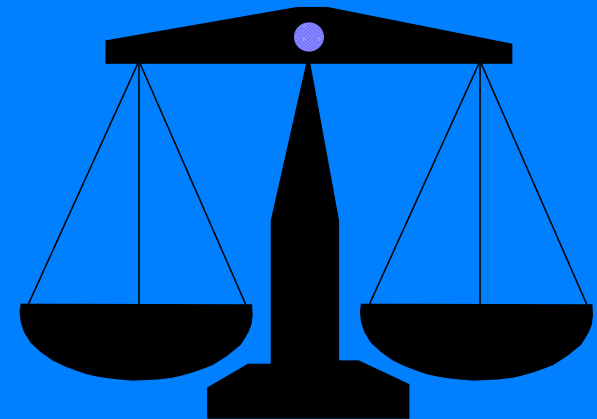


Measurement of Pollution

Fred Roberts, DIMACS, Rutgers
University



Some Questions We Will Ask

- There are many pollutants in the air.
- Is it possible to find one combined index of air pollution that takes into account all of them?



Some Questions We Will Ask

- Is an airplane louder than a motorcycle?
- Is it noisier?
- What is the difference?



Some Questions We Will Ask

- Given two devices to measure changes in water pollution level, which one does a better job?



MEASUREMENT

- We will observe that all of these questions have something to do with measurement.
- The answers are very relevant to public and private sector decision making.
- We will apply measurement theory to measurement of air, water, and noise pollution.



Outline

1. **Introduction to Measurement Theory**
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

MEASUREMENT

- *Measurement* has something to do with numbers.



- *Our approach: “Representational theory of measurement”*

- Assign numbers to “objects” being measured in such a way that certain empirical relations are “preserved.”

- In measurement of temperature, we preserve a relation “warmer than.”



- In measurement of mass, we preserve a relation “heavier than.”



MEASUREMENT

A : Set of Objects

R : Binary relation on A

aRb  a is “warmer than” b

aRb  a is “heavier than” b

$f: A$  

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

R could be *preference*. Then f is a *utility function (ordinal utility function)*.

R could be “louder than.” Then f is a measure of loudness.

MEASUREMENT

A : Set of Objects


R : Binary relation on A

aRb  a is “warmer than” b

aRb  a is “heavier than” b

$f: A$  

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

With *mass*, there is more going on. *There is an operation of combination of objects and mass is additive.* a  b means a combined with b .

$$f(a \text{  } b) = f(a) + f(b).$$

MEASUREMENT

- This can all be generalized using a formalism called a *homomorphism*.
- It will suffice to think of a homomorphism as a way of assigning numbers to objects being measured so that certain relations and operations among objects are reflected in comparable relations among the assigned numbers.
- Even more basically: Homomorphisms will be “acceptable” ways to assign numbers.
- We will be particularly interested in finding ways to transform one homomorphism (acceptable way to measure) into another.

Homomorphisms: A Formalism

- **Empirical Relational System** A

Set of objects A and relations R and operations $\boxed{\mathbb{W}}$ on A .

- **Numerical Relational System** B

Set of objects B where B is a set of real numbers, plus a relation R^* corresponding to each R on A and an operation $\boxed{\mathbb{W}}^*$ corresponding to each $\boxed{\mathbb{W}}$ on A .

- **Homomorphism** from A into B

A function $f: A \rightarrow B$ such that all relations and operations among elements in A are reflected in corresponding relations and operations among elements in B , e.g.,

$$aRb \iff f(a)R^*f(b)$$

$$f(a \boxed{\mathbb{W}} b) = f(a) \boxed{\mathbb{W}}^* f(b).$$

Outline

1. Introduction to Measurement Theory
- 2. Theory of Uniqueness of Scales of Measurement/
Scale Types**
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

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 homomorphism.

- Suppose f is a homomorphism from A into B .

- $f(A)$  B is called an *admissible transformation of f* if $f(A)$  f is again a homomorphism from A  B into B .

The Theory of Uniqueness

Admissible Transformations

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

Kilograms  Pounds:  $(x) = 2.2x$



The Theory of Uniqueness

- 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

<u>Class of Adm. Transfs.</u>	<u>Scale Type</u>	<u>Example</u>
$\boxed{W}(x) = \boxed{W}x, \boxed{W} > 0$	<i>ratio</i>	Mass Temp. (Kelvin) Time (intervals) Loudness (sones)? Brightness (brils)?
$\boxed{W}(x) = \boxed{W}x + \boxed{W}, \boxed{W} > 0$	<i>interval</i>	Temp (F,C) Time (calendar) IQ tests (standard

Some Common Scale Types

Class of Adm. Transfs.	Scale Type	Example
------------------------	------------	---------

$x \begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} y \begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} \begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} (x) \begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} \begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} (y)$		
--	--	--

$\begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix}$ strictly increasing		
--	--	--

ordinal

Preference?

Hardness

Grades of leather,

IQ tests (raw

$\begin{matrix} \text{PRIORITY USE} \\ \text{W} \\ \text{PRIORITY USE} \end{matrix} (x) = x$		
--	--	--

absolute

Counting

Outline

1. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
- 3. Meaningful Statements**
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

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.

Definition: 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 to 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

“This talk will be three times as long as the next talk.”

• Is this meaningful?

Meaningful Statements

“This talk will be three times as long as the next talk.”

•Is this meaningful?

I hope not!

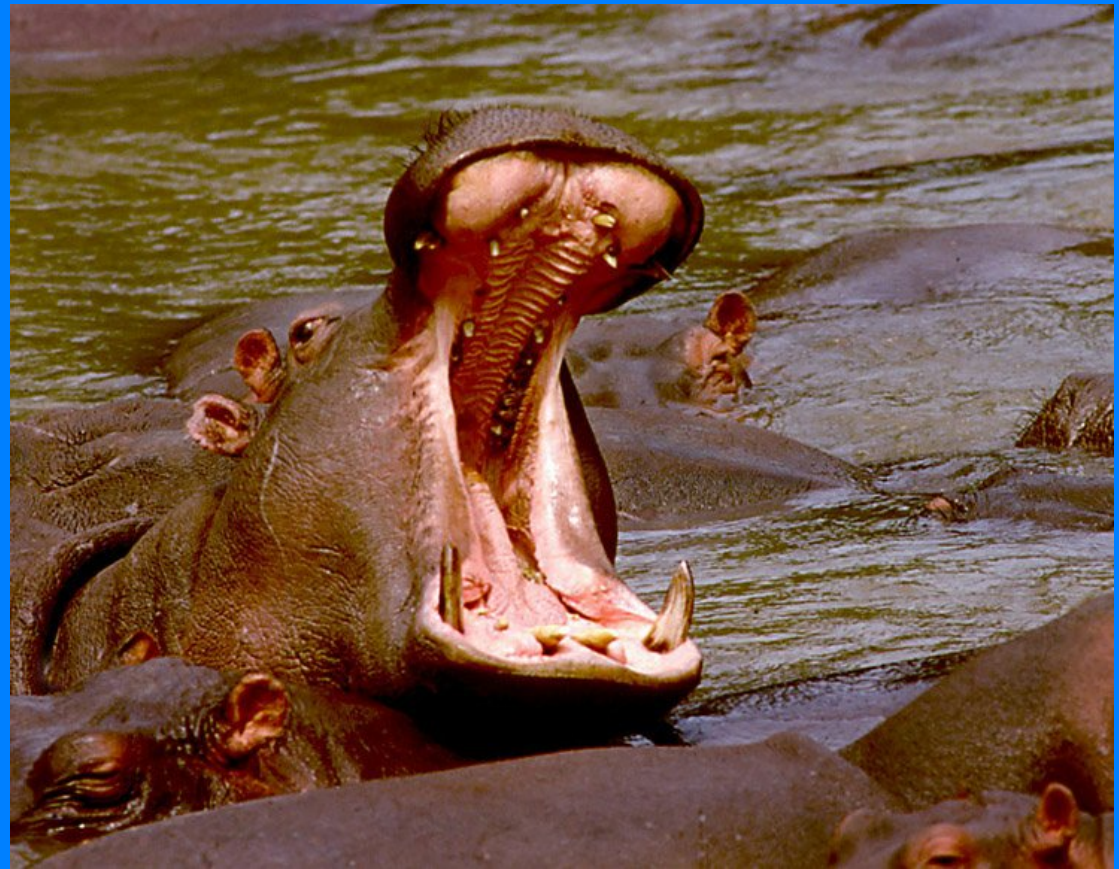


Meaningful Statements

“This talk will be three times as long as the next talk.”

• Is this meaningful?

Me too



Meaningful Statements

“This talk will be three times as long as the next talk.”

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

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

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

$$(2) \quad (\mathbb{W}\mathbb{W})f(a) = 3(\mathbb{W}\mathbb{W})f(b)$$

- But (2) becomes

$$(3) \quad \mathbb{W}f(a) = 3\mathbb{W}f(b)$$

- (1) \mathbb{W} (3) since $\mathbb{W} > 0$.

Meaningful Statements

“The high temperature today was five percent higher than the high temperature yesterday.”

- Is this meaningful?



Meaningful Statements

“The high temperature today was five percent higher than the high temperature yesterday.”

$$f(a) = 1.05f(b)$$

- 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 the Statue of Liberty weighs.”

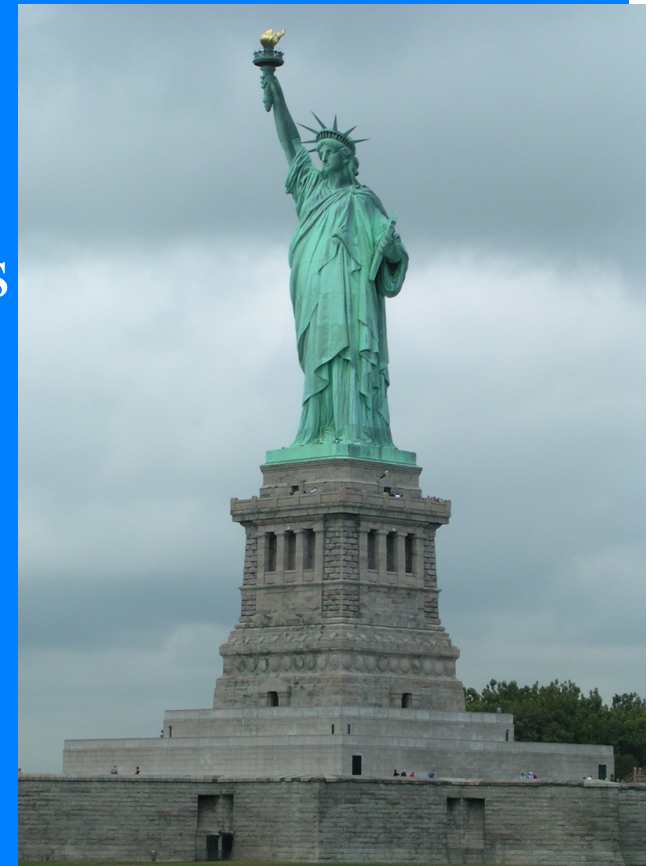
- Is this meaningful?



Meaningful Statements

“I weigh 1000 times what the Statue of Liberty 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. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
- 4. Averaging Judgments of Loudness**
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

Average Loudness

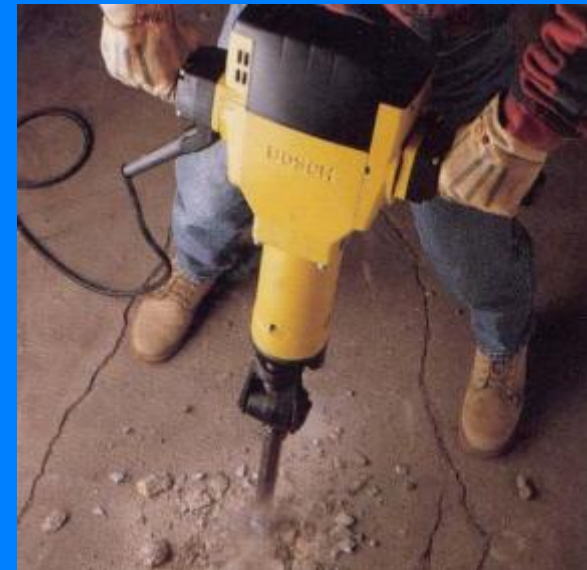
- Study two groups of machines.
- $f(a)$ is the loudness of machine a .
- **Data suggests that the average loudness of machines in the first group is higher than the average loudness of machines in the second group.**

a_1, a_2, \dots, a_n machines in first group

b_1, b_2, \dots, b_m machines in second group.

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

- We are comparing *arithmetic means*.



Average Loudness

•Statement (1) is meaningful iff for all admissible transformations of scale $\left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right]$, (1) holds iff

$$(2) \quad \frac{1}{\bar{n}} \sum_{i=1}^n \pm f(a_i) > \frac{1}{\bar{m}} \sum_{i=1}^m \pm f(b_i)$$

•*Some argue that loudness (sones) define a ratio scale.*
(More on this later.)

•Thus, $\left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right](x) = \left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right]x$, $\left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right] > 0$, so (2) becomes

$$(3) \quad \frac{1}{\bar{n}} \sum_{i=1}^n \textcircled{R}f(a_i) > \frac{1}{\bar{m}} \sum_{i=1}^m \textcircled{R}f(b_i)$$

•Then $\left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right] > 0$ implies (1) $\left[\begin{smallmatrix} \text{PROHIBIT USE} \\ \mathbb{W} \\ \text{PROHIBIT USE} \end{smallmatrix} \right]$ (3). Hence, (1) is meaningful

Average Loudness

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

$$(4) \quad \frac{1}{n} \sum_{i=1}^n [\mathbb{R}f(a_i) + \mathbb{W}] > \frac{1}{m} \sum_{i=1}^m [\mathbb{R}f(b_i) + \mathbb{W}]$$

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

Average Loudness

- To show that comparison of arithmetic means can be meaningless for ordinal scales, suppose we are asking experts for a subjective judgment of loudness.
- Suppose $f(a)$ is measured on an ordinal scale, e.g., 5-point scale: 5=extremely loud, 4=very loud, 3=loud, 2=slightly loud, 1=quiet.
- 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 loudness of group 2 machines is higher.

Average Loudness

- Suppose $f(a)$ is measured on an ordinal scale, e.g., 5-point scale: 5=extremely loud, 4=very loud, 3=loud, 2=slightly loud, 1=quiet.
- 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 loudness of group 1 machines is higher.³⁶

Average Loudness

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

Average Loudness

- **Suppose each of n individuals is asked to rate each of a collection of alternative machines as to their relative loudness.**
- Or we rate alternatives on different criteria or against different benchmarks. (Similar results with performance ratings, importance ratings, brightness ratings, etc.)
- Let $f_i(a)$ be the rating of alternative a by individual i (under criterion i). Is it meaningful to assert that the average rating of alternative a is higher than the average rating of alternative b ?

Average Loudness

- Let $f_i(a)$ be the rating of alternative a by individual i (under criterion i). Is it meaningful to assert that the average rating of alternative a is higher than the average rating of alternative b ?
- A similar question arises in performance ratings, brightness ratings, importance ratings, etc.

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

Average Loudness

- If each f_i is a ratio scale, then we consider for $\left[\begin{smallmatrix} \text{PRIORITY USE} \\ \mathbb{W} \\ \text{PRIORITY USE} \end{smallmatrix} \right] > 0$,

$$(2) \quad \frac{1}{\bar{n}} \sum_{i=1}^n \textcircled{R}f_i(a) > \frac{1}{\bar{n}} \sum_{i=1}^n \textcircled{R}f_i(b)$$

- Clearly, (1) $\left[\begin{smallmatrix} \text{PRIORITY USE} \\ \mathbb{W} \\ \text{PRIORITY USE} \end{smallmatrix} \right]$ (2), so (1) is meaningful.

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

$$(3) \quad \frac{1}{\bar{n}} \sum_{i=1}^n \textcircled{R}_i f_i(a) > \frac{1}{\bar{n}} \sum_{i=1}^n \textcircled{R}_i f_i(b)$$

- It is easy to see that there are $\left[\begin{smallmatrix} \text{PRIORITY USE} \\ \mathbb{W} \\ \text{PRIORITY USE} \end{smallmatrix} \right]_i$ so that (1) holds and (3) fails. Thus, (1) is meaningless.

Average Loudness

Motivation for considering different  i :

$n = 2$, $f_1(a) = \text{weight of } a$, $f_2(a) = \text{height of } a$. 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.



Average Loudness

Motivation for considering different  i :

$n = 2$, $f_1(a) = \text{weight of } a$, $f_2(a) = \text{height of } a$. 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 Loudness

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

$$\sqrt[n]{\prod_{i=1}^n f_i(a)} > \sqrt[n]{\prod_{i=1}^n f_i(b)} \quad \& \quad \sqrt[n]{\prod_{i=1}^n \textcircled{R} f_i(a)} > \sqrt[n]{\prod_{i=1}^n \textcircled{R} f_i(b)}$$

all $\textcircled{W}_i > 0$.

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

Application of this Idea



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 energy demand 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.

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

<u>Variable</u>	<u>Rel. Import. Rating</u>
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. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
- 5. Measurement of Air Pollution: A Combined Pollution Index**
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

MEASUREMENT OF AIR POLLUTION



MEASUREMENT OF AIR POLLUTION

- Various pollutants are present in the air:
- Carbon monoxide (CO), hydrocarbons (HC), nitrogen oxides (NOX), sulfur oxides (SOX), particulate matter (PM).
- Also damaging: Products of chemical reactions among pollutants. E.g.: Oxidants such as ozone produced by HC and NOX reacting in presence of sunlight.
- Some pollutants are more serious in presence of others, e.g., SOX are more harmful in presence of PM.
- *Can we measure pollution with one overall measure?*

MEASUREMENT OF AIR POLLUTION

- To compare pollution control policies, need to compare effects of different pollutants. We might allow increase of some pollutants to achieve decrease of others.
- **One single measure could give indication of how bad pollution level is and might help us determine if we have made progress.**

Combining Weight of Pollutants:

- Measure total weight of emissions of pollutant i over fixed period of time and sum over i .

$e(i,t,k)$ = total weight of emissions of pollutant i (per cubic meter) over t th time period and due to k th source or measured in k th location.

$$A(t;k) = \sum_i e(i;t;k)$$

MEASUREMENT OF AIR POLLUTION

- Early uses of this simple index A in the early 1970s led to the conclusions:
 - (A) Transportation is the largest source of air pollution, with stationary fuel combustion (especially by electric power plants) second largest.
 - (B) Transportation accounts for over 50% of all air pollution.
 - (C) CO accounts for over half of all emitted air pollution.
- Are these meaningful conclusions?

MEASUREMENT OF AIR POLLUTION

- Early uses of this simple index A in the early 1970s led to the conclusions:

(A) Transportation is the largest source of air pollution, with stationary fuel combustion (especially by electric power plants) second largest.

- Are these meaningful conclusions?

$$A(t; k) > A(t; k^0)$$



MEASUREMENT OF AIR POLLUTION

- Early uses of this simple index A in the early 1970s led to the conclusions:

(B) Transportation accounts for over 50% of all air pollution.

- Are these meaningful conclusions?

$$A(t; k_r) > \sum_{k \in k_r} A(t; k)$$



MEASUREMENT OF AIR POLLUTION

- Early uses of this simple index A in the early 1970s led to the conclusions:

(C) CO accounts for over half of all emitted air pollution.

- Are these meaningful conclusions?

$$\sum_{t;k} X_{t;k} e(i;t;k) > \sum_{t;k} \sum_{j \neq i} X_{t;k} e(j;t;k)$$



MEASUREMENT OF AIR POLLUTION

$$A(t; k) > A(t; k^0)$$

$$A(t; k_r) > X A(t; k)$$

$$X e(i; t; k) > \frac{k \in k_r}{X} X e(j; t; k)$$

$t; k$ $t; k$ $j \in i$

All these conclusions are meaningful if we measure all $e(i, t, k)$ in same units of mass (e.g., milligrams per cubic meter) and so admissible transformation means multiply $e(i, t, k)$ by same constant.

MEASUREMENT OF AIR POLLUTION

- *These comparisons are meaningful in the technical sense.*
- *But: Are they meaningful comparisons of pollution level in a practical sense?*
- A unit of mass of CO is far less harmful than a unit of mass of NOX. U.S. Environmental Protection Agency standards based on health effects for 24 hour period allow 7800 units of CO to 330 units of NOX.
- These are *Minimum acute toxicity effluent tolerance factors* (MATE criteria).
- *Tolerance factor* is level at which adverse effects are known. Let $\left[\frac{W}{W} \right] (i)$ be tolerance factor for i th pollutant.
- *Severity factor*: $\left[\frac{W}{W} \right] (\text{CO}) / \left[\frac{W}{W} \right] (i)$ or $1 / \left[\frac{W}{W} \right] (i)$

MEASUREMENT OF AIR POLLUTION

- One idea (Babcock and Nagda, Walther, Caretto and Sawyer): Weight the emission levels (in mass) by severity factor and get a weighted sum. This amounts to using the indices

Degree of hazard: $\frac{1}{\bar{z}(t)} e(i; t; k)$

and the combined index

Pindex: $B(t; k) = \sum_i \frac{1}{\bar{z}(t)} e(i; t; k)$

- Under pindex, transportation is still the largest source of pollutants, but now accounting for less than 50%. Stationary sources fall to fourth place. CO drops to bottom of list of pollutants, accounting for just over 2% of the total.

MEASUREMENT OF AIR POLLUTION

- These conclusions are again meaningful if all emission weights are measured in the same units. For an admissible transformation multiplies $\left[\frac{W}{e} \right]$ and e by the same constant and thus leaves the degree of hazard unchanged and pindex unchanged.

- Pindex was introduced in the San Francisco Bay Area in the 1960s.



- *But, are comparisons using pindex meaningful in the practical sense?*

MEASUREMENT OF AIR POLLUTION

- Pindex amounts to: For a given pollutant, take the percentage of a given harmful level of emissions that is reached in a given period of time, and add up these percentages over all pollutants. (Sum can be greater than 100% as a result.)
- If 100% of the CO tolerance level is reached, this is known to have some damaging effects. Pindex implies that the effects are equally severe if levels of five major pollutants are relatively low, say 20% of their known harmful levels.

MEASUREMENT OF AIR POLLUTION

- *Severity tonnage* of pollutant i due to a given source is actual tonnage times the severity factor $1/W(i)$.
- In early air pollution measurement literature, severity tonnage was considered a measure of how severe pollution due to a source was.
- Data from Walther 1972 suggests the following. Interesting exercise to decide which of these conclusions are meaningful.



MEASUREMENT OF AIR POLLUTION

1. HC emissions are more severe (have greater severity tonnage) than NOX emissions.
2. Effects of HC emissions from transportation are more severe than those of HC emissions from industry. (Same for NOX.).
3. Effects of HC emissions from transportation are more severe than those of CO emissions from industry.
4. Effects of HC emissions from transportation are more than 20 times as severe as effects of CO emissions from transportation.
5. The total effect of HC emissions due to all sources is more than 8 times as severe as total effect of NOX emissions due to all sources.

Outline

1. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
- 6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”**
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

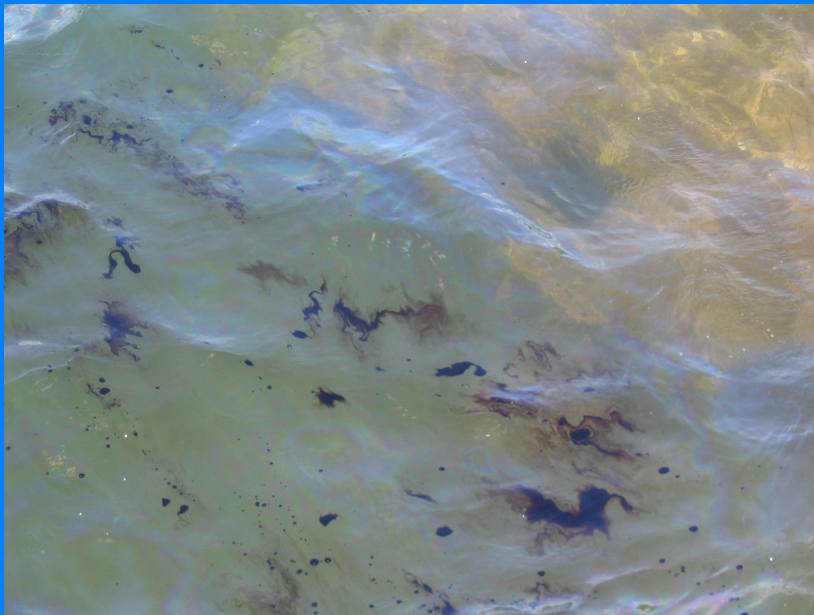
Evaluation of Water Testing Equipment

- How do we evaluate alternative possible water testing systems? Or pollution control systems for oil, chemicals, ...
- A number of systems are tested on different benchmarks.
- Their scores on each benchmark are normalized relative to the score of one of the systems.
- The normalized scores of a system are combined by some averaging procedure.
- If the averaging is the arithmetic mean, then the statement **“one system has a higher arithmetic mean normalized score than another system”** is meaningless:
The system to which scores are normalized can determine which has the higher arithmetic mean.



Evaluation of Water Testing Equipment

- Similar methods are used in comparing performance of alternative computer systems or other types of machinery.
- The following example has numbers taken out of the computer science literature from an article comparing computer systems. However, the same applies to pollution control equipment or other types of equipment.
- The example is due to Fleming and Wallace (1986).



Equipment Evaluation

Evaluation of Water Testing Equipment

BENCHMARK

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

Data from Heath, *Comput. Archit. News* (1984)

Equipment Evaluation

Normalize Relative to System R

BENCHMARK

		E	F	G	H	I
SYSTEM	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

Equipment Evaluation

Take Arithmetic Mean of Normalized Scores

		BENCHMARK					Arithmetic Mean
		E	F	G	H	I	
S Y S T E M	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

Equipment Evaluation

Take Arithmetic Mean of Normalized Scores

		BENCHMARK					Arithmetic Mean
		E	F	G	H	I	
SYSTEM	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 system Z is best

Equipment Evaluation

Now Normalize Relative to System M

BENCHMARK

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

Equipment Evaluation

Take Arithmetic Mean of Normalized Scores

BENCHMARK

Arithmetic
Mean

		E	F	G	H	I	
		S Y S T E M	R	417 1.71	83 1.19	66 .43	
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

Equipment Evaluation

Take Arithmetic Mean of Normalized Scores

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

Conclude that system R is best

Equipment Evaluation

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

$$\sqrt[n]{\prod_{i=1}^n S(x_i)}$$

Equipment Evaluation

Normalize Relative to System R

BENCHMARK

Geometric Mean

S
Y
S
T
E
M

R

M

Z

	E	F	G	H	I		
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 system R is best

Equipment Evaluation

Now Normalize Relative to System M

		BENCHMARK					Geometric Mean
		E	F	G	H	I	
SYSTEM	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	
	Z	134 .55	70 1.00	135 .88	66,000 1.97	369 1.00	

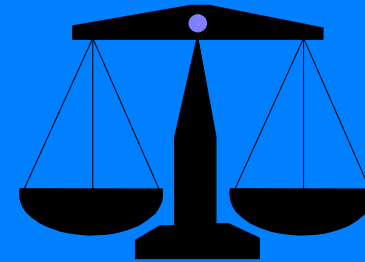
Still conclude that system R is best

Equipment Evaluation

- In this situation, it is easy to show that *the conclusion that a given system has highest geometric mean normalized score is a meaningful conclusion.*
- *Even meaningful: A given system 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.

Equipment Evaluation

Message from measurement theory:



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

Equipment Evaluation

- We have seen that in some situations, comparing arithmetic means is not a good idea and comparing geometric means is.
- We will see that there are situations where the reverse is true.
- Can we lay down some guidelines as to when to use what averaging procedure?
- More on this later.

Outline

1. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
- 7. Optimization Problems in Pollution Measurement**
8. Measurement of Noise: Introduction to Psychophysical Scaling
9. How to Average Scores

Climate Change and Pollution

- Some early warning signs of climate change include extreme heat events, commonly associated with air pollution events:

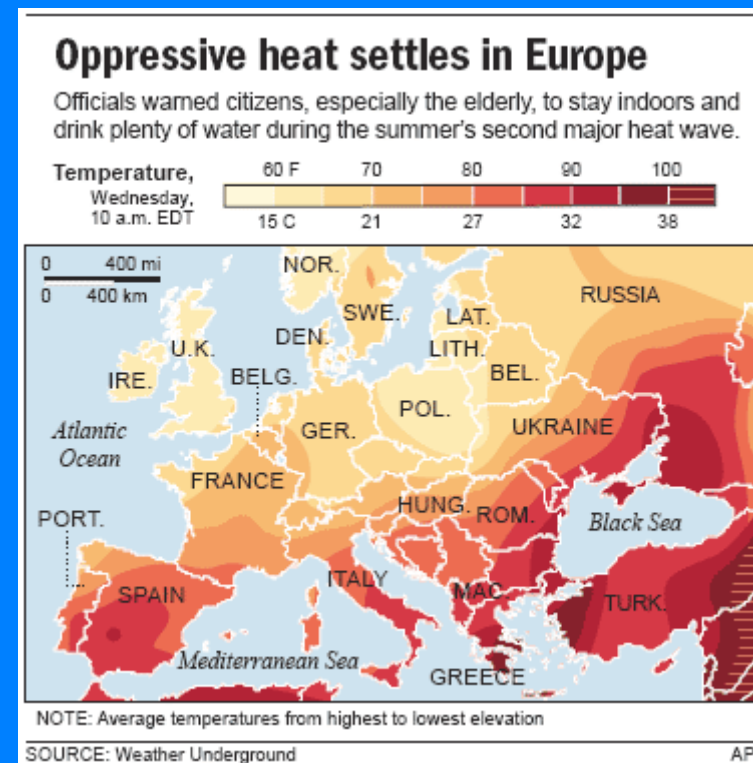
- 1995 extreme heat event in Chicago

- 514 heat-related deaths

- 3300 excess emergency admissions

- 2003 heat wave in Europe

- 35,000 deaths



Extreme Events due to Global Warming

- We anticipate an increase in number and severity of extreme events due to global warming.
- More heat waves and associated air pollution events.
- More floods, hurricanes.



Extreme Pollution Events: Evacuation

- One response to such extreme pollution events: evacuation of most vulnerable individuals to climate controlled environments.
- Modeling challenges:
 - Where to locate the evacuation centers?
 - Whom to send where?
 - Goals include minimizing travel time, keeping facilities to their maximum capacity, etc.
 - Relevance of mathematical tools of operations research – location theory, assignment problems, etc.

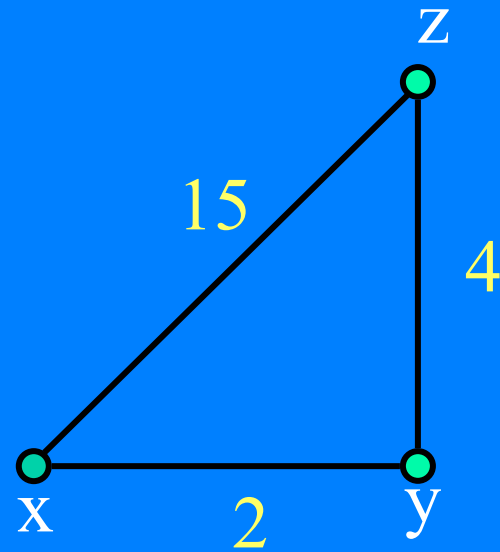


One Approach to Evacuation: Find the Shortest Route from Home to Evacuation Center



Optimization Problems

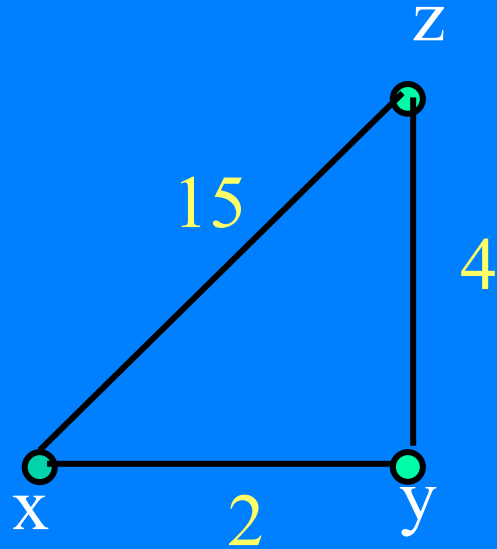
Shortest Path Problem



Numbers = some sort of weights or lengths

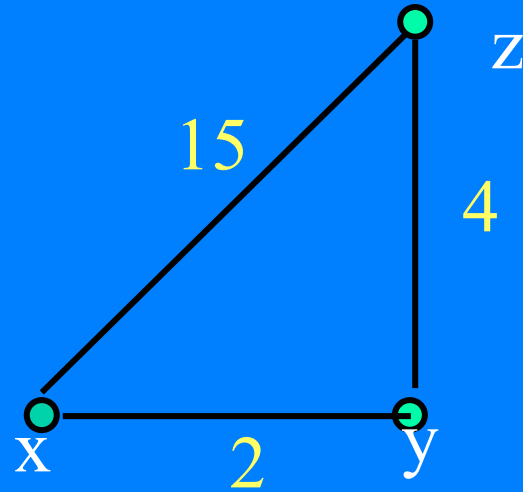
- *Problem: Find the shortest path from x to z in the network.*
- Widely applied problem.
 - ✓ US Dept. of Transportation alone uses it billions of times a year.

Shortest Path Problem



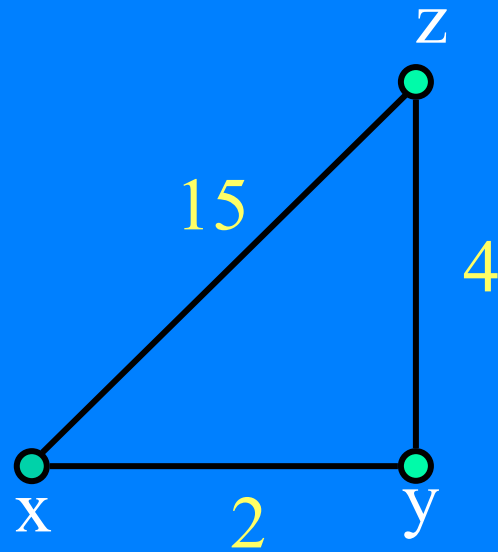
- The shortest path from x to z is the path x to y to z.
- Is this conclusion meaningful?
- It is if the numbers define a ratio scale.
- The numbers define a ratio scale if they are distances.

Shortest Path Problem



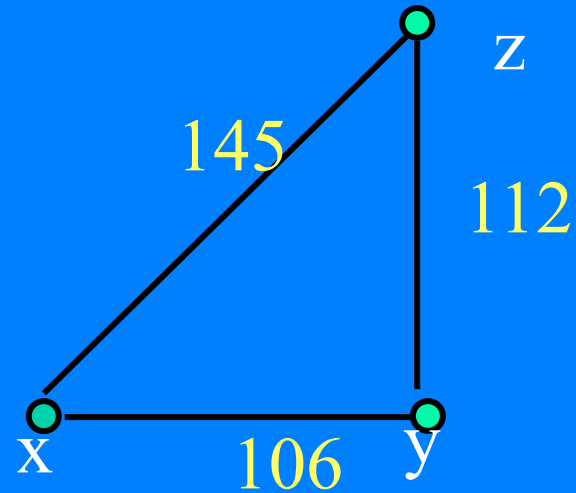
- However, what if the numbers define an interval scale?

Shortest Path Problem



- Consider the admissible transformation $\mathbb{W}(x) = 3x + 100$.

Shortest Path Problem

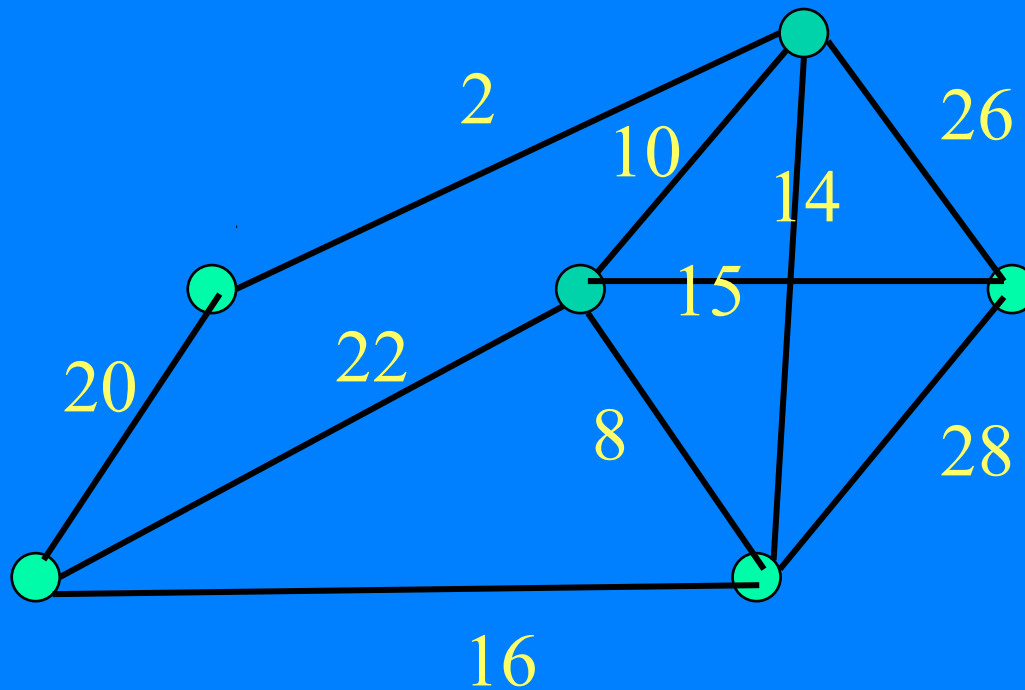


- Consider the admissible transformation $\mathbb{W}(x) = 3x + 100$.
- Now we get the above numbers on the edges.
- Now the shortest path is to go directly from x to z.
- The original conclusion was meaningless.

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.*
- How many people realize that?
- Note that linear programming is widely used in practical applications, e.g., to solve problems like:
 - ✓ Optimizing inventories of pollution control equipment
 - ✓ Assigning workers to pollution control jobs
 - ✓ Optimizing the size of a pollution control facility
 - ✓ Determining the amount to invest in alternative pollution control measures

Related Example: Minimum Spanning Tree Problem



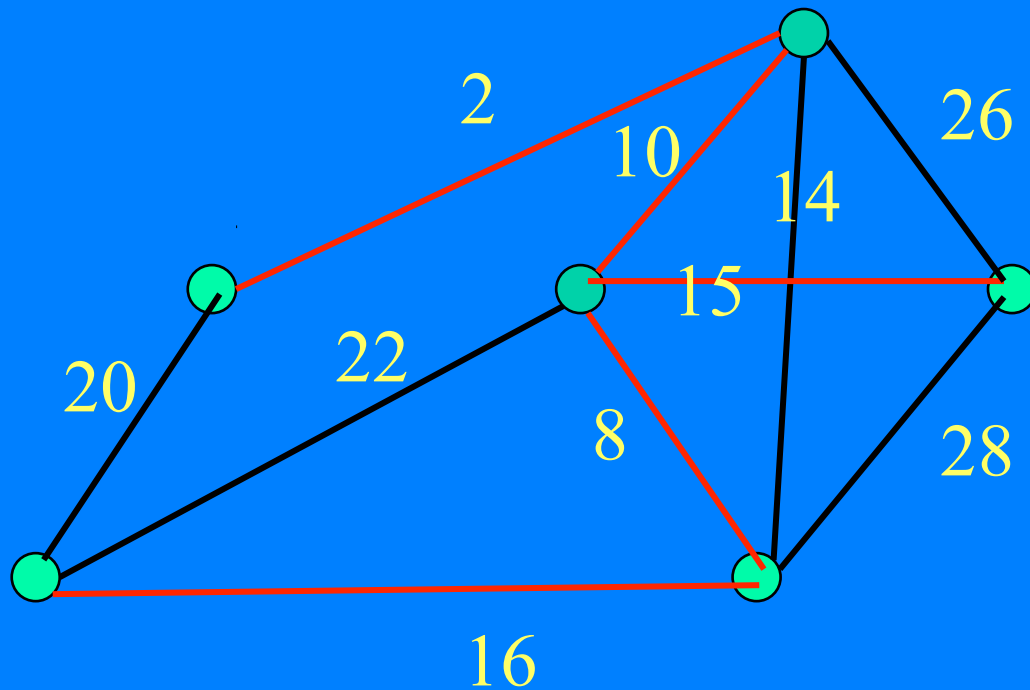
- A spanning tree is a tree using the edges of the graph and containing all of the vertices.
- It is minimum if the sum of the numbers on the edges used is as small as possible.

Related Example: Minimum Spanning Tree Problem

- Minimum spanning trees arise in many applications.
- One example: Given a road network, find usable roads that allow you to go from any vertex to any other vertex, minimizing the lengths of the roads used.
- This problem also arises in extreme events due to global warming: Find a usable road network for emergency vehicles in case extreme events leave flooded roads.

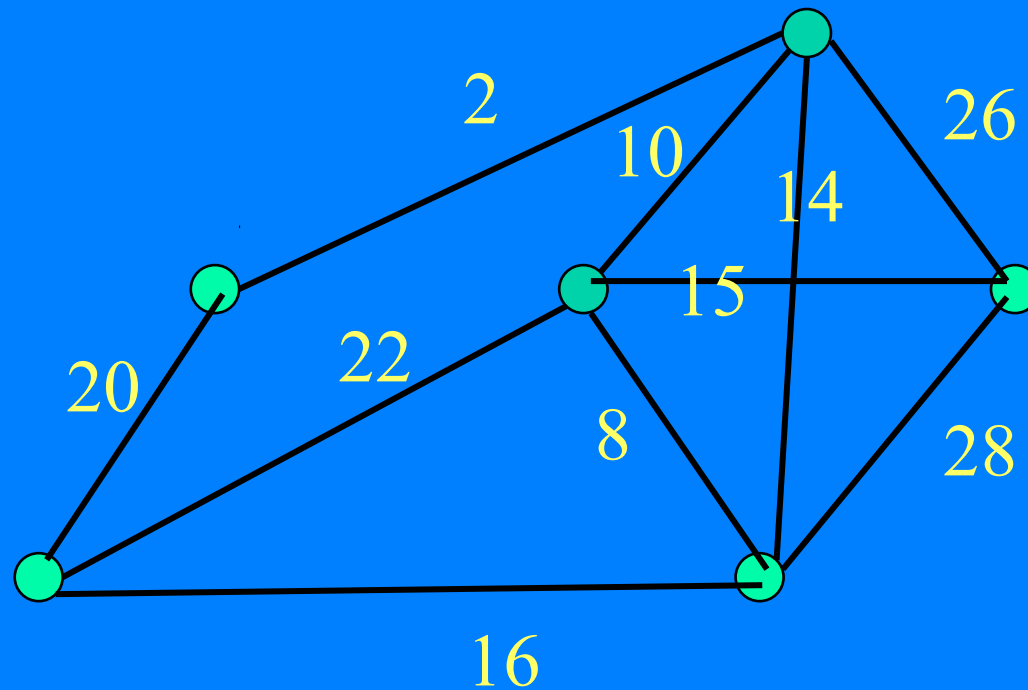


Related Example: Minimum Spanning Tree Problem



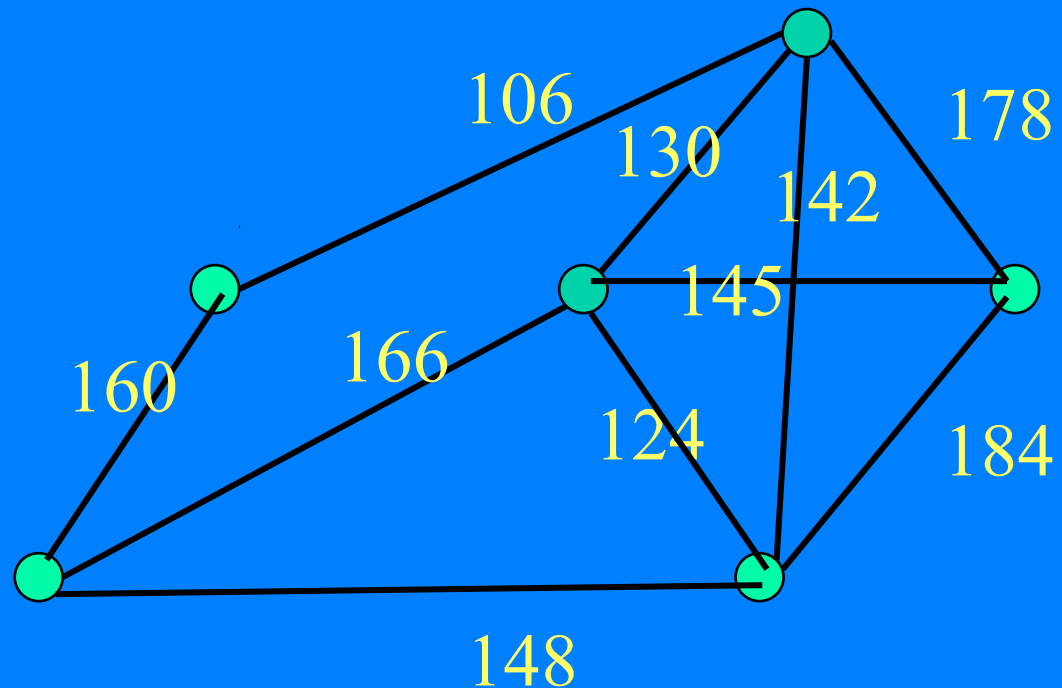
- Red edges define a minimum spanning tree.
- Is it meaningful to conclude that this is a minimum spanning tree?

Related Example: Minimum Spanning Tree Problem



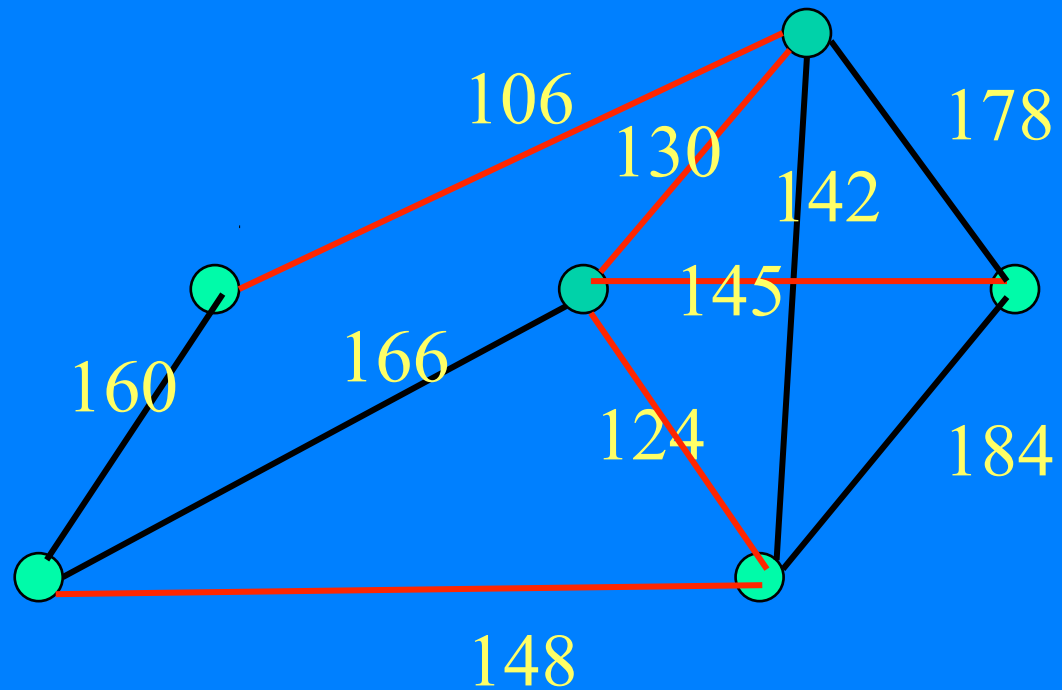
- Consider the admissible transformation $\mathbb{W}(x) = 3x + 100$.

Related Example: Minimum Spanning Tree Problem



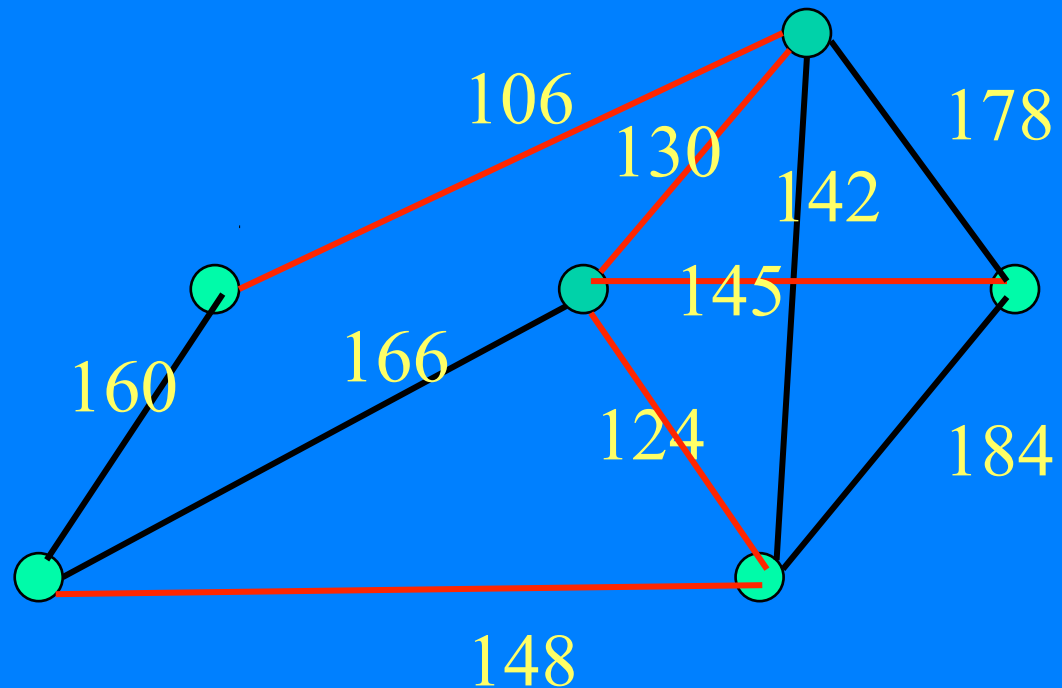
- Consider the admissible transformation $\mathbb{W}(x) = 3x + 100$.
- We now get the above numbers on edges.

Related Example: Minimum Spanning Tree Problem



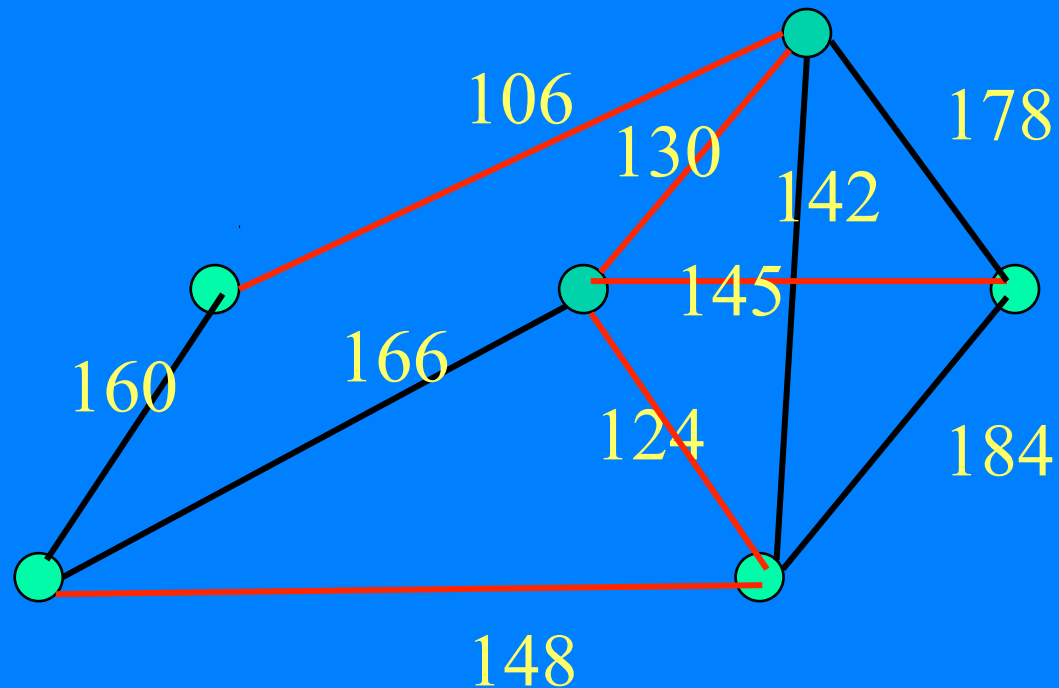
- The minimum spanning tree is the same.

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.

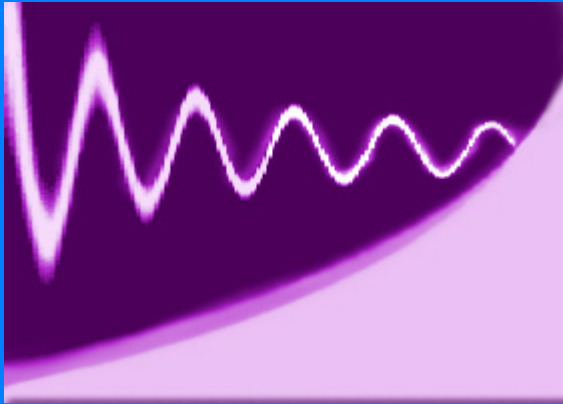
Related Example: Minimum Spanning Tree Problem

- 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 the conclusion that a particular solution is optimal may be an accident of the way things are measured.*

Outline

1. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
- 8. Measurement of Noise: Introduction to Psychophysical Scaling**
9. How to Average Scores

Measurement of Noise



A sound has physical characteristics:

- Intensity (energy transported)
- Frequency (in cycles per second)
- Duration

A sound also has psychological characteristics:

- How loud does it seem?
- What emotional meaning does it portray?
- What images does it suggest?

Measurement of Noise

- *Since middle of 19th century, scientists have tried to study the relationships between physical characteristics of stimuli like sounds and their psychological characteristics.*
- *Psychophysics* is the discipline that studies **psychological sensations** such as loudness, brightness, apparent length, apparent duration, and their relations to **physical stimuli**.
- Not all psychological characteristics have clear relationships to physical ones. E.g., emotional meaning.
- However, some seem to.
- We will concentrate on loudness.

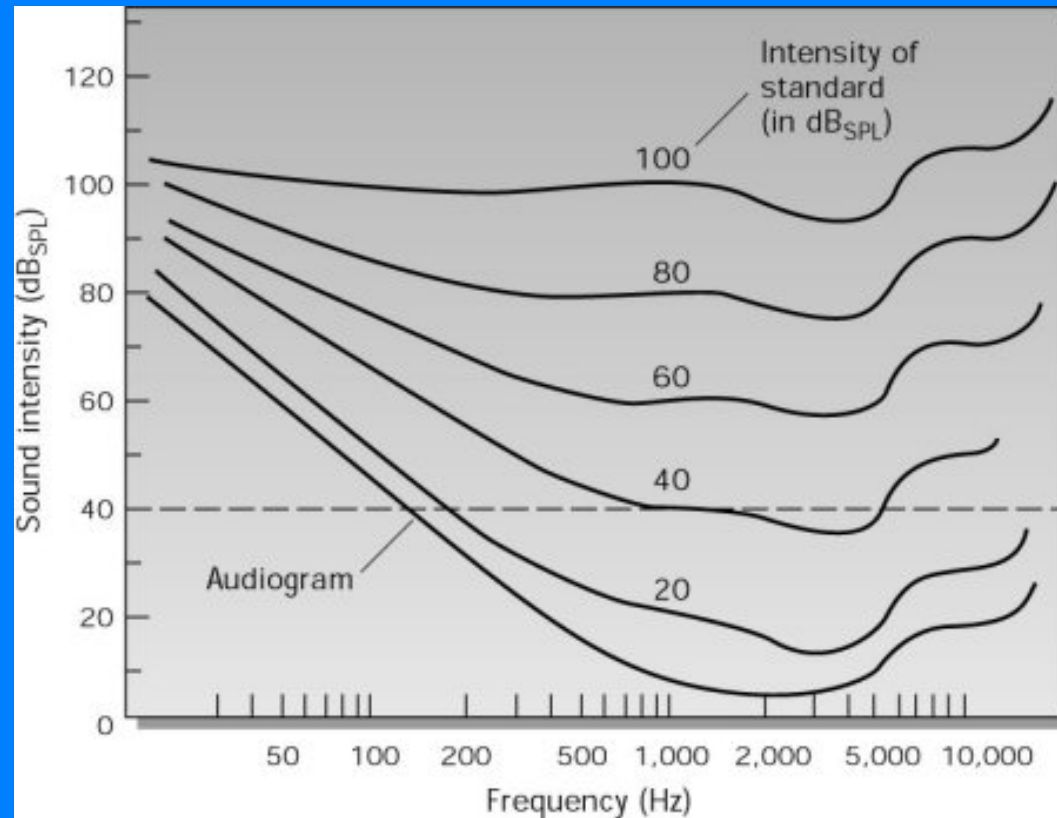
Measurement of Noise

- Loudness of a sound is different from its disturbing effect.
- This disturbing effect is called *noise*.
- We will concentrate on loudness and equate noise with loudness.
- Noise has more than just disturbing effects.
- It has physiological effects too:
 - Affects hearing
 - Affects cardiovascular system
 - May be related to stomach problems
 - May even be related to infertility



Measurement of Noise

- Subjective judgments of loudness depend on physical intensity and frequency.
- Equal loudness contour:



- Duration of a sound may also enter into loudness.

Measurement of Noise

- To simplify matters, one tries to eliminate all physical factors but one.
- Deal with *pure tones*, sounds of constant intensity at a fixed frequency.
- Present them for fixed duration of time.
- Let $I(a)$ denote the *intensity* of pure tone a .
- $I(a)$ is proportional to the root-mean-square pressure $p(a)$.
- The common unit of measurement of intensity is the *decibel (dB)*.
- This is $10 \log_{10}(I/I_0)$, where I_0 is a reference sound.
- A sound of 1 dB is essentially the lowest audible sound.

Measurement of Noise

Some Sample Decibel Levels:

Uncomfortably Loud:

Oxygen torch (121 dB)

Snowmobile (113 dB)

Riveting machine (110 dB)

Rock band (108-114 dB)

Jet takeoff at 1000 ft. (110 dB)

Jet flyover at 1000 ft. (103 dB)



Measurement of Noise

Some Sample Decibel Levels:

Very Loud:

Electric furnace (100 dB)

Power mower (96 dB)

Rock drill at 50 ft. (95 dB)

Motorcycle at 50 ft. (90 dB)

Snowmobile at 50 ft. (90 dB)

Food blender (88 dB)



Measurement of Noise

Some Sample Decibel Levels:

Moderately Loud:

Power mower at 50 ft. (85 dB)

Diesel truck at 50 ft. (85 dB)

Diesel train at 50 ft. (85dB)

Garbage disposal (80 dB)

Washing machine (78 dB)

Dishwasher (75 dB)

Passenger car at 50 ft. (75 dB)

Air conditioning unit at 50 ft. (60 dB)



Measurement of Noise

- Loudness of a sound $a = L(a)$.
- Unit of measurement of loudness = the *sones*
(1 sone = loudness of 1000 cps pure tone at 40 dB)
- Loudness is a psychological scale.
- *What is the relation between $L(a)$ and the physical intensity of a , $I(a)$?*
- This relation usually called the *psychophysical law*.

$$L(a) = \Psi(I(a))$$

- Ψ is called the *psychophysical function*.
- A basic goal of psychophysics is to find the general form of the psychophysical function that applies in many cases.

Measurement of Noise

Fechner's Law

- First attempt to specify Ψ for large class of psychological variables: Gustav Fechner (1860).
- Fechner argued that:

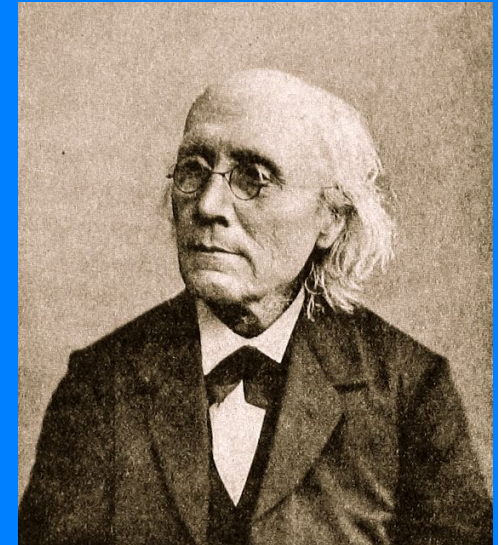
$$\Psi(x) = c \log x + k$$

c, k constant.

- This is called ***Fechner's Law***
- For loudness, this would say:

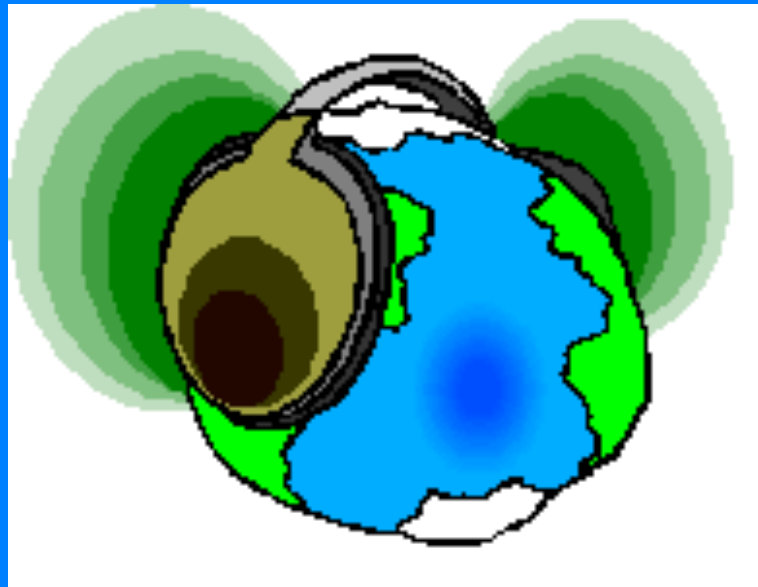
$$L(a) = c \log I(a) + k$$

- Decibel scale is a special case of Fechner's Law:
Base of log is 10, $c = 10$, $k = -10 \log_{10} I_0$.



Measurement of Noise

- *If $L(a) = dB(a)$, then a doubling of the dB level of a sound should lead to a doubling of the perceived loudness.*
- So, 100 dB should sound twice as loud as 50 dB.
- Is this the case?



Measurement of Noise

Fletcher and Munson (1933)

- Assumption: loudness proportional to number of auditory nerve impulses reaching the brain
- *Thus: sound delivered to 2 ears should appear twice as loud as same sound delivered to 1 ear.*
- F & M found that to sound equally loud, a pure tone delivered to 1 ear had to be about 10 dB higher than if it were delivered to 2 ears.
- *They concluded: subjective loudness doubles for each 10 dB increase in pressure.*
- Thus, an increase from 50 dB to 60 dB doubles loudness, not an increase from 50 dB to 100 dB.

Measurement of Noise

Fletcher and Munson (1933)

- This shows that dB does not measure loudness.
- Much data supports the F & M results.
- It also implies that Fechner's Law

$$L(a) = c \log I(a) + k$$

can't hold.

Measurement of Noise

The Power Law

- S.S. Stevens (many papers): *The fundamental psychophysical law for many psychological and physical variables is a power law:*

$$\Psi(x) = cx^k$$

- Many experiments have estimated that for loudness and intensity, the exponent k is approximately 0.3.
- Is this consistent with the Fletcher-Munson observation?

Measurement of Noise

The Power Law

- Note that $0.3 \frac{\text{W}}{\text{m}^2} \frac{\text{W}}{\text{m}^2} \log_{10} 2$

- Power law:

$$L(x) = cI(x)^{\log_{10} 2}$$

$$\text{dB}(x) = 10 \log_{10} [I(x)/I_0]$$

$$I(x) = I_0 10^{(1/10)\text{dB}(x)}$$

$$(**) \quad L(x) = cI_0^{\log_{10} 2} [10^{(1/10)\text{dB}(x)}]^{\log_{10} 2}$$

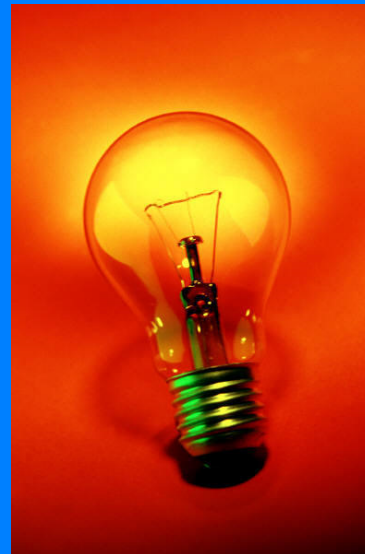
If $\text{dB}(b) = \text{dB}(a) + 10$, then letting $x = b$ in lhs of (**)
and taking $\text{dB}(x) = \text{dB}(a) + 10$ in rhs, we find that

$$L(b) = 2L(a)$$

Measurement of Noise

The Power Law

- Data seems suggests that the power law holds for more than 2 dozen variables (at least approximately and in limited intervals), including:
 - Brightness (in brils)
 - Smell
 - Taste (in gusts)
 - Judged temperature
 - Judged duration
 - Pressure on palm
 - Judged heaviness
 - Force of handgrip
 - Vocal effort



Measurement of Noise

The Power Law

- Power law fairly widely accepted for certain psychological/physical variables.
- It fails for things like pitch as function of frequency, apparent inclination, etc.
- We will discuss a theory that allows us to derive the power law.

Outline

1. Introduction to Measurement Theory
2. Theory of Uniqueness of Scales of Measurement/Scale Types
3. Meaningful Statements
4. Averaging Judgments of Loudness
5. Measurement of Air Pollution: A Combined Pollution Index
6. Evaluation of Water Testing Equipment: “Merging Normalized Scores”
7. Optimization Problems in Pollution Measurement
8. Measurement of Noise: Introduction to Psychophysical Scaling
- 9. How to Average Scores**

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?*
- Methods we will describe will help and also help with the possible psychophysical laws.

How Should We Average Scores?

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

• Let a_1, a_2, \dots, a_n be “scores” or ratings, e.g., scores on benchmarks for water or air pollution equipment, loudness ratings, etc.

• Let $u = F(a_1, a_2, \dots, a_n)$

• F is an unknown averaging function – sometimes called a *merging function*, and u is the average or merged score.

How Should We Average Scores?

- Approaches to finding acceptable merging functions F :
 - (1) axiomatic
 - (2) scale types
 - (3) meaningfulness

How Should We Average Scores?

An Axiomatic Approach

Theorem (Fleming and Wallace). Suppose $F: (\mathbb{W}^+)^n \rightarrow \mathbb{W}^+$ has the following properties:

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

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

(3). *Multiplicativity*:

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

Then F is the geometric mean. And conversely.

How Should We Average Scores?

A Functional Equations Approach Using Scale Type or Meaningfulness Assumptions

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

We will use an idea due to R. Duncan Luce that he called the “*Principle of Theory Construction*”

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



How Should We Average Scores?

A Functional Equations Approach Using Scale Type or Meaningfulness Assumptions

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

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 .

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*:

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

How Should we Average Scores?

• This leads to the functional equation:

$$(\&) \quad F\left(\frac{w}{W}a\right) = A\left(\frac{w}{W}\right)F(a), \quad \frac{w}{W} > 0, A\left(\frac{w}{W}\right) > 0.$$

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.

What are the Possible Psychophysical Laws?

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, in psychophysical scaling, if both the physical and psychological variables are ratio scales, the only possible way to relate them is by a power law.

$$\boxed{W}(x) = cx^k$$

What are the Possible Psychophysical Laws?

Thus, in psychophysical scaling, if both the physical and psychological variables are ratio scales, the only possible way to relate them is by a power law.

$$\Psi(x) = cx^k$$

- The functional equations approach can be viewed as a derivation of the power law in psychophysics. In particular, it holds if loudness defines a ratio scale.
- So how do you know that loudness defines a ratio scale?
- One of Stevens' arguments: Because subjects can do "magnitude estimation" and are comfortable with ratios of sounds (a sounds twice as loud as b).

The Possible Scientific Laws

- This result is also 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 Cont'd

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

$$F: (\mathbb{W}^+)^n \rightarrow \mathbb{W}^+$$

$$F(a_1, a_2, \dots, a_n) = u \cdot F(\mathbb{W}_1 a_1, \mathbb{W}_2 a_2, \dots, \mathbb{W}_n a_n) = \mathbb{W} u,$$

$\mathbb{W}_1 > 0, \mathbb{W}_2 > 0, \dots, \mathbb{W}_n > 0, \mathbb{W} > 0$, \mathbb{W} depends on a_1, a_2, \dots, a_n .

• Thus we get the functional equation:

$$(*) \quad F(\mathbb{W}_1 a_1, \mathbb{W}_2 a_2, \dots, \mathbb{W}_n a_n) = A(\mathbb{W}_1, \mathbb{W}_2, \dots, \mathbb{W}_n) F(a_1, a_2, \dots, a_n),$$

How Should We Average Scores?

A Functional Equations Approach

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

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

Theorem (Luce 1964): If $F: (\mathbb{R}^+)^n \rightarrow \mathbb{R}^+$ is continuous and

satisfies (*), then there are $\alpha_i > 0, c_1, c_2, \dots, c_n$ so that

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

How Should We Average Scores?

Theorem (Aczél and Roberts 1989): If in addition F satisfies reflexivity and symmetry, then $\left[\frac{W}{\mathbb{W}} \right] = 1$ and $c_1 = c_2 = \dots = c_n = 1/n$, so F is the geometric mean.

Janos Aczél
“Mr. Functional Equations”



How Should We Average Scores?

Sometimes You Get the Arithmetic Mean

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

Functional Equation:

$$(\text{****}) \quad F\left(\frac{W}{W}a_1 + \frac{W_1}{W}1, \frac{W}{W}a_2 + \frac{W_2}{W}2, \dots, \frac{W}{W}a_n + \frac{W_n}{W}n\right) = \\ A\left(\frac{W}{W}, \frac{W_1}{W}, \frac{W_2}{W}, \dots, \frac{W_n}{W}\right)F(a_1, a_2, \dots, a_n) + B\left(\frac{W}{W}, \frac{W_1}{W}, \frac{W_2}{W}, \dots, \frac{W_n}{W}\right)$$

$$A\left(\frac{W}{W}, \frac{W_1}{W}, \frac{W_2}{W}, \dots, \frac{W_n}{W}\right) > 0$$

How Should We Average Scores?

Functional Equation:

$$\begin{aligned}
 (***) \quad & F\left(\frac{a_1}{W}, \frac{a_1}{W_1}, \frac{a_2}{W}, \frac{a_2}{W_2}, \dots, \frac{a_n}{W}, \frac{a_n}{W_n}\right) = \\
 & A\left(\frac{a_1}{W}, \frac{a_1}{W_1}, \frac{a_2}{W}, \frac{a_2}{W_2}, \dots, \frac{a_n}{W}, \frac{a_n}{W_n}\right) F(a_1, a_2, \dots, a_n) + B\left(\frac{a_1}{W}, \frac{a_1}{W_1}, \frac{a_2}{W}, \dots, \frac{a_n}{W_n}\right)
 \end{aligned}$$

$$A\left(\frac{a_1}{W}, \frac{a_1}{W_1}, \frac{a_2}{W}, \frac{a_2}{W_2}, \dots, \frac{a_n}{W}, \frac{a_n}{W_n}\right) > 0$$

Solutions to (***) (Even Without Continuity Assumed)
 (Aczél, Roberts, and Rosenbaum) \times^n

$$F\left(\frac{a_1}{W}; \frac{a_2}{W_2}; \dots; \frac{a_n}{W_n}\right) = \sum_{i=1}^n \frac{a_i}{W_i} + b$$

$\frac{a_1}{W_1}, \frac{a_2}{W_2}, \dots, \frac{a_n}{W_n}, b$ arbitrary constants

How Should We Average Scores?

Theorem (Aczél and Roberts):

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

$$\prod_{i=1}^n s_i = 1, b = 0:$$

(2). If in addition F satisfies reflexivity and symmetry, then $\left[\frac{W}{n} \right]_i = 1/n$ 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, 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 xxxxxxxx by the assumption that a certain statement involving u is meaningful.

How Should We Average Scores?

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

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

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

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

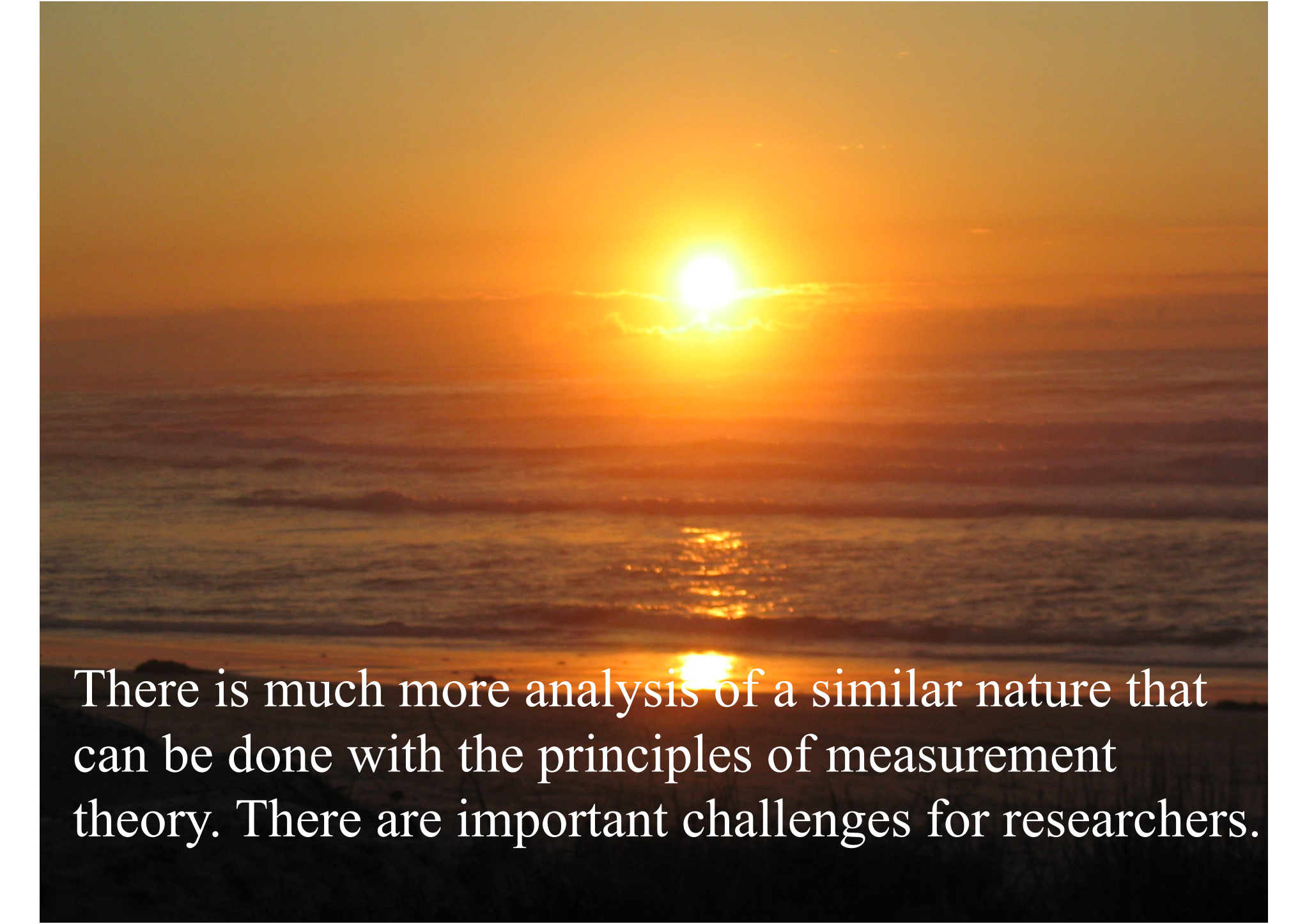
$$F(a_1; a_2; \dots; a_n) = \sqrt[n]{a_1^{c_1} a_2^{c_2} \dots a_n^{c_n}}$$

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

How Should We Average Scores?

- Averaging of measurements or judgments or estimates is commonly carried out in a variety of applied areas.
- It is certainly relevant not only to air, water, and noise pollution, but to decision making about other kinds of pollution, such as visual pollution, thermal pollution, land pollution, radioactive pollution, etc.
- Thus is it important in many applications to know what averaging procedures lead to meaningful conclusions.





There is much more analysis of a similar nature that can be done with the principles of measurement theory. There are important challenges for researchers.