」

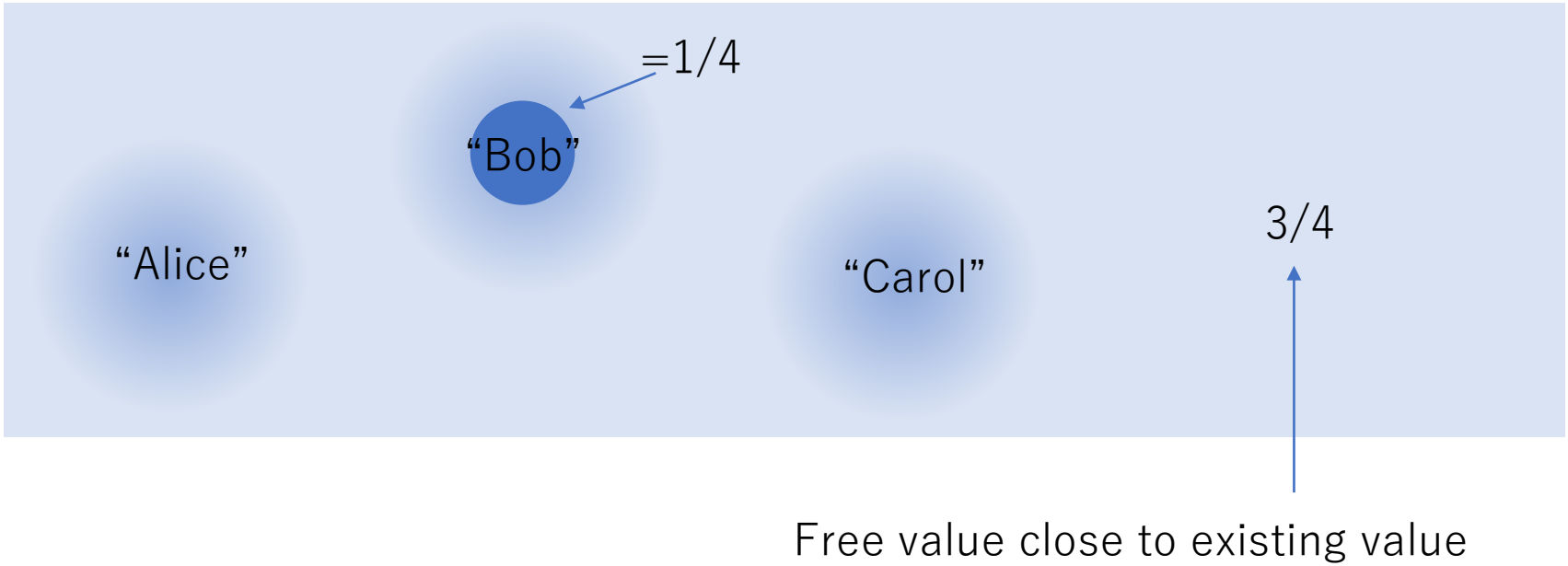
Matches the proportion of known values that occurred
(≧1/4)

Let's consider the case where values in a known sample are duplicated.
We will calculate the cases where the values are "1, 2, 2, 3, ?".
"-∞ to 1" = 1/8, "1 to 2" = 1/4, "2 to 2" = 1/4, "2 to 3" = 1/4, "3 to +∞" = 1/8.
Here, "2 to 2" = 1/4 is assumed to exactly match "2".
To verify, we will count the frequency at which values that have already appeared re-appear.
In the first sample, it does not re-appear, so we will consider it the base value and exclude it, and count it.
In the second to fifth samples, it re-appears at least once in four times.
The duplicate frequency of the estimated values matched the duplicate frequency of the known samples.

Nominal inference

Objective variable : “Alice, Bob, Bob, Carol, ?”

“～Alice”=1/4　　“～Bob”=1/4　　“Bob～Bob”=1/4　　“～Carol”=1/4



Let's consider the case where the known sample is on a nominal scale.
Suppose the values are "Alice, Bob, Bob, Carol, ?".
"Bob ~ Bob" = 1/4, which is the same as when the values are discrete.
The remaining 3/4 cannot be arranged in ascending order because it is on a nominal scale.
As an "unknown" sample, there is a possibility that unknown values will appear.
There is also a possibility that existing values will reappear.
Therefore, the 3/4 part can be considered free values that are close to the existing values.

Inference from 0 sample

Objective variable : “?”

4-choice question　　Each option=25%

Objective variable≧0　　Uniform distribution of “0～+∞”

The answer options determined by the questioner can be considered as a sample.

Question: What is the sugar content of an apple?

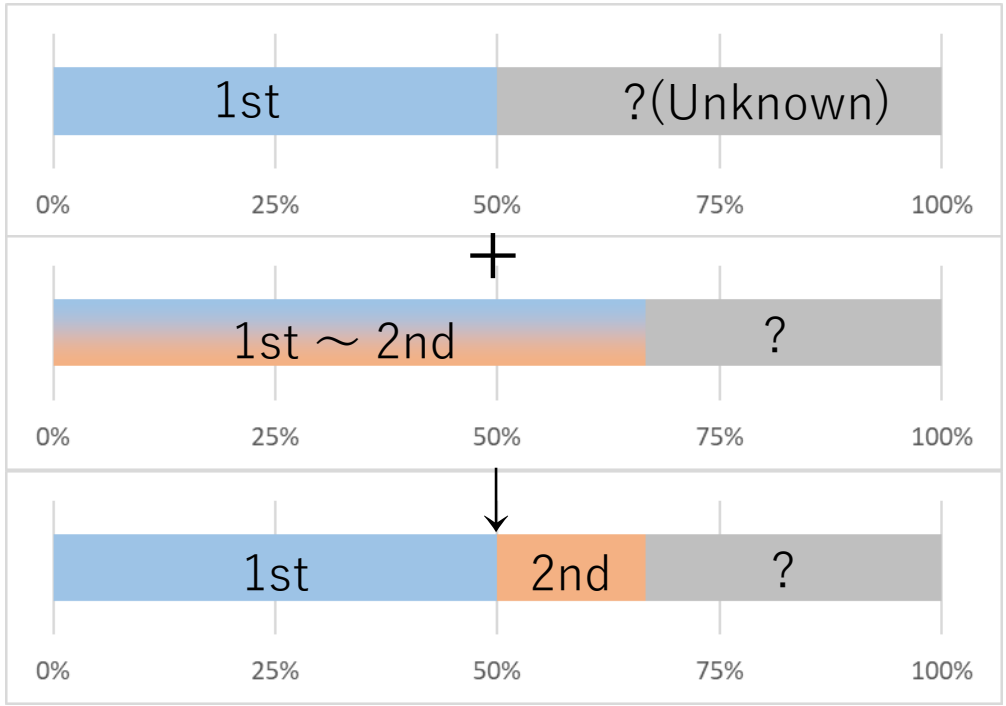
If there is no actual sugar content value

Theoretically possible sugar content ranges are taken as samples.

If you feel like you don't have enough specimens,
just use specimens of lower quality.

Let's consider the case where there are no known samples.
If it's a 4-choice question, we infer that each option has 25% probability.
Furthermore, if the objective variable is greater than or equal to 0, we infer that it is a uniform distribution of real numbers greater than or equal to 0.
This considers the answer options set by the questioner to be the sample as they are.
For example, consider the case where you want to predict the sugar content of an apple.
Suppose there is no data on actual measured sugar content.
In that case, we will take the entire theoretically possible range of sugar content as our sample.
If you feel that your sample is insufficient, you can just use a sample of lower quality.

Inference Synthesis (1)

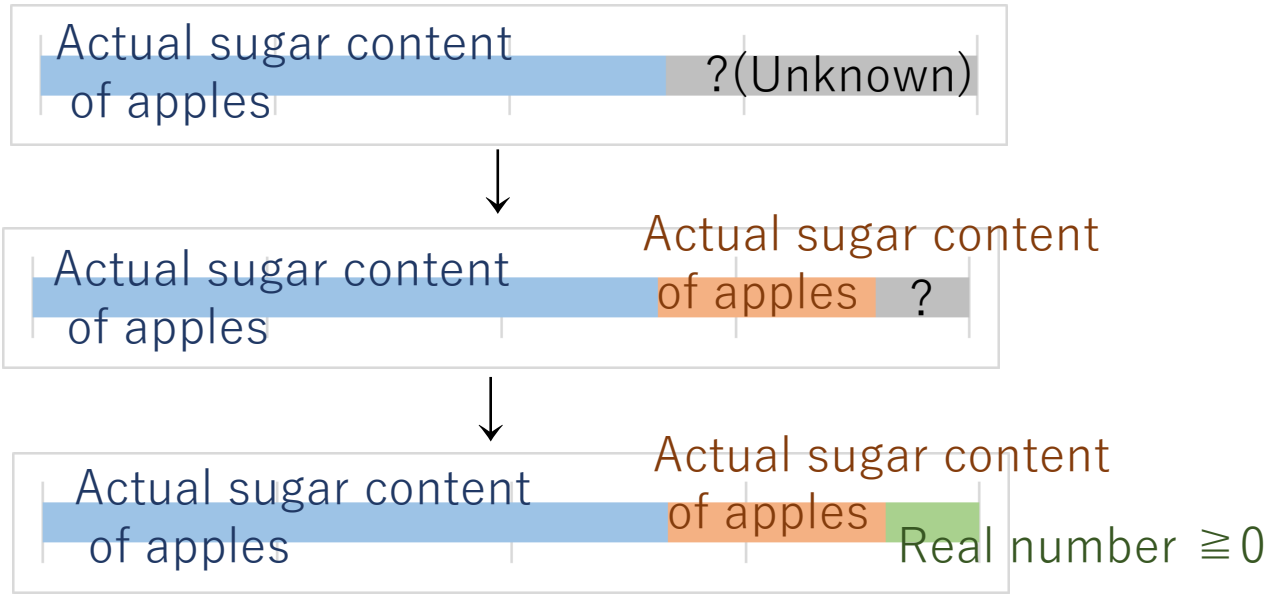


Other inferences can be assigned to the "unknown" parts
as long as they are not contradictory.

The "unknown" part of an inference result can be assigned other inference results as long as it does not contradict.
When calculating ideal induction, such inferences were synthesized.
Everything except "unknown" was kept, and a worse quality inference result was assigned to "unknown".
Since "unknown" is the worst inference result, if there are other inference results, they should be used rather than ignored.
In this example, the "unknown" parts common to the two inferences are kept.
Also, the first place part is not added, only the second place part is added.
This is to prevent contradictions such as exceeding the possible proportions with a single sample.

Inference Synthesis (2)

Question: What is the sugar content of an apple?



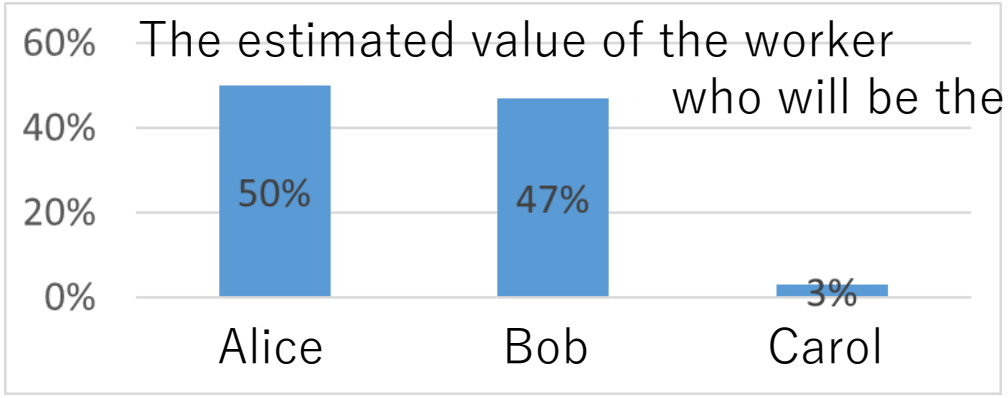
Results from unused samples are assigned as "unknown."

By taking possible values as a sample, we can eliminate the "unknown."

For example, if there is one sample, 50% of the guesses will be "unknown".
Since "unknown" is the worst-case value, if there is another inference result, it will be assigned to the "unknown" part.
However, assignments that would cause contradictions are prohibited.
You cannot assign the "unknown" part of an inference to the inference itself.
This would exceed the limit of the percentage that a single sample can occupy, which is determined by ideal induction.
It is possible to assign to the "unknown" part guesses of only the second or subsequent samples that have not yet been used.
If the assigned inference also contains "unknown", then "unknown" will remain.
It is also possible to assign a uniform distribution of the values that the objective variable can take as the guess.
In that case, it is possible to completely eliminate "unknown".

Quantizing the answer (1)

Question: Choose one worker to maximize your profits



Answer (A) Select "Alice" with the highest guess value.

Even if the difference is only within the margin of error,
you should choose the better option.

Answer (B) Determined by random numbers.

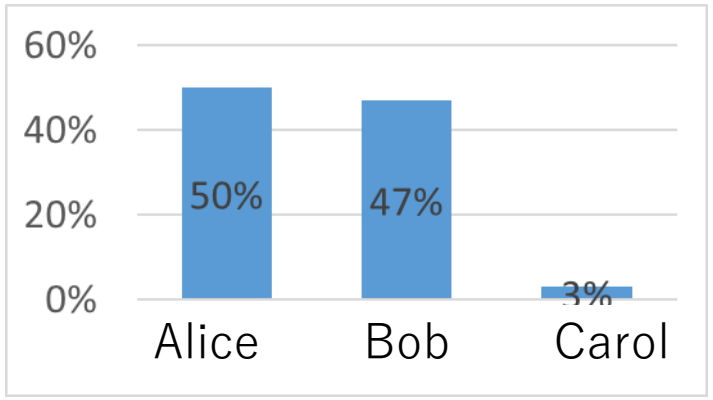
There is a 3% chance that "Carol" will be selected.
Even if the probability is low, it is a mistake to choose a bad guess.

The result of inductive inference is calculated as a probability distribution.
However, there are times when you must choose one option to answer a question.
For example, suppose you have to choose a single worker to maximize profits.
For answer (A), you would choose "Alice", who you guess will bring in the most profit.
You believe that you should make the best choice, even if the margin of error is within the range of error.
Answer (B) is decided by a random number.
"Carol" will be selected with a 3% probability.
Even if the probability is low, it is an error to choose someone you guess is bad.

Quantizing the answer (2)

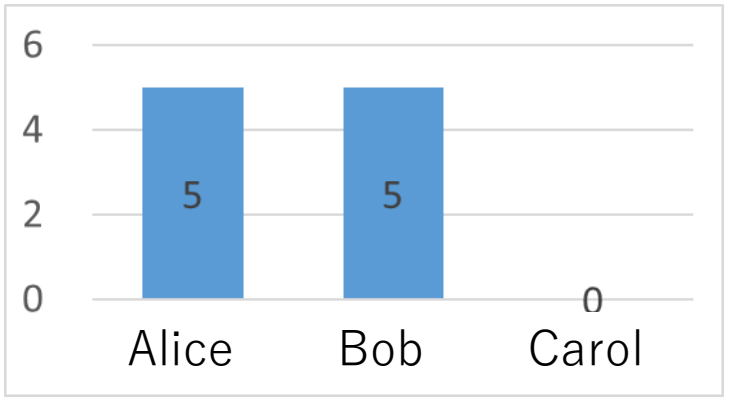
Question: Choose 10 workers to maximize your profits

The estimated value of the worker who will be the top earner



Quantize

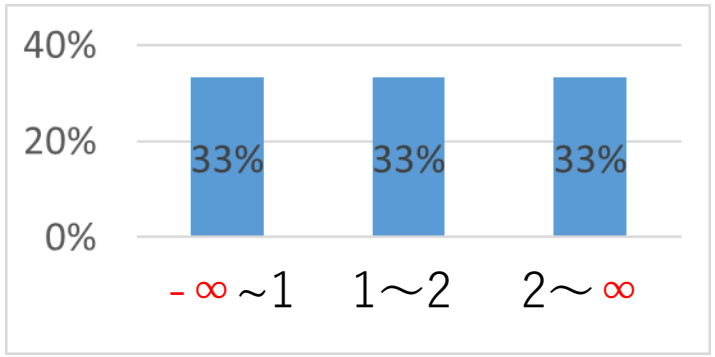
Answer

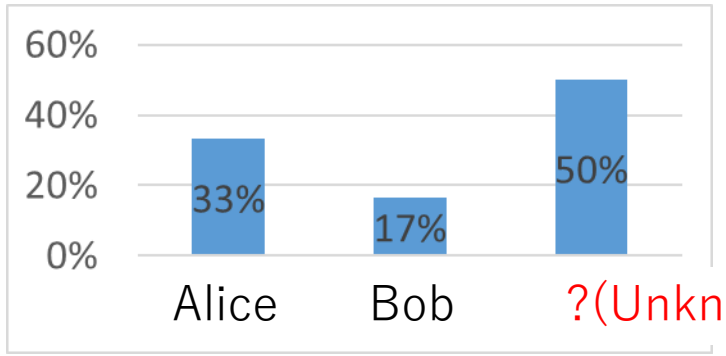


Quantize the probability distribution to maximize its "match rate"

Next, consider the case where ten people are chosen instead of one.
It is dangerous to choose ten “Alice”, who are in the top position.
This inference itself may not be accurate.
It would be better to choose some Bobs as well, as insurance in case Alice’s performance is unexpectedly poor.
Intuitively, it is optimal to choose an equal number of “Alice” and “Bob”.
When choosing options, the probability distribution is quantized.
It is optimal to quantize the probability distribution so as to maximize its "match rate"

Quantizing the answer (3)





1.5

Choose one value
as your answer

Alice

Even if "unknown" or "∞" is included,
it can be quantized to maximize the "match rate"

You can make inferences
even if you cannot calculate "mean", "expected value", "variance", etc.

The objective variable may include "unknown" or "∞".
This makes it difficult to calculate the mean and variance.
We can also assign a probability distribution of all possible values to "unknown".
Even then, it may not be possible to calculate the variance successfully.
However, even if "unknown" or "∞" are included, it is possible to quantize to maximize the "match rate".
Inference can be made without the need to calculate "mean", "expected value", "variance", etc.

Summary of Statistics 2.0

(1) As long as you can count numbers, you can freely set a "unit sample".

(2) As long as you don't use "OR" or "NOT", you can freely group and rank samples.

(3) The target for inference is grouped like other samples as an "unknown" sample.

(4) The "match rate" or "sample replacement number" for multiple sample groups is calculated to determine the sample weights.

(5) For the sample weight distribution, the "ideal rate" is calculated as the match rate with the “ideal induction”.

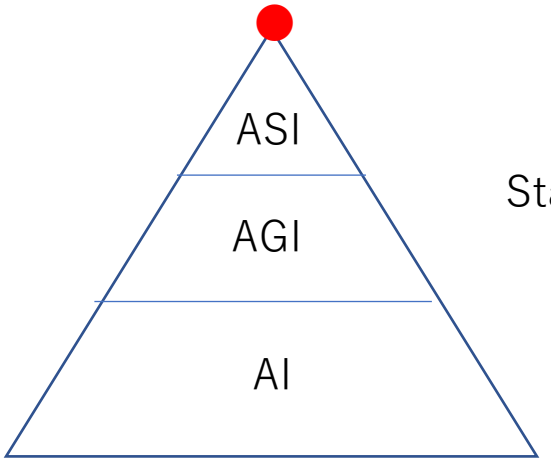
(6) The smaller the evaluation value obtained by multiplying the "distance" and "non-ideal rate" for each sample, the better the "sample selection strategy".

(7) The distribution of the sample's target variable is left "undegraded" as the inference result, without going through the population.

To summarize Statistics 2.0.
(1) As long as you can count numbers, you can freely set a "unit sample".
(2) As long as you don't use "OR" or "NOT", you can freely group and rank samples.
(3) The target for inference is grouped like other samples as an "unknown" sample.
(4) The "match rate" or "sample replacement number" for multiple sample groups is calculated to determine the sample weights.
(5) For the sample weight distribution, the "ideal rate" is calculated as the match rate with the “ideal induction”.
(6) The smaller the evaluation value obtained by multiply "distance" and "non-ideal rate" for each sample, the better the "sample selection strategy".
(7) The distribution of the sample's target variable is left "undegraded" as the inference result, without going through the population.

Statistics 2.0 and Ultimate AGI

Ultimate AGI: Get the optimal solution for every problem, no matter how long it takes.



Necessary basic theory completed.

Statistics 2.0:

Defining an objective optimal solution for every problem.

It is non-subjective and the same for everyone.

The ultimate goal, common to all those aiming for the ultimate AI,
has been set as Statistics 2.0.

If you understand Statistics 2.0, you have the foundation for creating the ultimate AGI.

Statistics 2.0 has defined the objective optimal solution for all problems.
In other words, the basic theory necessary for the ultimate AGI has been completed.
The definition of the ultimate AGI is obtaining the optimal solution for every problem, no matter how long it takes.
A program that follows this theory is guaranteed to arrive at the optimal solution.
The efficiency of the calculations depends on the skill of the programmer.
On the other hand, the optimal solution is free of subjectivity, so it is the same for everyone.
With Statistics 2.0, the ultimate goal common to all those aiming to create the ultimate AI has been set.
If you understand Statistics 2.0, you have the foundation for creating the ultimate AGI.

References

- Previous video
- And You

Here are some references.
Please also refer to the previous video.
And you.
If you understand intelligence, you can optimize your own intelligence.
If you understand that there are biases in your thinking, you can avoid them.
If you've understood the videos up to this point, you have incredible intelligence.
Even if you are entering unexplored research areas, believe in your own intelligence and move forward.

Afterword

This research is being carried out on volunteers.

There is no goal to replace labor and make a profit.

It would be more valuable if the ultimate AGI could make the impossible possible.

However, if we correctly understand intelligence, humans can reach ultimate intelligence.

The purpose of the research is to understand intelligence.

There are no plans to use the ultimate AGI for anything.

There is no need to benchmark it to prove its superiority.

So turning it into executable program code will also be postponed.

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

This research is being carried out on volunteers.
There is no goal to replace labor and make a profit.
It would be more valuable if the ultimate AGI could make the impossible possible.
However, if we correctly understand intelligence, humans can reach ultimate intelligence.
The purpose of the research is to understand intelligence.
There are no plans to use the ultimate AGI for anything.
There is no need to benchmark it to prove its superiority.
So turning it into executable program code will also be postponed.

Next Episode

The next video has not yet been decided.

If you understand the basic theories described in the videos so far,

you can already create the ultimate AGI.

If you want to be the first in the world to create the ultimate AGI for profit,

you should start applied research immediately.

If you wait for the next video, you may be overtaken by someone else.

From here on, research will likely focus on ways to efficiently arrive at the optimal solution.

It does not need to be the most efficient,

so it is probably less important than the research so far.

Please go ahead and create the world's first ultimate AGI with your own hands.

The next video has not yet been decided.
If you understand the basic theories described in the videos so far, you can already create the ultimate AGI.
If you want to be the first in the world to create the ultimate AGI for profit,
you should start applied research immediately.
If you wait for the next video, you may be overtaken by someone else.
From here on, research will likely focus on ways to efficiently arrive at the optimal solution.
It does not need to be the most efficient, so it is probably less important than the research so far.
Please go ahead and create the world's first ultimate AGI with your own hands.

Contact Information

For inquiries,

please contact us here.

https://ultagi.org/