Meaningless Statements in Performance Measurement for Intelligent Machines

Fred Roberts, DIMACS

Credit: Toby Hudson, wikimedia commons (no changes)

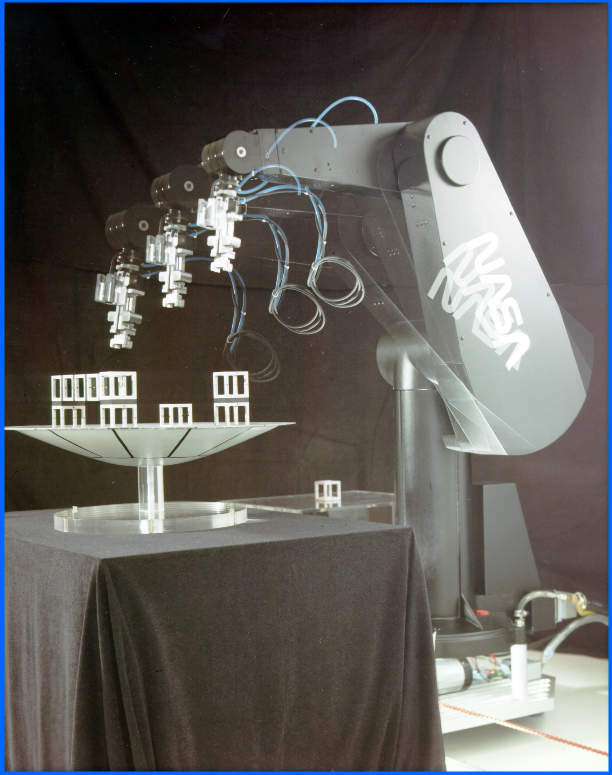
My Message

- **"Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules." - Fenton and Pfleeger, *Software Measurement***

- Message: Unless we are careful, statements using scales of measurement can be meaningless (in a precise sense).
- That is specifically relevant to measuring the performance of intelligent machines.

Some Questions We Will Ask

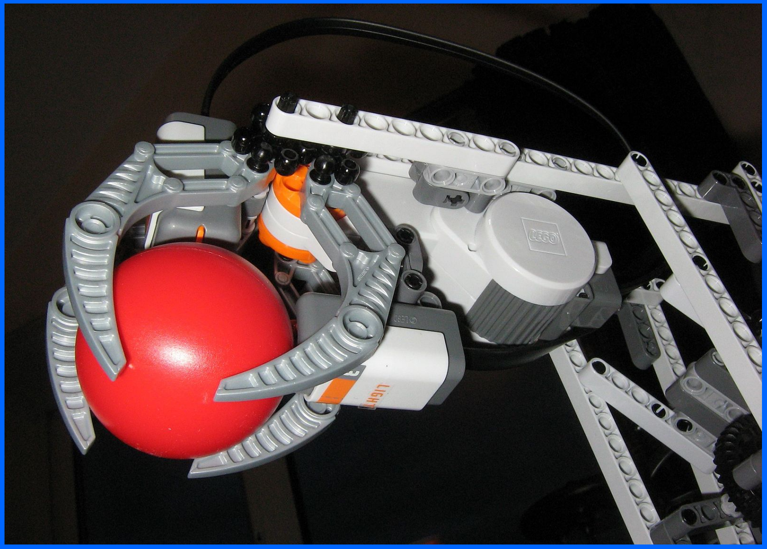
- A machine can complete one task three times as fast as another task.
- Is this a meaningful conclusion?



Credit: NASA/Dominic Hart, wikimedia commons

Some Questions We Will Ask

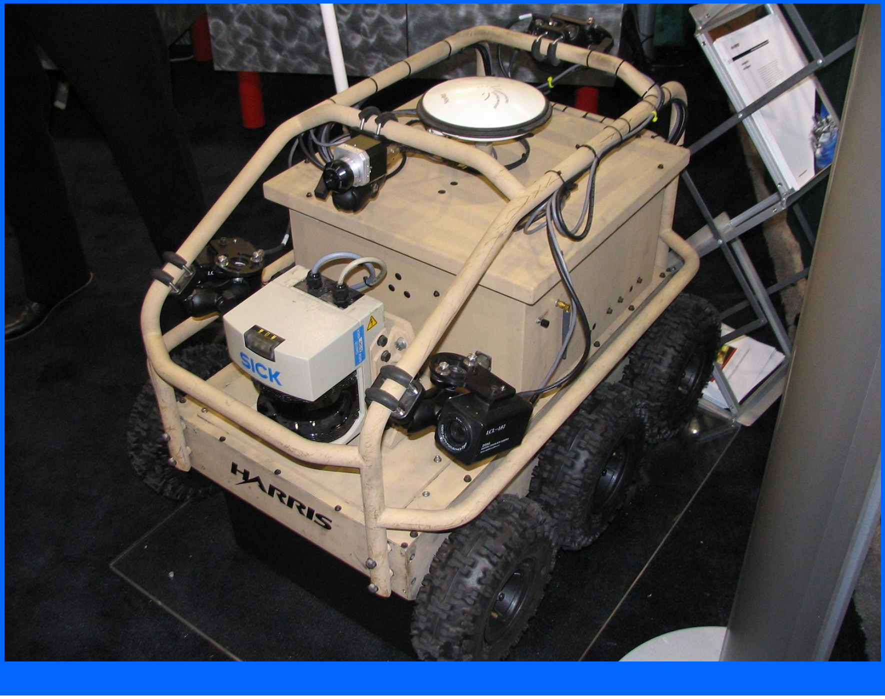
- A robot is tested on a variety of tasks involving lifting objects and then on a variety of tasks involving moving objects.
- Is it meaningful to say that the average weight of items lifted is greater than the average weight of items moved?



Credit: Crispin Semmens, wikimedia commons (no changes)

Some Questions We Will Ask

- Two unmanned vehicles have to move an item from one location to another.
- Is it meaningful to say that the path one vehicle takes is shorter than that of the other?

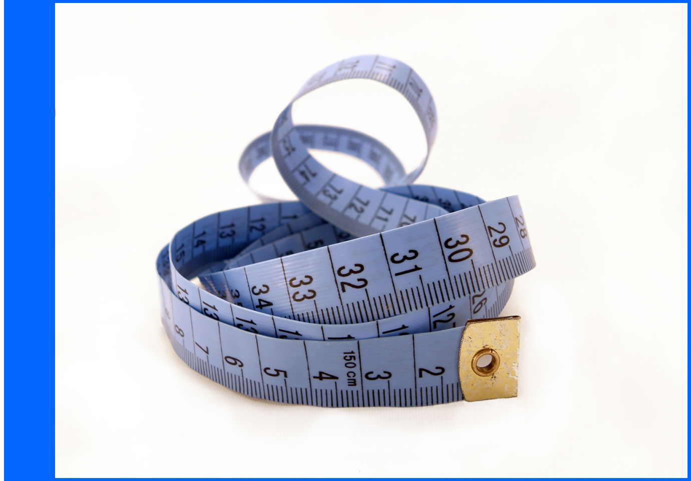


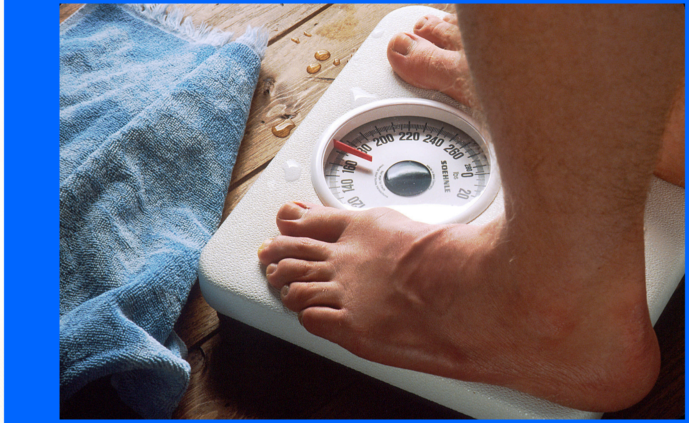
Credit: Johnny Zoo wikimedia commons

MEASUREMENT

- All of these questions have something to do with measurement.

- We will discuss applications of the theory of measurement to measurement in robotics and more generally for intelligent machines.





Credits: L ponizej, wikimedia commons C: Wikimedia commons R: OpenStax, wikimedia commons

MEASUREMENT

- ***Measurement*** has something to do with numbers.
- The theory of measurement was developed by mathematical social scientists to put measurement on a firm mathematical foundation.
- Think of starting with a set *A* of objects that we want to measure.
- We shall think of a ***scale of measurement*** as a function *f* that assigns a real number *f(a)* to each element *a* of *A* (or more generally assigns a number *f(a)* in some other set *B*).
- The representational theory of measurement gives conditions under which a function is an ***acceptable scale*** of measurement.
- Formalized through study of homomorphisms from one relational system to another.

Outline

1. **Theory of Uniqueness of Scales of Measurement/Scale Types**
2. Meaningful Statements
3. Average Machine Performance
4. Normalized Performance Scores
5. Optimization Problems for Intelligent Machines
6. How to Average Scores
7. Meaningfulness of Statistical Tests

The Theory of Uniqueness

Admissible Transformations

•An ***admissible transformation*** sends one acceptable scale into another.

Centigrade → Fahrenheit
Kilograms → Pounds

•In most cases one can think of an admissible transformation as defined on the range of a scale of measurement.

•Suppose f is an acceptable scale on A taking values in B .

•$\Phi{:}f(A) \rightarrow B$ is called an ***admissible transformation of f*** if $\Phi \circ f$ is again an acceptable scale.

The Theory of Uniqueness

Admissible Transformations Φ

Centigrade → Fahrenheit: $\Phi(x) = (9/5)x + 32$

Kilograms → Pounds: $\Phi(x) = 2.2x$

•A classification of scales is obtained by studying the class of admissible transformations associated with the scale.

•This defines the ***scale type***. (S.S. Stevens)

Some Common Scale Types

Class of Adm. Transfs.	Scale Type	Example
$\Phi(x) = \alpha x,\ \alpha > 0$	***ratio***	Mass Temp. (Kelvin) Time (intervals) Length Volume Loudness (sones)?
$\Phi(x) = \alpha x+\beta,\ \alpha > 0$	***interval***	Temp (F,C) Time (calendar)

Some Common Scale Types

Class of Adm. Transfs.	Scale Type	Example
$x \geq y \leftrightarrow \Phi(x) \geq \Phi(y)$ Φ strictly increasing	***ordinal***	Preference? Hardness Grades of leather, wool, etc. Subjective judgments: cough, fatigue,...
$\Phi(x) = x$	***absolute***	Counting

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. **Meaningful Statements**
3. Average Machine Performance
4. Normalized Performance Scores
5. Optimization Problems for Intelligent Machines
6. How to Average Scores
7. Meaningfulness of Statistical Tests

Meaningful Statements

•In measurement theory, we speak of a statement as being ***meaningful*** if its truth or falsity is not an artifact of the particular scale values used.

•The following definition is due to Suppes 1959 and Suppes and Zinnes 1963.

<u>Definition</u>: A statement involving numerical scales is ***meaningful*** if its truth or falsity is unchanged after any (or all) of the scales is transformed (independently?) by an admissible transformation.

Meaningful Statements

•A slightly more informal definition:

Alternate Definition: A statement involving numerical scales is ***meaningful*** if its truth or falsity is unchanged after any (or all) of the scales is (independently?) replaced by another acceptable scale.

•In some practical examples, for example those involving preference judgments or judgments "louder than" under the "semiorder" model, it is possible to have two scales where one can't go from one to the other by an admissible transformation, so one has to use this definition.

Meaningful Statements

- We will avoid the long literature of more sophisticated approaches to meaningfulness.

- Situations where this relatively simple-minded definition may run into trouble will be disregarded.

- Emphasis is to be on applications of the "invariance" motivation behind the theory of meaningfulness.

Meaningful Statements

"A machine can complete task *a* three times as fast as task *b*."

- Is this meaningful?

Credit: ICAPlants, Wikimedia commons (no changes)

Meaningful Statements

"A machine can complete task *a* three times as fast as task *b*."

- Is this meaningful?
- We have a ratio scale (time intervals).

(1) $$f(a) = 3f(b).$$

- This is meaningful if f is a ratio scale. For, an admissible transformation is $\Phi(x) = \alpha x$, $\alpha > 0$. We want (1) to hold iff

(2) $$(\Phi \circ f)(a) = 3(\Phi \circ f)(b)$$

- But (2) becomes

(3) $$\alpha f(a) = 3\alpha f(b)$$

- (1) ↔ (3) since $\alpha > 0$.

Meaningful Statements

“After completing its task, the machine’s temperature will be 2 per cent higher than it was at the beginning.”

- Is this meaningful?

Meaningful Statements

“After completing its task, the machine’s temperature will be 2 per cent higher than it was at the beginning.”

$$f(a) = 1.02f(b)$$

- Meaningless. It could be true with Fahrenheit and false with Centigrade, or vice versa.

Meaningful Statements

In general:

- For ratio scales, it is meaningful to compare ratios:

$$f(a)/f(b) > f(c)/f(d)$$

- For interval scales, it is meaningful to compare intervals:

$$f(a) - f(b) > f(c) - f(d)$$

- For ordinal scales, it is meaningful to compare size:

$$f(a) > f(b)$$

Meaningful Statements

"I weigh 1000 times what that elephant weighs."

- Is this meaningful?

Meaningful Statements

"I weigh 1000 times what that elephant weighs."

- Meaningful. It involves ratio scales.
It is false no matter what the unit.

- ***Meaningfulness is different from truth.***

- It has to do with what kinds of assertions it makes sense to make, which assertions are not accidents of the particular choice of scale (units, zero points) in use.

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. **Average Machine Performance**
4. Normalized Performance Scores
5. Optimization Problems for Intelligent Machines
6. How to Average Scores
7. Meaningfulness of Statistical Tests

Average Machine Performance

- Compare the performance of a machine on two groups of tasks to see which it is better at.

- **Data suggests that the average performance on tasks in the first group is higher than the average performance on tasks in the second group.**

- A robot lifts some items and moves some items.
- Average weight of items lifted is greater than average weight of items moved.

- Is this meaningful?

Average Machine Performance

- Compare the performance of a machine on two groups of tasks to see which it is better at.

- *f(a)* is machine performance on task *a*

- **Data suggests that the average performance on tasks in the first group is higher than the average performance on tasks in the second group.**

$a_1, a_2, \ldots, a_n$ tasks in first group
$b_1, b_2, \ldots, b_m$ tasks in second group

$$(1) \quad ({}^{1}/_{n}) \sum_{i=1}^{n} f(a_i) > ({}^{1}/_{m}) \sum_{i=1}^{m} f(b_i)$$

- We are comparing ***arithmetic means***.

Average Machine Performance

- Statement (1) is meaningful iff for all admissible transformations of scale Φ, (1) holds iff

$$(2) \quad (^1/_n) \sum_{i=1}^{n} (\Phi \circ f)(a_i) > (^1/_m) \sum_{i=1}^{m} (\Phi \circ f)(b_i)$$

- **If machine performance defines a ratio scale:**
- Then, $\Phi(x) = \alpha x$, $\alpha > 0$, so (2) becomes

$$(3) \quad (^1/_n) \sum_{i=1}^{n} \alpha f(a_i) > (^1/_m) \sum_{i=1}^{m} \alpha f(b_i)$$

- Then $\alpha > 0$ implies (1) $\leftrightarrow$ (3). Hence, (1) is meaningful.
- So this kind of comparison would work if we were comparing weights of objects lifted or moved.

Average Machine Performance

- Note: **(1) is still meaningful if *f* is an interval scale.**
- For example, we could be comparing achieved temperatures *f(a)*.
- Here, $\Phi(x) = \alpha x + \beta$, $\alpha > 0$. Then (2) becomes

$$(4) \quad ({}^1/_n) \sum_{i=1}^{n} \alpha f(a_i)+\beta > ({}^1/_m) \sum_{i=1}^{m} \alpha f(b_i)+\beta$$

- This readily reduces to (1).
- However, **(1) is meaningless if *f* is just an ordinal scale.**

Average Machine Performance

- Suppose that *f(a)* is measured on an ordinal scale, e.g., 5-point scale: 5=very good, 4=good, 3=good, 2=bad, 1=very bad.

- **In such a scale, the numbers may not mean anything; only their order matters.**

Group 1: 5, 3, 1 average 3
Group 2: 4, 4, 2 average 3.33

- Conclude: average performance on group 2 tasks is higher.

Average Machine Performance

•Suppose that *f(a)* is measured on an ordinal scale, e.g., 5-point scale: 5=very good, 4=good, 3=good, 2=bad, 1=very bad.

•In such a scale, the numbers may not mean anything; only their order matters.

Group 1: 5, 3, 1 average 3
Group 2: 4, 4, 2 average 3.33 (greater)

•Admissible transformation: 5 → 100, 4 → 75, 3 → 65, 2 → 40, 1 → 30

•New scale conveys the same information. New scores:

Group 1: 100, 65, 30 average 65
Group 2: 75, 75, 40 average 63.33

Conclude: average performance on group 1 tasks is higher.

Average Machine Performance

- **Thus, comparison of arithmetic means can be meaningless for ordinal data.**

- Of course, you may argue that in the 5-point scale, at least *equal spacing* between scale values is an inherent property of the scale. In that case, the scale is *not* ordinal and this example does not apply.

- Note: C**omparing *medians* is meaningful with ordinal scales**: To say that one group has a higher median than another group is preserved under admissible transformations.

Average Machine Performance II

•Suppose each of *n* observers is asked to rate each of a collection of machines as to their performance on a given task.

•Similarly if we judge the machines as to performance on *n* different criteria.

•Let $f_i(a)$ be the rating of machine a on the task by judge i (under criterion i). Is it meaningful to assert that the average rating of machine a is higher than the average rating of machine b?

Average Machine Performance II

•Let $f_i(a)$ be the rating of machine a on a given task by judge i (under criterion i). Is it meaningful to assert that the average rating of machine a is higher than the average rating of machine b?

(1) $$(^1/_n) \sum_{i=1}^{n} f_i(a) > (^1/_n) \sum_{i=1}^{n} f_i(b)$$

Average Machine Performance II

- Let $f_i(a)$ be the rating of machine a on a given task by judge i (under criterion i). Is it meaningful to assert that the average rating of machine a is higher than the average rating of machine b?

$$(1) \quad ({}^1/_n) \sum_{i=1}^{n} f_i(a) > ({}^1/_n) \sum_{i=1}^{n} f_i(b)$$

- If each f_i is a ratio scale, then we consider for $\alpha > 0$,

$$(2) \quad ({}^1/_n) \sum_{i=1}^{n} \alpha f_i(a) > ({}^1/_n) \sum_{i=1}^{n} \alpha f_i(b)$$

- Clearly, (1) $\leftrightarrow$ (2), so (1) is meaningful.

Average Machine Performance II

- If each f_i is a ratio scale, then we consider for $\alpha > 0$,

$$(2) \quad (^1/_n) \sum_{i=1}^{n} \alpha f_i(a) > (^1/_n) \sum_{i=1}^{n} \alpha f_i(b)$$

- Clearly, (1) ↔ (2), so (1) is meaningful.

- Problem: $f_1, f_2, \ldots, f_n$ might have ***independent units***. In this case, we want to allow independent admissible transformations of the f_i. Thus, we must consider

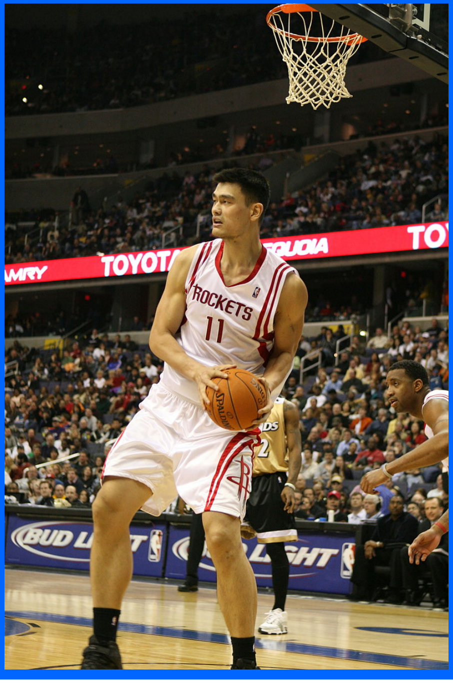
$$(3) \quad (^1/_n) \sum_{i=1}^{n} \alpha_i f_i(a) > (^1/_n) \sum_{i=1}^{n} \alpha_i f_i(b)$$

- It is easy to see that there are α_i so that (1) holds and (3) fails. Thus, (1) is meaningless.

Average Machine Performance II

Motivation for considering different α_i:
Machine = person, task is to play football

$n = 2$, $f_1(a)$ = weight of a, $f_2(a)$ = height of a. Then (1) says that the average of a's weight and height is greater than the average of b's weight and height. This could be true with one combination of weight and height scales and false with another.

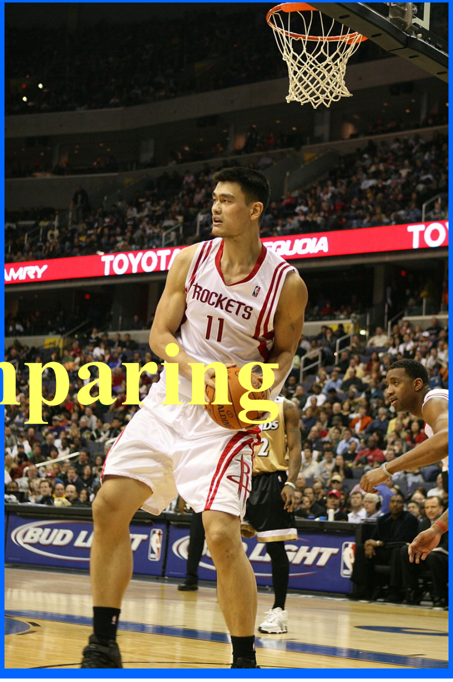


Credit:
L: Miguel Castaneda: wikimedia commons
R: Keith Allison wikimedia commons (no changes)

Average Machine Performance II

Motivation for considering different α_i:
Machine = person, task is to play football

$n = 2$, $f_1(a)$ = weight of a, $f_2(a)$ = height of a. Then (1) says that the average of a's weight and height is greater than the average of b's weight and height. This could be true with one combination of weight and height scales and false with another.



- **Conclusion: Be careful when comparing arithmetic mean ratings.**

Average Machine Performance II

•In this context, it is safer to compare ***geometric means*** (Dalkey).

$$\sqrt[n]{\Pi\, f_i(a)} > \sqrt[n]{\Pi\, f_i(b)} \leftrightarrow \sqrt[n]{\Pi\, \alpha_i f_i(a)} > \sqrt[n]{\Pi\, \alpha_\mathrm{i}\mathrm{f}_\mathrm{i}(b)}$$

all $\alpha_i > 0$.

•Thus, if each f_i is a ratio scale, if individuals can change performance rating scales independently, then ***comparison of geometric means is meaningful while comparison of arithmetic means is not.***

Application of this Idea

Role of Air Pollution in Health.

•In a study of air pollution and related energy use in San Diego, a panel of experts each estimated the relative importance of variables relevant to air pollution using the ***magnitude estimation procedure***. Roberts (1972, 1973).

•***Magnitude estimation***: Most important gets score of 100. If half as important, score of 50. And so on.

•If magnitude estimation leads to a ratio scale -- Stevens presumes this -- then comparison of geometric mean importance ratings is meaningful.

•However, comparison of arithmetic means may not be. Geometric means were used.

Credits:
Uipper:Welp.sk wikimedia commons (no changes)
Lower: Kentaro IEMOTO, wikimedia commons (no changes)

Magnitude Estimation by One Expert of Relative Importance for Air Pollution of Variables Related to Commuter Bus Transportation in a Given Region

Variable	Rel. Import. Rating
1. No. bus passenger mi. annually	80
2. No. trips annually	100
3. No. miles of bus routes	50
4. No. miles special bus lanes	50
5. Average time home to office	70
6. Average distance home to office	65
7. Average speed	10
8. Average no. passengers per bus	20
9. Distance to bus stop from home	50
10. No. buses in the region	20
11. No. stops, home to office	20

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Average Machine Performance
4. **Normalized Performance Scores**
5. Optimization Problems for Intelligent Machines
6. How to Average Scores
7. Meaningfulness of Statistical Tests

Normalized Performance Scores

- A widely used method in hardware measurement starts with scoring performance of different machines (systems) under different criteria or benchmarks.
- The scores on each criterion are normalized relative to the score of one of the machines.
- The normalized scores are combined by some averaging procedure and normalized scores are compared.
- The machine with the highest average normalized score is chosen.
- Fleming and Wallace show that the outcome can depend on the choice of the base system.
- So it is meaningless in the sense of measurement theory.

Performance Scores

		CRITERION				
		E	F	G	H	I
MACHINE	R	417	83	66	39,449	772
	M	244	70	153	33,527	368
	Z	134	70	135	66,000	369

Performance Scores

Normalize Relative to Machine R

		CRITERION				
		E	F	G	H	I
MACHINE	R	417 1.00	83 1.00	66 1.00	39,449 1.00	772 1.00
	M	244 .59	70 .84	153 2.32	33,527 .85	368 .48
	Z	134 .32	70 .85	135 2.05	66,000 1.67	369 .45

Performance Scores

Take Arithmetic Mean of Normalized Scores

		CRITERION					Arithmetic Mean
		E	F	G	H	I	
MACHINE	R	417 1.00	83 1.00	66 1.00	39,449 1.00	772 1.00	1.00
	M	244 .59	70 .84	153 2.32	33,527 .85	368 .48	1.01
	Z	134 .32	70 .85	135 2.05	66,000 1.67	369 .45	1.07

Performance Scores

Take Arithmetic Mean of Normalized Scores

		CRITERION					
		E	F	G	H	I	Arithmetic Mean
MACHINE	R	417 1.00	83 1.00	66 1.00	39,449 1.00	772 1.00	1.00
	M	244 .59	70 .84	153 2.32	33,527 .85	368 .48	1.01
	Z	134 .32	70 .85	135 2.05	66,000 1.67	369 .45	1.07

Conclude that machine Z is best

Performance Scores

Now Normalize Relative to Machine M

		CRITERION				
		E	F	G	H	I
MACHINE	R	417 1.71	83 1.19	66 .43	39,449 1.18	772 2.10
	M	244 1.00	70 1.00	153 1.00	33,527 1.00	368 1.00
	Z	134 .55	70 1.00	135 .88	66,000 1.97	369 1.00

Performance Scores

Take Arithmetic Mean of Normalized Scores

		CRITERION					Arithmetic Mean
		E	F	G	H	I	
MACHINE	R	417 1.71	83 1.19	66 .43	39,449 1.18	772 2.10	1.32
	M	244 1.00	70 1.00	153 1.00	33,527 1.00	368 1.00	1.00
	Z	134 .55	70 1.00	135 .88	66,000 1.97	369 1.00	1.08

Performance Scores

Take Arithmetic Mean of Normalized Scores

		CRITERION					
		E	F	G	H	I	Arithmetic Mean
MACHINE	R	417 1.71	83 1.19	66 .43	39,449 1.18	772 2.10	1.32
	M	244 1.00	70 1.00	153 1.00	33,527 1.00	368 1.00	1.00
	Z	134 .55	70 1.00	135 .88	66,000 1.97	369 1.00	1.08

Conclude that machine R is best

Normalized Scores

- So, the conclusion that a given machine is best by taking arithmetic mean of normalized scores is meaningless in this case.
- Above example from Fleming and Wallace (Communications of the ACM, 1986), data from Heath (1984) (in a computing machine application)
- Sometimes, ***geometric mean*** is helpful.

Performance Scores

Normalize Relative to Machine R

		CRITERION					Geometric Mean
		E	F	G	H	I	
MACHINE	R	417 1.00	83 1.00	66 1.00	39,449 1.00	772 1.00	**1.00**
	M	244 .59	70 .84	153 2.32	33,527 .85	368 .48	**.86**
	Z	134 .32	70 .85	135 2.05	66,000 1.67	369 .45	**.84**

Conclude that treatment R is best

Performance Scores

Now Normalize Relative to Machine M

CRITERION

		E	F	G	H	I	Geometric Mean
MACHINE	R	417 1.71	83 1.19	66 .43	39,449 1.18	772 2.10	1.17
	M	244 1.00	70 1.00	153 1.00	33,527 1.00	368 1.00	1.00
	Z	134 .55	70 1.00	135 .88	66,000 1.97	369 1.00	.99

Still conclude that treatment R is best

Normalized Scores

- In this situation, it is easy to show that ***the conclusion that a given machine has highest geometric mean normalized score is a meaningful conclusion.***
- ***Even meaningful: A given machine has geometric mean normalized score 20% higher than another machine.***
- Fleming and Wallace give general conditions under which comparing geometric means of normalized scores is meaningful.
- Research area: what averaging procedures make sense in what situations? Large literature.

Treatment Evaluation

Message from measurement theory:

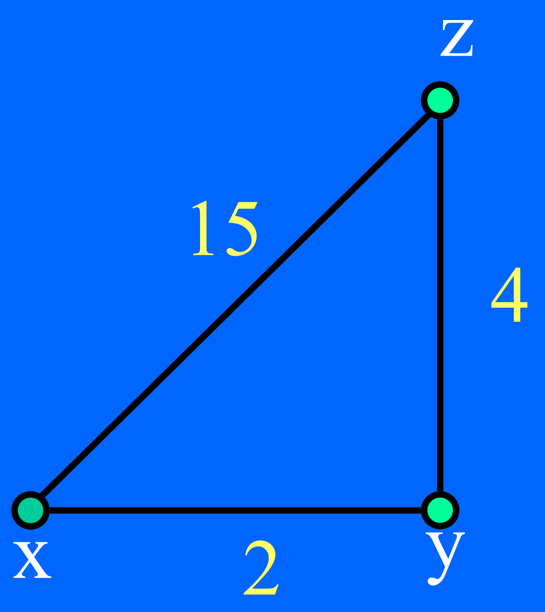
Do not perform arithmetic operations on data without paying attention to whether the conclusions you get are meaningful.

Credit: Toby Hudson, wikimedia commons (no changes)

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Average Machine Performance
4. Normalized Performance Scores
5. **Optimization Problems for Intelligent Machines**
6. How to Average Scores
7. Meaningfulness of Statistical Tests

Optimization Problems for Intelligent Machines: Shortest Path Problem

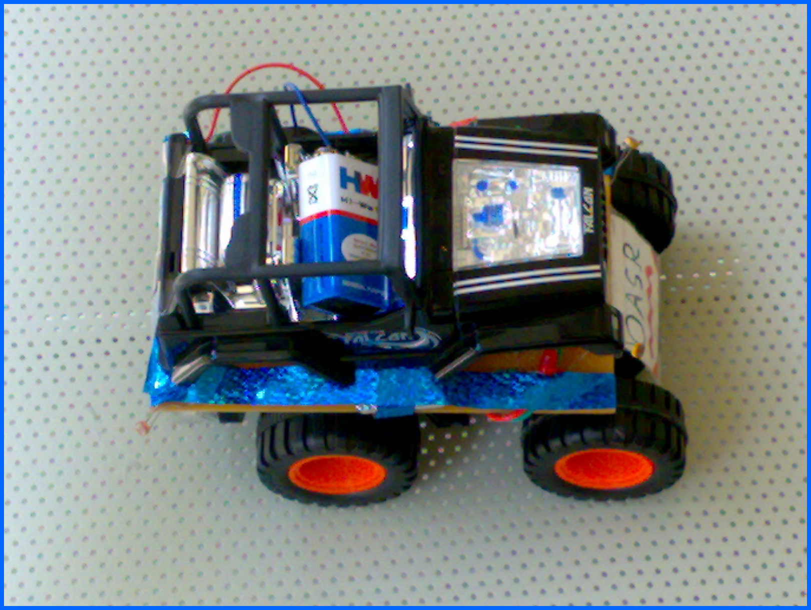


Numbers = some sort of weights

- *Problem: Find the shortest path from x to z.*

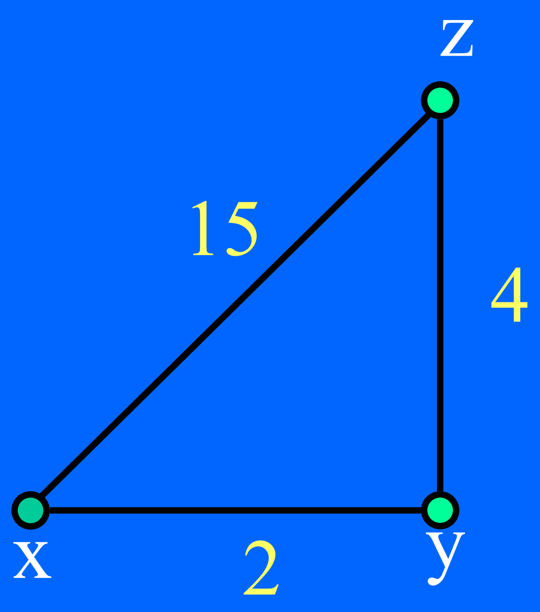
Optimization Problems for Intelligent Machines: Shortest Path Problem

- This problem frequently arises in robotics, e.g., in finding a path that minimizes time or energy expended, and it certainly arises with unmanned vehicles.



Credit: Mittgaurav: wikimedia commons

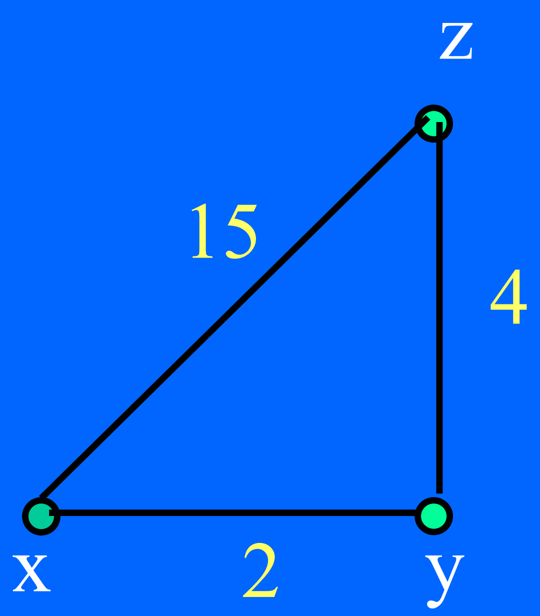
Shortest Path Problem



Numbers = some sort of weights

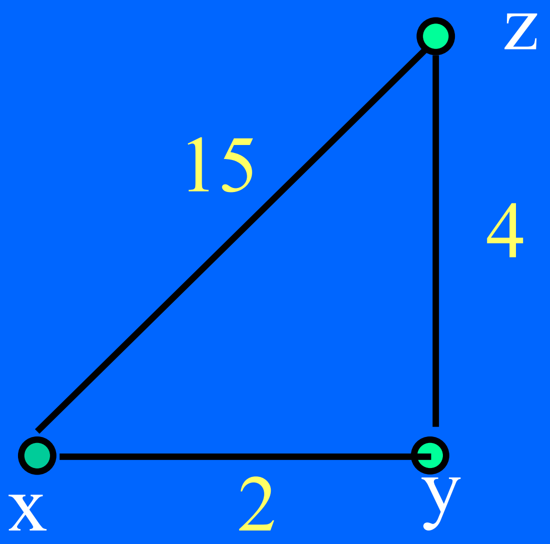
- *So what is the shortest path from x to z?*

Shortest Path Problem



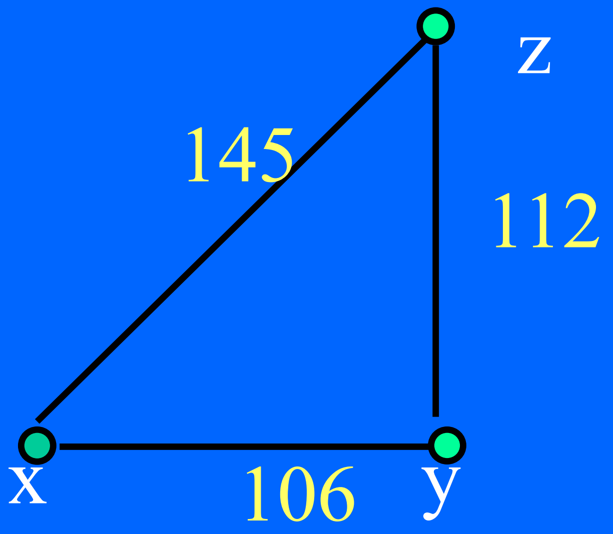
- The shortest path from x to z is the path x to y to z.
- Is this conclusion meaningful?
- It is if the numbers define a ratio scale.
- The numbers define a ratio scale if they are distances.

Shortest Path Problem



- However, what if the numbers define an interval scale?
- For example, the numbers could be costs in terms of utility (or disutility) assigned to a route, and these might only define an interval scale.

Shortest Path Problem

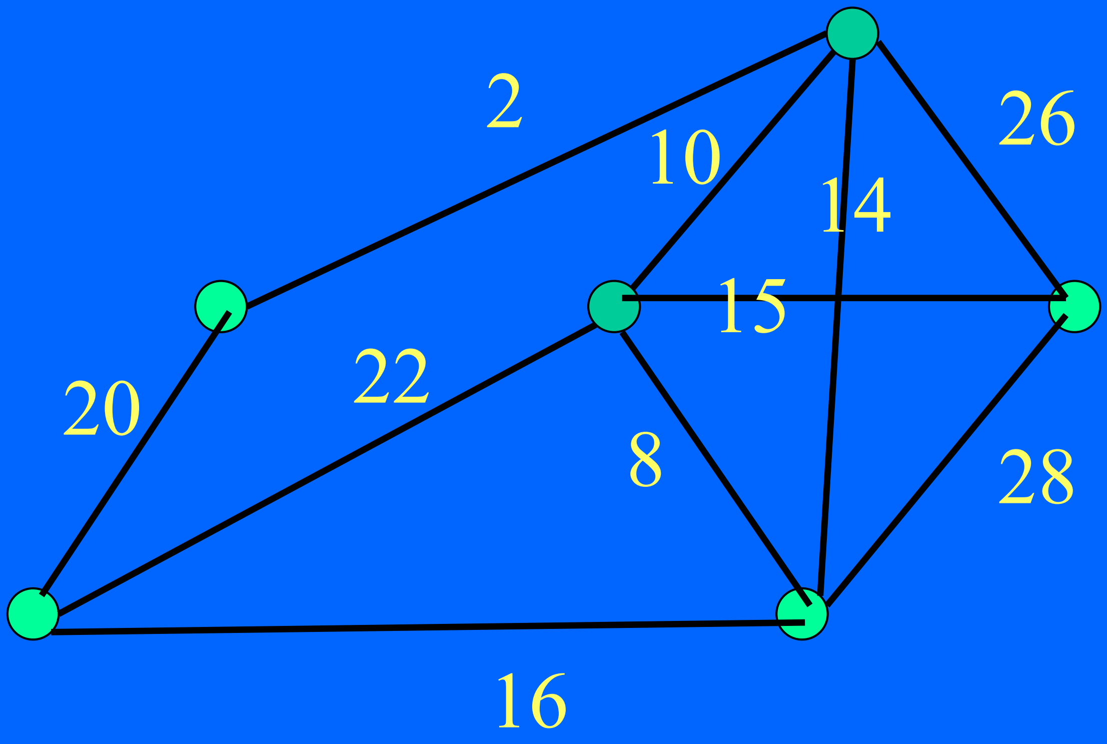


- Consider the admissible transformation $\Phi(x) = 3x + 100$.
- Now we get the above numbers on the edges.
- Now the shortest path is to go directly from x to z.
- The original conclusion was meaningless.

Linear Programming

- The shortest path problem can be formulated as a linear programming problem.
- ***Thus: The conclusion that A is the solution to a linear programming problem can be meaningless if cost parameters are measured on an interval scale.***

Related Example: Minimum Spanning Tree Problem



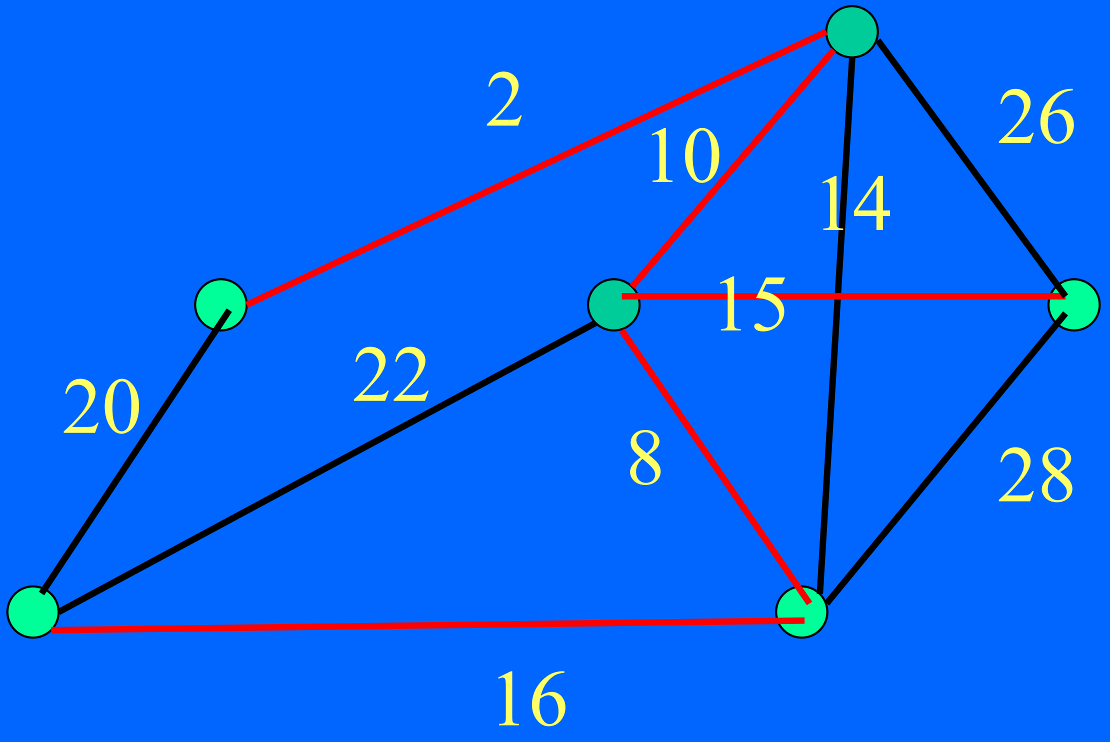
- A spanning tree is a tree using the edges of the graph and containing all of the vertices.
- It is minimum if the sum of the numbers on the edges used is as small as possible.

Related Example: Minimum Spanning Tree Problem

- Minimum spanning trees arise in many applications.
- It arises in construction applications where automated robots pick up and drop off building blocks located at vertices (nodes) of a network.

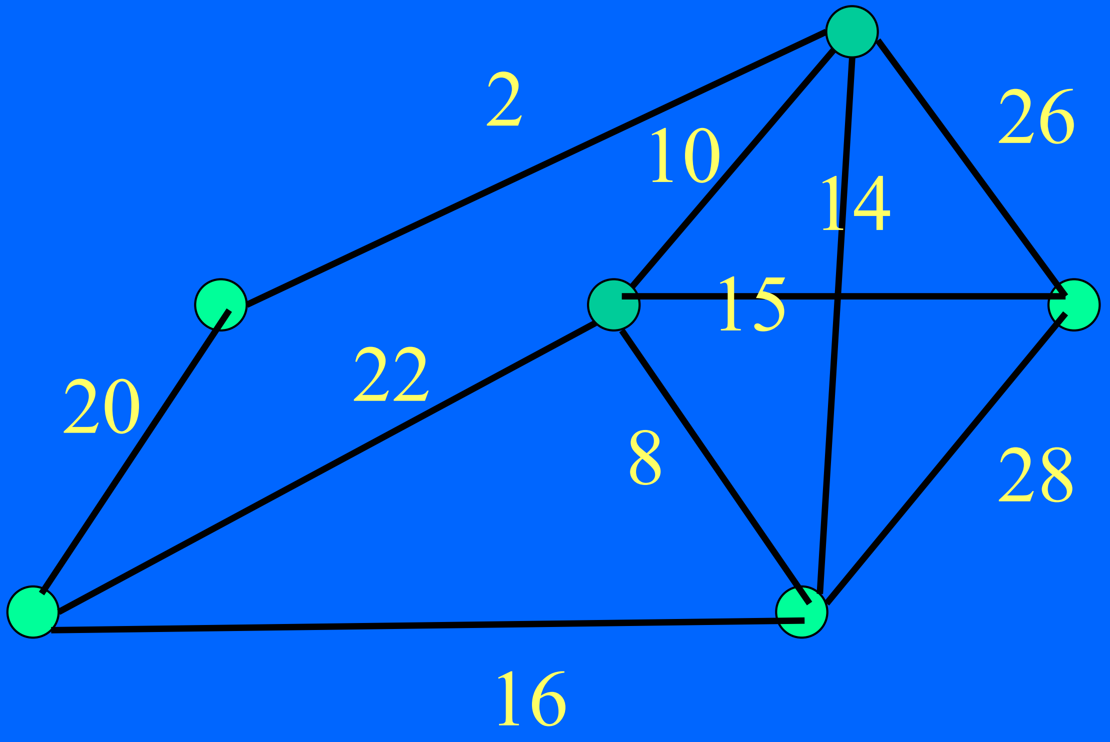
Credit: Wikimedia commons

Related Example: Minimum Spanning Tree Problem



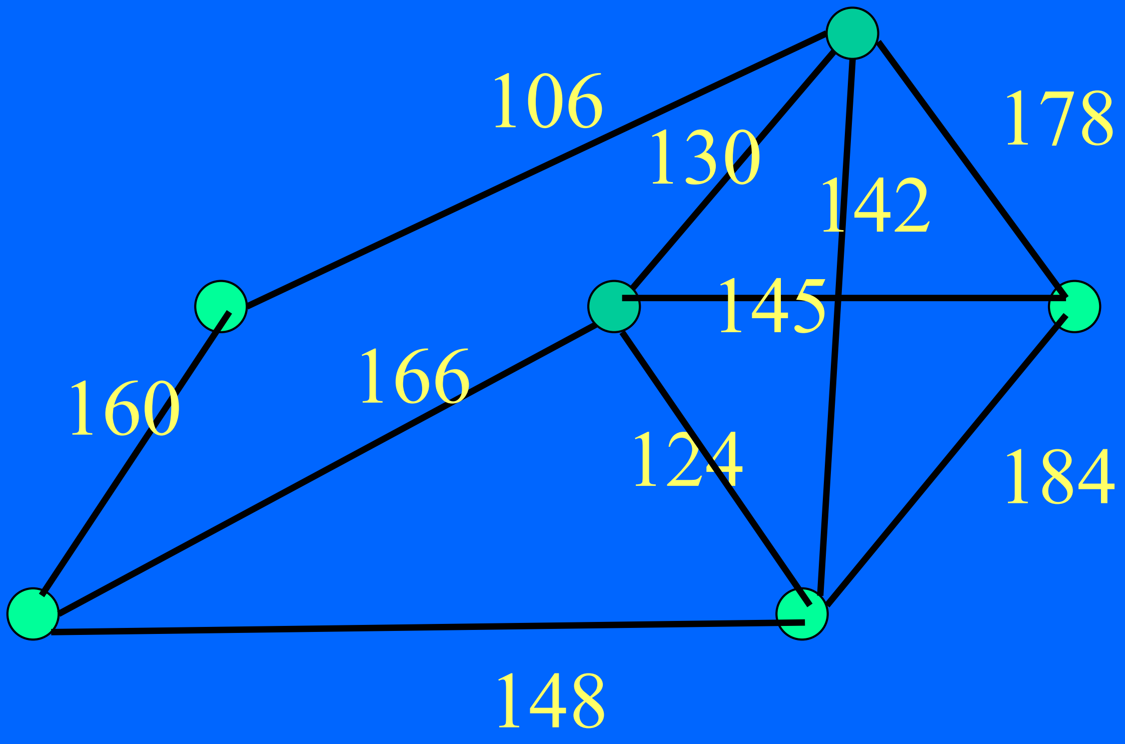
- Red edges define a minimum spanning tree.
- Is it meaningful to conclude that this is a minimum spanning tree?

Related Example: Minimum Spanning Tree Problem



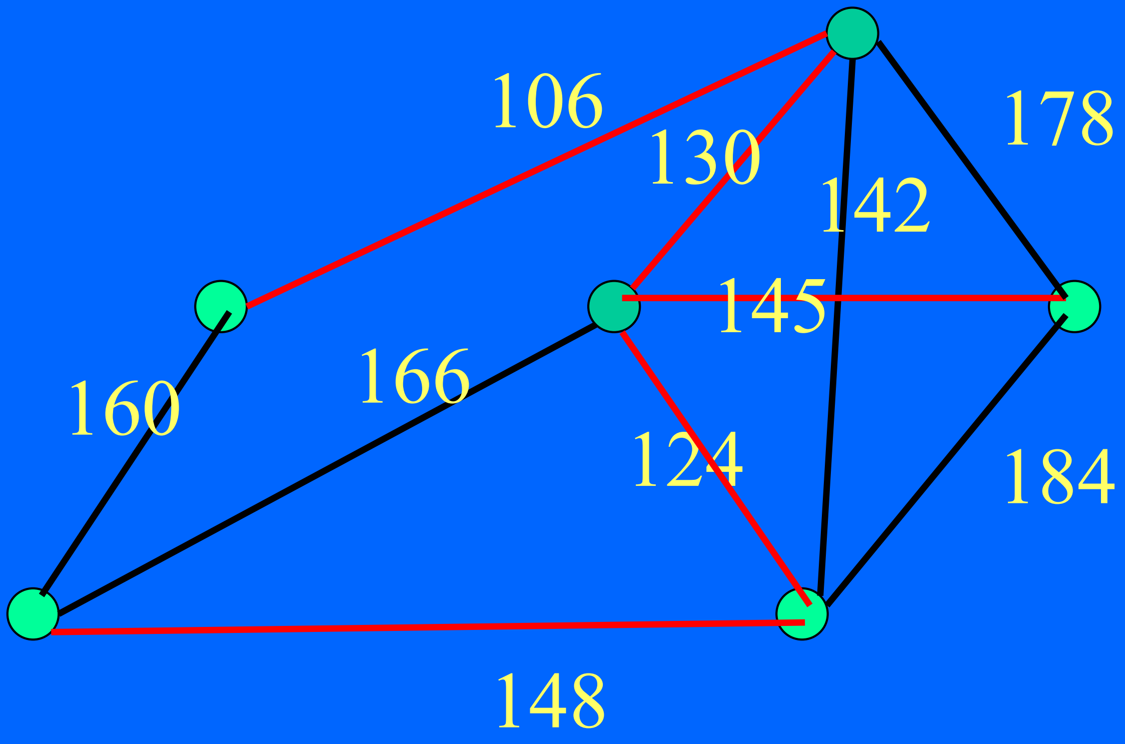
- Consider the admissible transformation $\varphi(x) = 3x + 100$.

Related Example: Minimum Spanning Tree Problem



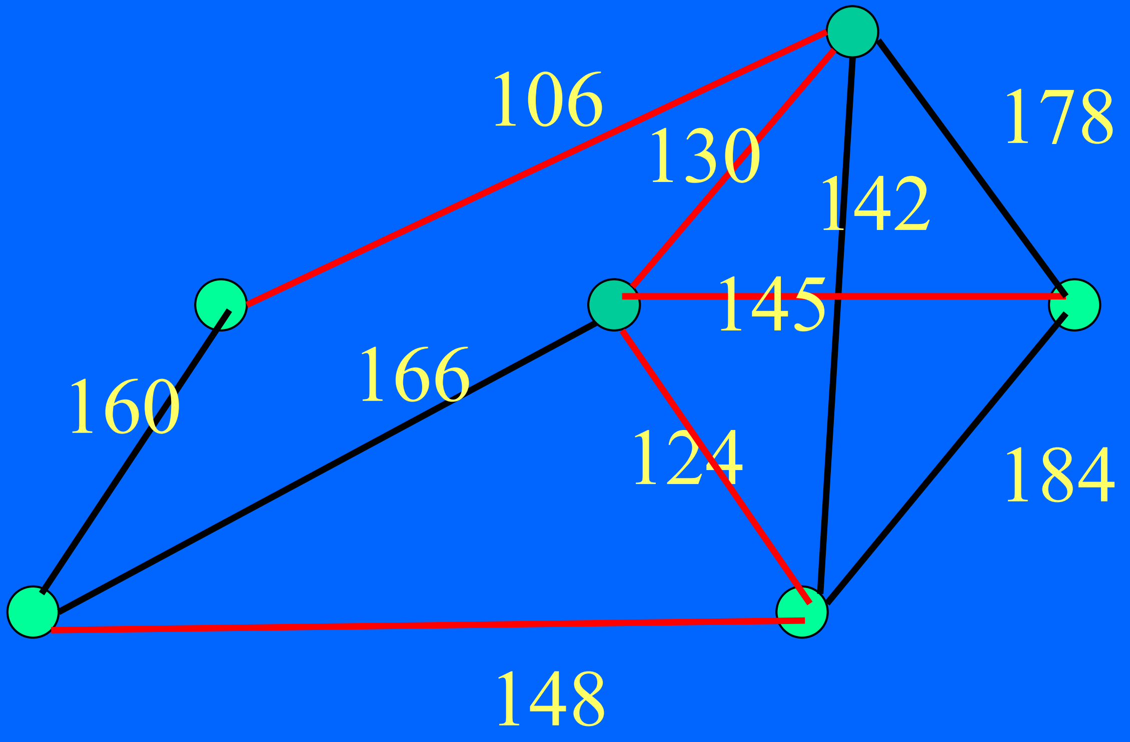
- Consider the admissible transformation $\varphi(x) = 3x + 100$.
- We now get the above numbers on edges.

Related Example: Minimum Spanning Tree Problem



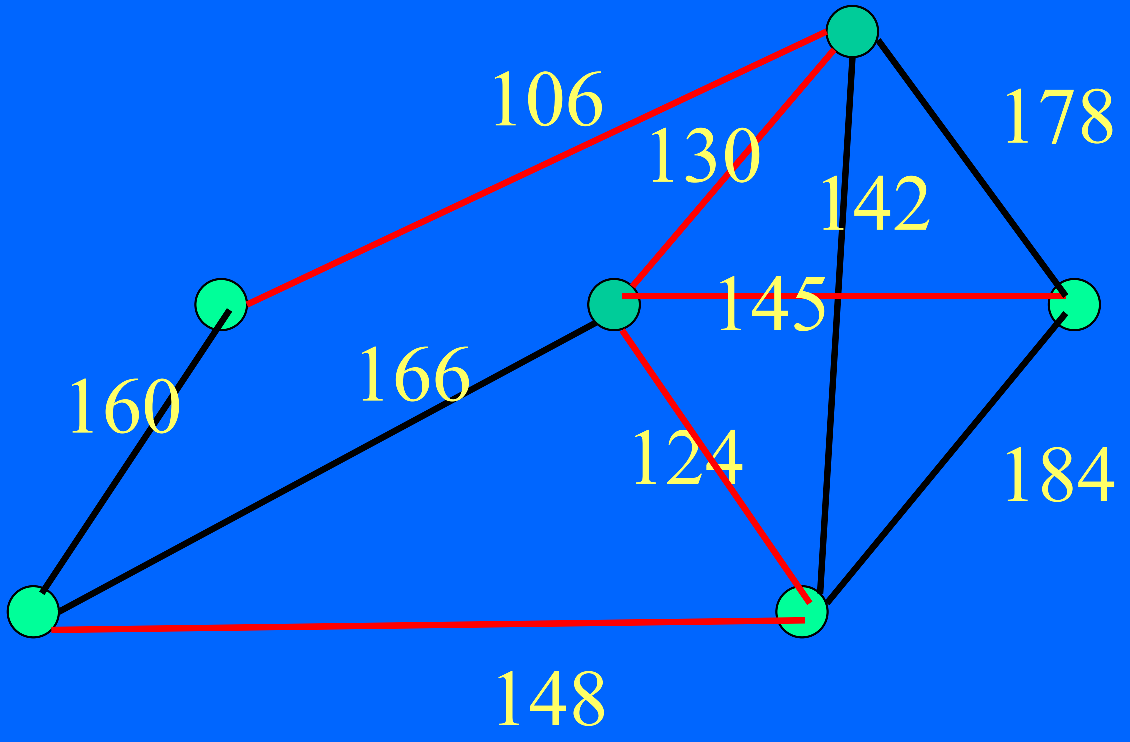
- The minimum spanning tree is the same.

Related Example: Minimum Spanning Tree Problem



- Is this an accident?
- No: By Kruskal's algorithm for finding the minimum spanning tree, even an ordinal transformation will leave the minimum spanning tree unchanged.

Related Example: Minimum Spanning Tree Problem



- Kruskal's algorithm:
  - ✓ Order edges by weight.
  - ✓ At each step, pick least-weight edge that does not create a cycle with previously chosen edges.

Related Example: Minimum Spanning Tree Problem

- Many practical decision making problems involve the search for an optimal solution as in Shortest Path and Minimum Spanning Tree.
- ***Little attention is paid to the possibility that conclusion that a particular solution is optimal may be an accident of the way things are measured.***

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Average Machine Performance
4. Normalized Performance Scores
5. Optimization Problems for Intelligent Machines
6. **How to Average Scores**
7. Meaningfulness of Statistical Tests

How Should We Average Scores?

- Sometimes arithmetic means are not a good idea.
- Sometimes geometric means are.

- Are there situations where the opposite is the case? Or some other method is better?

- ***Can we lay down some guidelines about when to use what averaging or merging procedure?***

- Let $a_1, a_2, \ldots, a_n$ be "scores" or ratings, e.g., scores on criteria for evaluating machines.
- Let $u = F(a_1,a_2, \ldots, a_n)$
- F is an unknown averaging function – sometimes called a ***merging function***, and u is the average or merged score.

How Should We Average Scores?

An Axiomatic Approach

Theorem (Fleming and Wallace). Suppose $F:(\mathfrak{R}^+)^n \to \mathfrak{R}^+$ has the following properties:

(1). ***Reflexivity***: $F(a,a,...,a) = a$

(2). ***Symmetry***: $F(a_1,a_2,...,a_n) = F(a_{\pi(1)},a_{\pi(2)}, ...,a_{\pi(n)})$ for all permutations π of $\{1,2,…,n\}$

(3). ***Multiplicativity***:

$F(a_1b_1,a_2b_2,…,a_nb_n) = F(a_1,a_2,...,a_n)\ F(b_1,b_2,...,b_n)$

Then F is the geometric mean. And conversely.

How Should We Average Scores?

Unknown function $u = F(a_1, a_2, \ldots, a_n)$

Luce's idea ("***Principle of Theory Construction***"): If you know the scale types of the a_i and the scale type of u and you assume that an admissible transformation of each of the a_i leads to an admissible transformation of u, you can derive the form of F.

(We will disregard some of the restrictions on applicability of this principle, including those given by Luce.)

This gets us into functional equations.

How Should we Average Scores?

A Functional Equations Approach

Example: $u = F$(a). Assume a and u are ratio scales.

- Admissible transformations of scale: multiplication by a positive constant.
- Multiplying the independent variable by a positive constant α leads to multiplying the dependent variable by a positive constant A that depends on α.
- This leads to the functional equation:

$$(\&) \qquad F(\alpha a) = A(\alpha)F(a),\ A(\alpha) > 0.$$

How Should we Average Scores?

- This leads to the functional equation:

(&) $$F(\alpha a) = A(\alpha)F(a),\ \alpha > 0,\ A(\alpha) > 0.$$

By solving this functional equation, Luce proved the following theorem:

Theorem (Luce 1959): Suppose the averaging function F is continuous and suppose a takes on all positive real values and F takes on positive real values. Then

$$F(a) = ca^k$$

Thus, if both the independent and dependent variables are ratio scales, the only possible way to relate them is by a power law.

The Possible Scientific Laws

- This result is very general.
- It can be interpreted as limiting in very strict ways the ***"possible scientific laws"***
- Other examples of power laws:
  - $V = (4/3)\pi r^3$ Volume V, radius r are ratio scales
  - **Newton's Law of gravitation**: $F = G(mm^*/r^2)$, where F is force of attraction, G is gravitational constant, m, m^* are fixed masses of bodies being attracted, r is distance between them.
  - **Ohm's Law**: Under fixed resistance, voltage is proportional to current (voltage, current are ratio scales)

How Should We Average Scores?

A Functional Equations Approach

Example: a_1, a_2, ..., a_n are independent ratio scales, u is a ratio scale.

$F: (\mathcal{R}^+)^n \rightarrow \mathcal{R}^+$

$F(a_1,a_2,\ldots,a_n) = u \rightarrow F(\alpha_1 a_1,\alpha_2 a_2,\ldots,\alpha_n a_n) = \alpha u,$

$\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_n > 0$, $\alpha > 0$, α depends on a_1, a_2, …, a_n.

- Thus we get the functional equation:

(*) $F(\alpha_1 a_1,\alpha_2 a_2,\ldots,\alpha_n a_n) = A(\alpha_1,\alpha_2,\ldots,\alpha_n)F(a_1,a_2,\ldots,a_n),$

$A(\alpha_1,\alpha_2,\ldots,\alpha_n) > 0$

How Should We Average Scores?

A Functional Equations Approach

(*) $F(\alpha_1 a_1, \alpha_2 a_2, \ldots, \alpha_n a_n) = A(\alpha_1, \alpha_2, \ldots, \alpha_n) F(a_1, a_2, \ldots, a_n)$,

$A(\alpha_1, \alpha_2, \ldots, \alpha_n) > 0$

Theorem (Luce 1964): If $F: (\Re^+)^n \rightarrow \Re^+$ is continuous and satisfies (*), then there are $\lambda > 0$, c_1, c_2, …, c_n so that

$$F(a_1, a_2, \ldots, a_n) = \lambda a_1^{c_1} a_2^{c_2} \ldots a_n^{c_n}$$

λ, c_1, c_2, …, c_n constants

How Should We Average Scores?

$$F(a_1,a_2,\ldots,a_n) = \lambda a_1^{c_1} a_2^{c_2} \ldots a_n^{c_n}$$

Theorem (Aczél and Roberts 1989): If in addition F satisfies reflexivity and symmetry, then $\lambda = 1$ and $c_1 = c_2 = \ldots = c_n = 1/n$, so F is the geometric mean.

How Should We Average Scores?

Sometimes You Get the Arithmetic Mean

Example: $a_1, a_2, \ldots, a_n$ interval scales with the same unit and independent zero points; u an interval scale.

Functional Equation:

$$(****) \quad F(\alpha a_1+\beta_1,\alpha a_2+\beta_2,\ldots,\alpha a_n+\beta_n) = A(\alpha,\beta_1,\beta_2,\ldots,\beta_n)F(a_1,a_2,\ldots,a_n) + B(\alpha,\beta_1,\beta_2,\ldots,\beta_n)$$

$$A(\alpha,\beta_1,\beta_2,\ldots,\beta_n) > 0$$

How Should We Average Scores?

Functional Equation:

(****) $F(\alpha a_1+\beta_1,\alpha a_2+\beta_2,\ldots,\alpha a_n+\beta n) =$
$A(\alpha,\beta_1,\beta_2,\ldots,\beta_n)F(a_1,a_2,\ldots,a_n) + B(\alpha,\beta_1,\beta_2,\ldots,\beta_n)$

$$A(\alpha,\beta_1,\beta_2,\ldots,\beta_n) > 0$$

Solutions to (****) (Even Without Continuity Assumed)
(Aczél, Roberts, and Rosenbaum):

$$F(a_1,a_2,\ldots,a_n) = \sum_{i=1}^{n} \lambda_i a_i + b$$

$\lambda_1, \lambda_2, \ldots, \lambda_n, b$ arbitrary constants

How Should We Average Scores?

$$F(a_1,a_2,\ldots,a_n) = \sum_{i=1}^{n} \lambda_i a_i + b$$

Theorem (Aczél and Roberts):

(1). If in addition F satisfies reflexivity, then

$$\Sigma \lambda_i = 1, b = 0$$

(2). If in addition F satisfies reflexivity and symmetry, then $\lambda_i = 1/n$ for all i, and $b = 0$, i.e., F is the arithmetic mean.

How Should We Average Scores?

Meaningfulness Approach

- While it is often reasonable to assume you know the scale type of the independent variables $a_1, a_2, \ldots, a_n$, it is not so often reasonable to assume that you know the scale type of the dependent variable u.

- However, it turns out that you can replace the assumption that the scale type of u is xxxxxxx by the assumption that a certain statement involving u is meaningful.

How Should We Average Scores?

Back to Earlier Example: $a_1, a_2, \ldots, a_n$ are independent ratio scales. Instead of assuming u is a ratio scale, assume that the statement

$$F(a_1,a_2, \ldots, a_n) = kF(b_1, b_2, \ldots, b_n)$$

is meaningful for all $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ and $k > 0$. Then we get the same results as before:

Theorem (Roberts and Rosenbaum 1986): Under these hypotheses and continuity of F,

$$\mathrm{F}(a_1,a_2,\ldots,a_\mathrm{n}) = \lambda a_1^{c_1} a_2^{c_2} \ldots a_\mathrm{n}^{c_\mathrm{n}}$$

If in addition F satisfies reflexivity and symmetry, then F is the geometric mean.

Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Average Machine Performance
4. Normalized Performance Scores
5. Optimization Problems for Intelligent Machines
6. How to Average Scores
7. **Meaningfulness of Statistical Tests**

Meaningfulness of Statistical Tests

(joint work with Helen Marcus-Roberts)

•For > 50 years: considerable disagreement on limitations scales of measurement impose on statistical procedures we may apply.

•Controversy stems from Stevens (1946, 1951, 1959, ...):

- Foundational work
- Developed the classification of scales of measurement
- Provided rules for the use of statistical procedures: certain statistics are inappropriate at certain levels of measurement.

Meaningfulness of Statistical Tests

- The application of Stevens' ideas to ***descriptive statistics*** has been widely accepted
- Application to ***inferential statistics*** has been labeled by some a *misconception*.

Meaningfulness of Statistical Tests: Descriptive Statistics

- P = population whose distribution we would like to describe.
- Capture properties of P by finding a descriptive statistic for P or taking a sample S from P and finding a descriptive statistic for S.
- Our examples suggest: certain descriptive statistics appropriate only for certain measurement situations.
- This idea originally due to Stevens.
- Popularized by Siegel in his well-known book *Nonparametric Statistics* (1956).

Meaningfulness of Statistical Tests: Descriptive Statistics

- Our examples suggest the principle: Arithmetic means are "appropriate" statistics for interval scales, medians for ordinal scales.
- Other side of the coin: It is argued that it is *always* appropriate to calculate means, medians, and other descriptive statistics, no matter what the scale of measurement.

Frederic Lord: Famous football player example.

"The numbers don't remember where they came from."

Meaningfulness of Statistical Tests: Descriptive Statistics

- I agree: It is *always* appropriate to *calculate* means, medians, ...
- But: Is it appropriate to make certain statements using these descriptive statistics?

Meaningfulness of Statistical Tests: Descriptive Statistics

•My position: ***It is usually appropriate to make a statement using descriptive statistics iff the statement is meaningful.***

•A statement that is true but meaningless gives information that is an accident of the scale of measurement used, not information that describes the population in some fundamental way.

•So, it is appropriate to calculate the mean of ordinal data

•It is just not appropriate to say that the mean of one group is higher than the mean of another group.

Meaningfulness of Statistical Tests: Inferential Statistics

•Stevens' ideas have come to be applied to inferential statistics -- inferences about an unknown population *P*.

•They have led to such principles as the following:

(1). Classical parametric tests (e.g., t-test, Pearson correlation, analysis of variance) are inappropriate for ordinal data. They should be applied only to data that define an interval or ratio scale.

Meaningfulness of Statistical Tests: Inferential Statistics

(2). For ordinal scales, non-parametric tests (e.g., Mann-Whitney U, Kruskal-Wallis, Kendall's tau) can be used.

Not everyone agrees. Thus: <u>Controversy</u>

Meaningfulness of Statistical Tests: Inferential Statistics

My View:

- The validity of a statistical test depends on a ***statistical model***
  - This includes information about the distribution of the population and about the sampling procedure.
- The validity of the test does not depend on a ***measurement model***
  - This is concerned with the admissible transformations and scale type.

Meaningfulness of Statistical Tests: Inferential Statistics

•***The scale type enters in deciding whether the hypothesis is worth testing at all -- is it a meaningful hypothesis?***

•The issue is: If we perform admissible transformations of scale, is the truth or falsity of the hypothesis unchanged?

•Example: Ordinal data. Hypothesis: Mean is 0. Conclusion: This is a meaningless hypothesis.

Meaningfulness of Statistical Tests: Inferential Statistics

•Can we test meaningless hypotheses?
•Sure. But I question what information we get outside of information about the population as measured.

More details: Testing H_0 about *P* :
1). Draw a ***random sample*** *S* from *P*.
2). Calculate a ***test statistic*** based on *S*.
3). Calculate probability that the test statistic is what was observed given H_0 is true.
4). Accept or reject H_0 on the basis of the test.

Meaningfulness of Statistical Tests: Inferential Statistics

- Calculation of probability depends on a ***statistical model***, which includes information about the distribution of P and about the sampling procedure.
- But, validity of the test depends only on the statistical model, not on the measurement model.

Meaningfulness of Statistical Tests: Inferential Statistics

- Thus, you can apply parametric tests to ordinal data, provided the statistical model is satisfied.
- Model satisfied if the data is normally distributed.
- Where does the scale type enter?
- In determining if the hypothesis is worth testing at all. i.e., if it is meaningful.

Meaningfulness of Statistical Tests: Inferential Statistics

- For instance, consider ordinal data and

$$H_0\text{: mean is 0}$$

- The hypothesis is meaningless.
- But, if the data meets certain distributional requirements such as normality, we can apply a parametric test, such as the t-test, to check if the mean is 0.

Closing Comments

- Meaningfulness of statistical tests: There are considerable limitations that meaningfulness places on conclusions from statistical tests
  - Descriptive statistics – reasonably well accepted
  - Inferential statistics – considerable “discussion”

Closing Comments

Message: Do not perform arithmetic operations on data without paying attention to whether the conclusions you get are meaningful.

Questions:
froberts@dimacs.rutgers.edu

Credit: Toby Hudson, wikimedia commons (no changes)