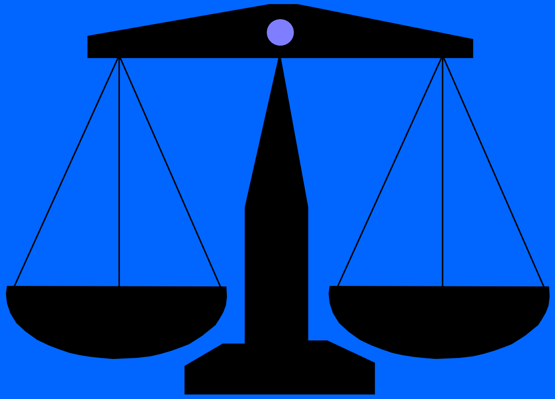
# Meaningless Statements in Epidemiology

## Fred Roberts, DIMACS and CCICADA, Rutgers University



# My Message

- The modern theory of measurement was developed to deal with measurement in the social and behavioral sciences where scales are not as readily defined as in the physical sciences.
  - Utility, noise, intelligence, …
- Traditional concepts of measurement theory are not well known in the public health arena.
- Problems of epidemiology and public health are providing new challenges for measurement theory.

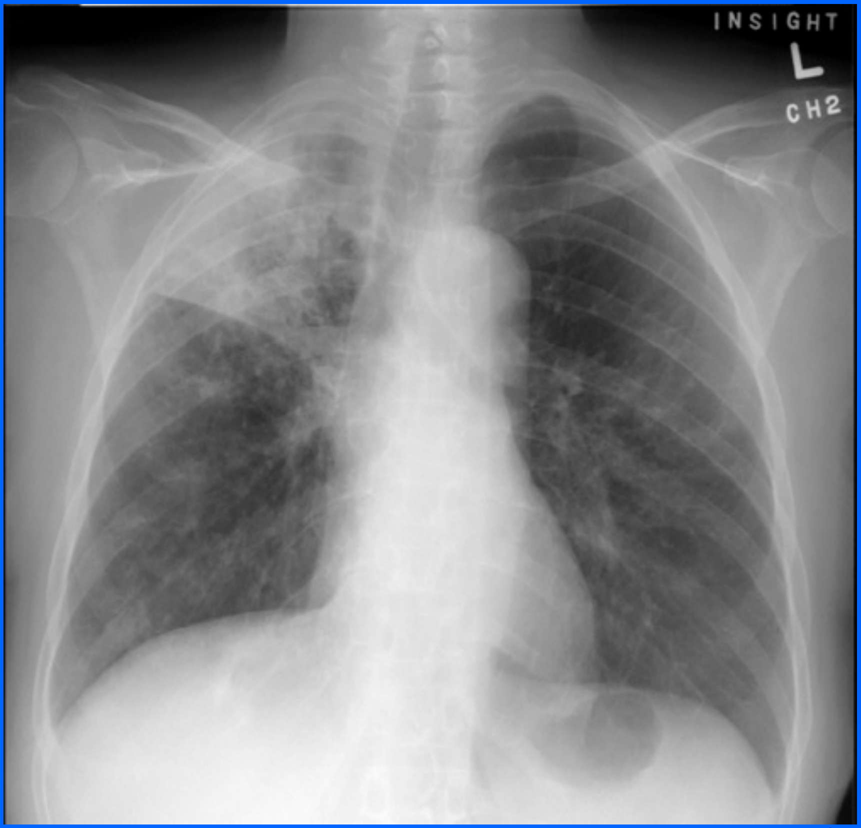
measles

# Some Questions We Will Ask

- Is it meaningful to say that the malaria parasite load has doubled?

# Some Questions We Will Ask

- Is the average cough score for one set of TB patients higher than that for another?



# Some Questions We Will Ask

- For controlling the spread of HIV, which of abstinence education, universal screening, and condom distribution are more effective?

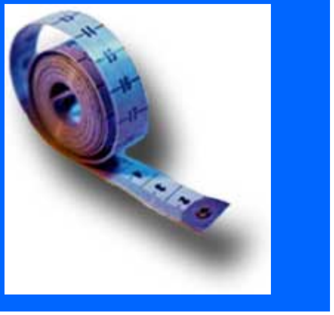
# MEASUREMENT

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

- We will discuss applications of the theory of measurement to measurement in epidemiology and public health.

# MEASUREMENT

- ***Measurement*** has something to do with numbers.
- 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. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. 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 scale of measurement.

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

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

# The Theory of Uniqueness

## Admissible Transformations $\varphi$

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

Kilograms → Pounds: $\varphi(x) = 2.2x$

# 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

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

# Some Common Scale Types

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

# Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. **Meaningful Statements**
3. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: “Merging Normalized Scores”
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. 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 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

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

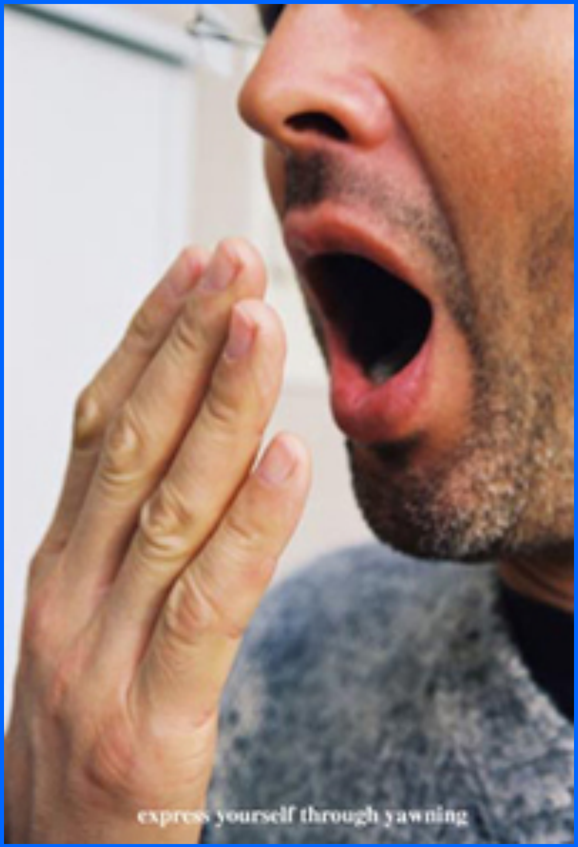
- Is this meaningful?

# Meaningful Statements

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

- Is this meaningful?

I hope not!

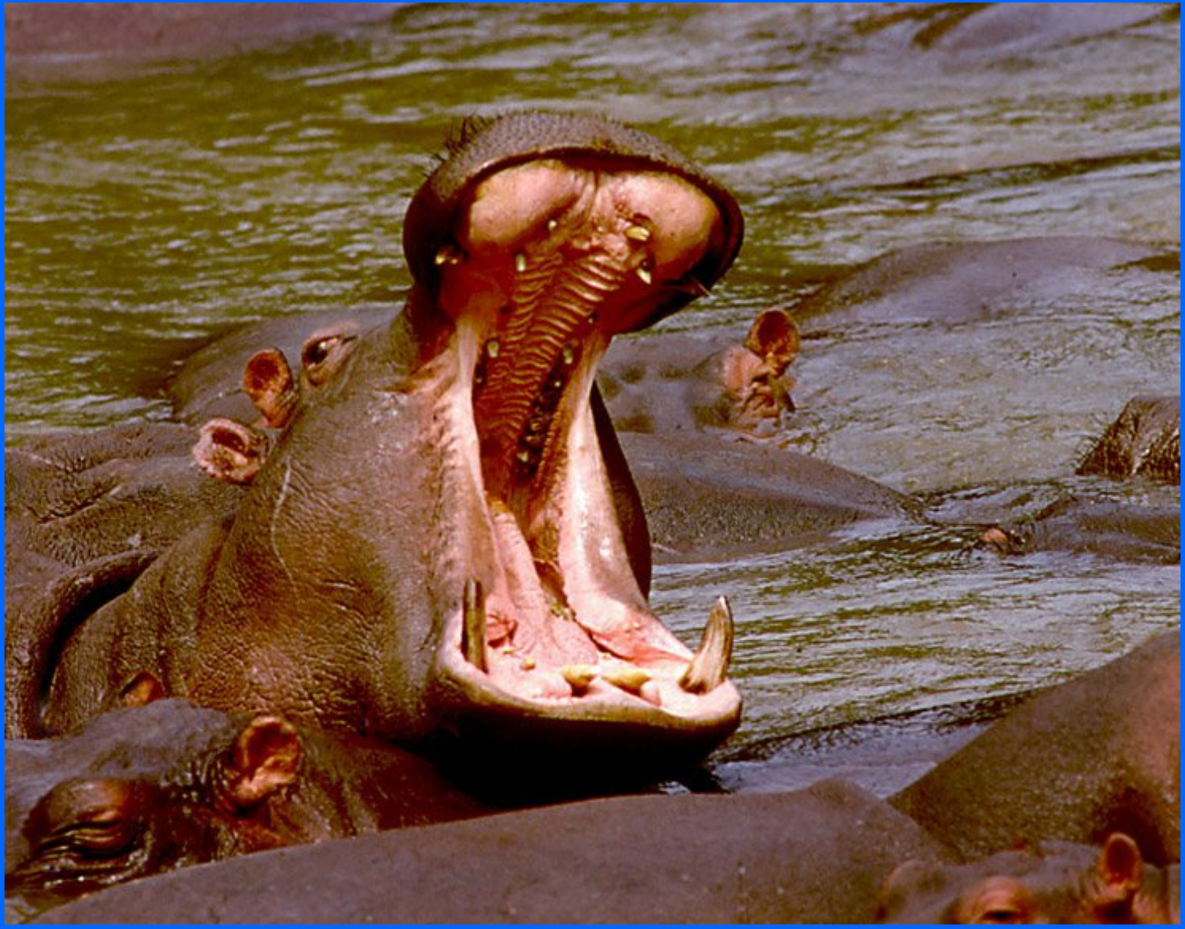


# Meaningful Statements

**“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) $$f(a) = 3f(b).$$

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

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

- But (2) becomes

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

- (1) ↔ (3) since $\alpha > 0$.

# Meaningful Statements

**“The patient’s temperature at 9AM today is 2 per cent higher than it was at 9 AM yesterday.”**

- Is this meaningful?

# Meaningful Statements

**"The patient's temperature at 9AM today is 2 per cent higher than it was at 9 AM yesterday."**

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

•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:

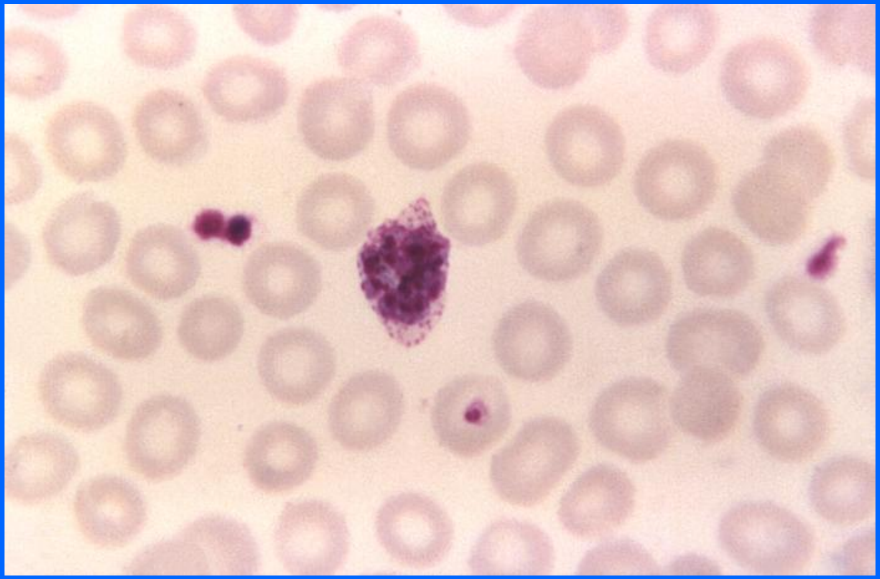
$$f(a) > f(b)$$

# Meaningful Statements

**Malaria parasite density is still mainly obtained by reading slides under microscopes.**

**"The parasite density in this slide is double the parasite density in that slide."**

- Is this meaningful?



# Meaningful Statements

**"The parasite density in this slide is double the parasite density in that slide."**

- Density is measured in number per microliter. So, if one slide has 100,000 per $\mu$L and another 50,000 per $\mu$ L, is it meaningful to conclude that the first slide has twice the density of the second?
- Meaningful. Volume involves ratio scales. And counts are absolute scales.

However: This disregards errors in measurement. A statement can be meaningful in the measurement theory sense but meaningless in a practical sense.

# 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. **Averaging Judgments of Cough Severity**
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. How to Average Scores

# Average Cough Severity

- Study two groups of patients with TB.
- *f(a)* is the cough severity of *a* as judged on one of the subjective cough severity scales (e.g., rate severity as 1 to 5).
- **Data suggests that the average cough severity for patients in the first group is higher than the average cough severity of patients in the second group.**

$a_1, a_2, \ldots, a_n$ patients in first group

$b_1, b_2, \ldots, b_m$ patients 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 Cough Severity

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

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

- **If cough severity defines a ratio scale:**

- Then, $\varphi(x) = \alpha x$, $\alpha > 0$, 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 TB patients.

# Average Cough Severity

- Note: **(1) is still meaningful if *f* is an interval scale.**

- For example, we could be comparing temperatures *f(a)*.

- Here, $\varphi(x) = \alpha x + \beta,\ \alpha > 0$. 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 Cough Severity

•To show that comparison of arithmetic means can be meaningless for ordinal scales, note that we are asking experts for a subjective judgment of cough severity.

•It seems that *f(a)* is measured on an ordinal scale, e.g., 5-point scale: 5=extremely severe, 4=very severe, 3=severe, 2=slightly severe, 1=no cough.

**•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 cough severity of group 2 patients is higher.

# Average Cough Severity

- Suppose *f(a)* is measured on an ordinal scale, e.g., 5-point scale: 5=extremely severe, 4=very severe, 3=severe, 2=slightly severe, 1=no cough.
- 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 severity of group 1 patients is higher.

# Average Cough Severity

•**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 Fatigue

- Fatigue is an important variable in measuring the progress of patients with serious diseases.
- One scale widely used in measuring fatigue is the Piper Fatigue Scale.
- It asks questions like:
  - On a scale of 1 to 10, to what degree is the fatigue you are feeling now interfering with your ability to complete your work or school activities? (1 = none, 10 = a great deal)
  - On a scale of 1 to 10, how would you describe the degree of intensity or severity of the fatigue which you are experiencing now? (1 = mild, 10 = severe)
- Similar analysis applies: Meaningless to compare means, meaningful to compare medians

# Average Cough Severity

**•Suppose each of *n* observers is asked to rate each of a collection of patients as to their relative cough severity.**

• Or we rate patients on different criteria or against different benchmarks. (Similar results with performance ratings, importance ratings, etc.)

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

# Average Cough Severity

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

- A similar question arises in fatigue ratings, ratings of brightness of rash, etc.

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

# Average Cough Severity

- 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 $\alpha_i$ so that (1) holds and (3) fails. Thus, (1) is meaningless.

# Average Cough Severity

Motivation for considering different $\alpha_i$:

$n = 2$, $f_1(a)$ = 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.

# Average Cough Severity

Motivation for considering different $\alpha_i$:

$n = 2$, $f_1(a)$ = 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 Cough Severity

•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 cough severity 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

## 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.

# 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. Averaging Judgments of Cough Severity
4. **Measurement of Air Pollution**
5. Evaluation of Alternative HIV Treatments: “Merging Normalized Scores”
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. How to Average Scores

# MEASUREMENT OF AIR POLLUTION

# MEASUREMENT OF AIR POLLUTION

- Close relationship between pollution and health
- 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 in order 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.

$$A(t,k) = \sum_{i=1}^{n} e(i,t,k)$$

# 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?

$$A(t,k) > A(t,k')$$

# 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?

$$A(t,k_r) > \sum_{k\neq k_r} A(t,k)$$

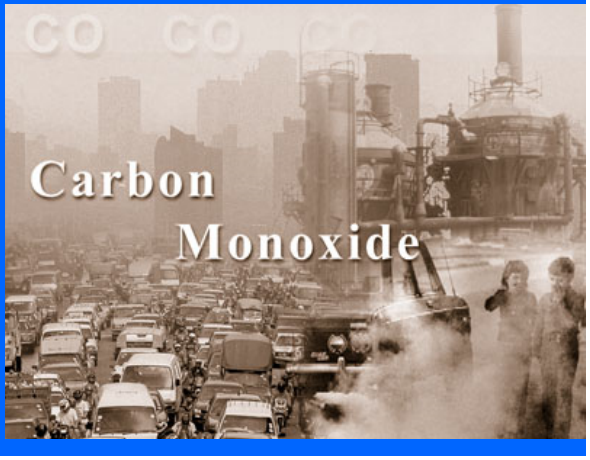
# MEASUREMENT OF AIR POLLUTION

- 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?

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



# MEASUREMENT OF AIR POLLUTION

$$A(t,k) > A(t,k')$$

$$A(t,k_r) > \sum_{k\neq k_r} A(t,k)$$

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

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. Early EPA standards based on health effects for 24 hour period allowed 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 $\tau(i)$ be tolerance factor for $i$th pollutant.

•***Severity factor***: $\tau(CO)/\tau(i)$ or $1/\tau(i)$

# 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***: $1/\tau(i) \, * \, e(i,t,k)$

and the combined index

***Pindex***: $B(t,k) = \sum_{i=1}^{n} [1/\tau(i) \, * \, e(i,t,k)]$

•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 $\tau$ 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 $1/\tau(i)$.

•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. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. **Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"**
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. How to Average Scores

# Evaluation of Alternative HIV Treatments

•How do we evaluate alternative possible treatment plans or interventions for a given disease?

• One common procedure: A number of treatments are compared on different criteria/benchmarks.

•Their scores on each criterion are normalized relative to the score of one of the treatments.

•The normalized scores of a treatment are combined by some averaging procedure and normalized scores are compared.

AIDS orphans

# Evaluation of Alternative HIV Treatments

•The normalized scores of a treatment are combined by some averaging procedure.

•If the averaging is the arithmetic mean, then the statement **“one treatment has a higher arithmetic mean normalized score than another system”** is meaningless:

•The treatment to which scores are normalized can determine which has the higher arithmetic mean.

AIDS street kids

# Evaluation of HIV Treatments

•Similar methods are used in comparing performance of alternative computer systems or other types of machinery.

•Consider a number of treatments/interventions:

- ✓Universal screening
- ✓Free condom distribution
- ✓Abstinence education
- ✓Male circumcision

•Consider a number of criteria/outcomes:

- ✓CD4 count
- ✓Days without symptoms of …
- ✓Number days hospitalized …

# Treatment Evaluation

## Evaluation of HIV Treatments

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

# Treatment Evaluation

## Normalize Relative to Treatment R

### CRITERION

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

# Treatment Evaluation

## Take Arithmetic Mean of Normalized Scores

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

# Treatment Evaluation

## Take Arithmetic Mean of Normalized Scores

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

Conclude that treatment Z is best

# Treatment Evaluation

## Now Normalize Relative to Treatment M

CRITERION

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

# Treatment Evaluation

## Take Arithmetic Mean of Normalized Scores

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

# Treatment Evaluation

## Take Arithmetic Mean of Normalized Scores

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

**Conclude that treatment R is best**

# Treatment Evaluation

- So, the conclusion that a given treatment 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) (in a computing machine application)
- Sometimes, ***geometric mean*** is helpful.
- Geometric mean is

$$\sqrt[n]{\Pi_i s(x_i)}$$

# Treatment Evaluation

## Normalize Relative to Treatment R

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

**Conclude that treatment R is best**

# Treatment Evaluation

## Now Normalize Relative to Treatment M

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

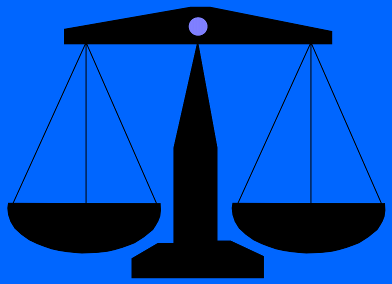
Still conclude that treatment R is best

# Treatment Evaluation

- In this situation, it is easy to show that ***the conclusion that a given treatment has highest geometric mean normalized score is a meaningful conclusion.***
- ***Even meaningful: A given treatment has geometric mean normalized score 20% higher than another treatment.***
- 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.***

# Treatment Evaluation

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

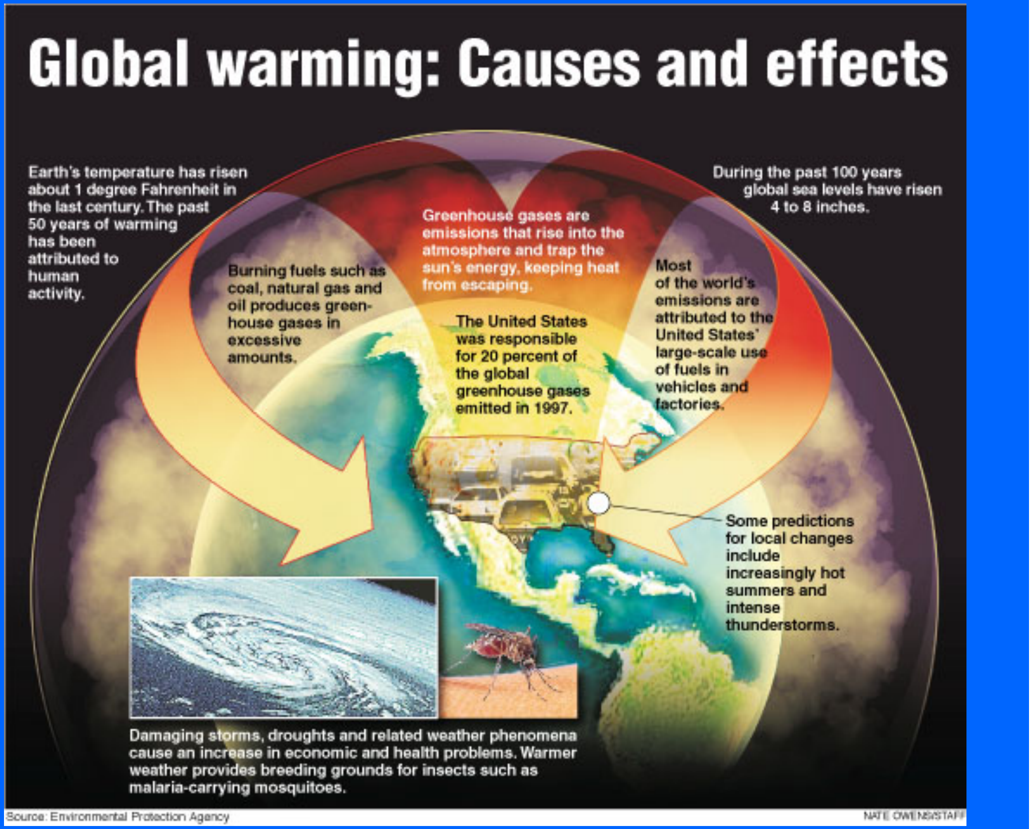
- A brief discussion follows later – if there is time.

# Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"
6. **Optimization Problems in Epidemiology**
7. Meaningfulness of Statistical Tests
8. How to Average Scores

# DIMACS Initiative on Climate and Health

- Spurred by concerns about global warming.
- Resulting impact on health
  - –Of people
  - –Of animals
  - –Of plants
  - –Of ecosystems



Global warming: Causes and effects

Earth's temperature has risen about 1 degree Fahrenheit in the last century. The past 50 years of warming has been attributed to human activity.

Burning fuels such as coal, natural gas and oil produces greenhouse gases in excessive amounts.

Greenhouse gases are emissions that rise into the atmosphere and trap the sun's energy, keeping heat from escaping.

The United States was responsible for 20 percent of the global greenhouse gases emitted in 1997.

During the past 100 years global sea levels have risen 4 to 8 inches.

Most of the world's emissions are attributed to the United States' large-scale use of fuels in vehicles and factories.

Some predictions for local changes include increasingly hot summers and intense thunderstorms.

Damaging storms, droughts and related weather phenomena cause an increase in economic and health problems. Warmer weather provides breeding grounds for insects such as malaria-carrying mosquitoes.

Source: Environmental Protection Agency

NATE OWENS/STAFF

# Extreme Events due to Global Warming

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

# DIMACS Project: Extreme Heat Events

- Extreme heat events: key test case when CDC rolled out its mathematical modeling initiative.
- Extreme heat events result in increased incidence of heat stroke, dehydration, cardiac stress, respiratory distress.
- Hyperthermia in elderly patients can lead to cardiac arrest.
- Effects not independent: Individuals under stress due to climate may be more susceptible to infectious diseases

# Extreme Heat Events: Evacuation

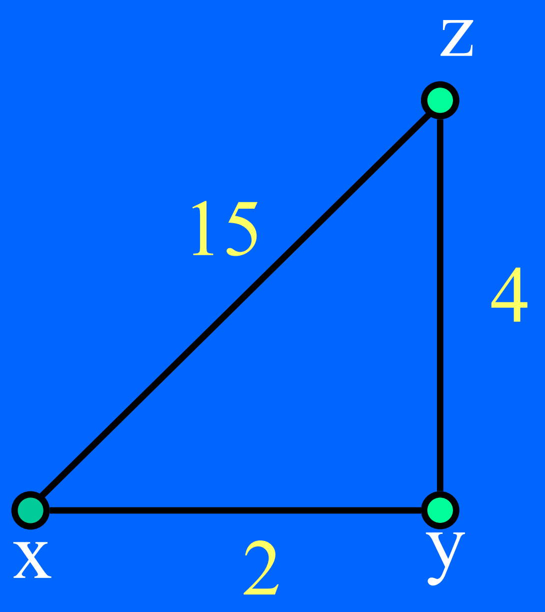
•One response to such events: evacuation of most vulnerable individuals to climate controlled environments.

•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

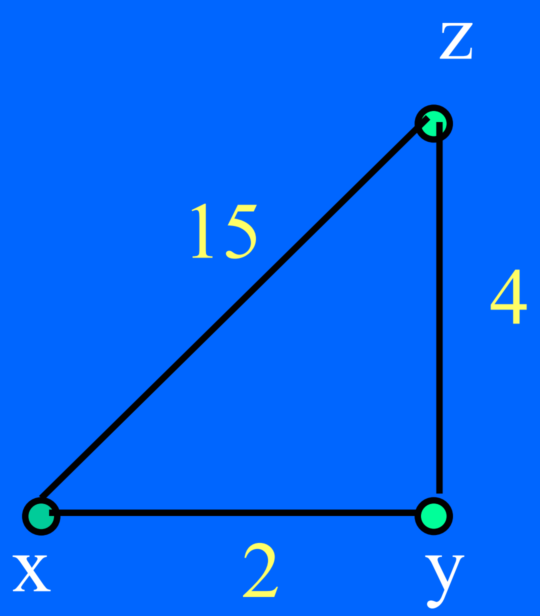
# Optimization Problems in Epidemiology: Shortest Path Problem



Numbers = some sort of weights or lengths

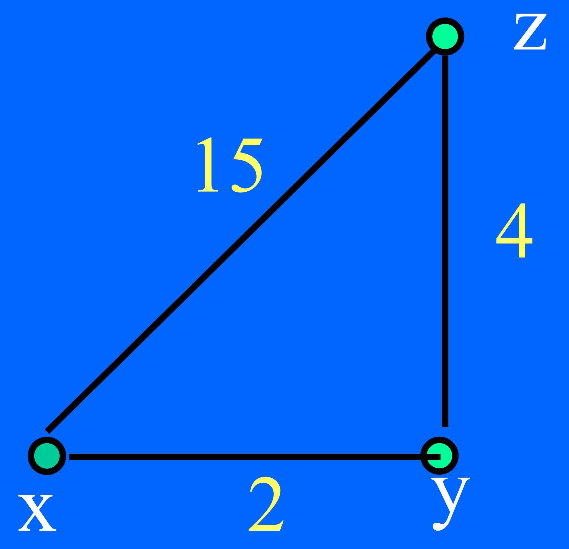
- *Problem: Find the shortest path from x to z in the network.*
- Widely applied problem.
  - US Dept. of Transportation alone uses it billions of times a year.

# Shortest Path Problem



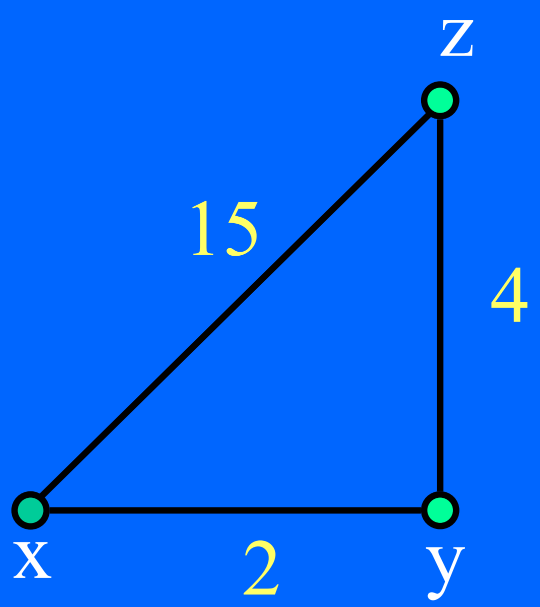
- The shortest path from x to z is the path x to y to z.
- Is this conclusion meaningful?
- It is if the numbers define a ratio scale.
- The numbers define a ratio scale if they are distances, as in the DIMACS Climate and Health project.

# Shortest Path Problem



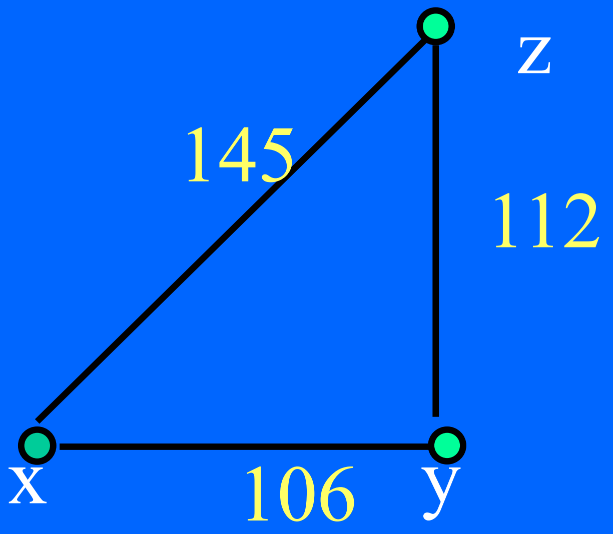
- However, what if the numbers define an interval scale?
- For example, the numbers could be costs in terms of utility (or disutility) assigned to a route, and these might only define an interval scale.

# Shortest Path Problem



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

# Shortest Path Problem

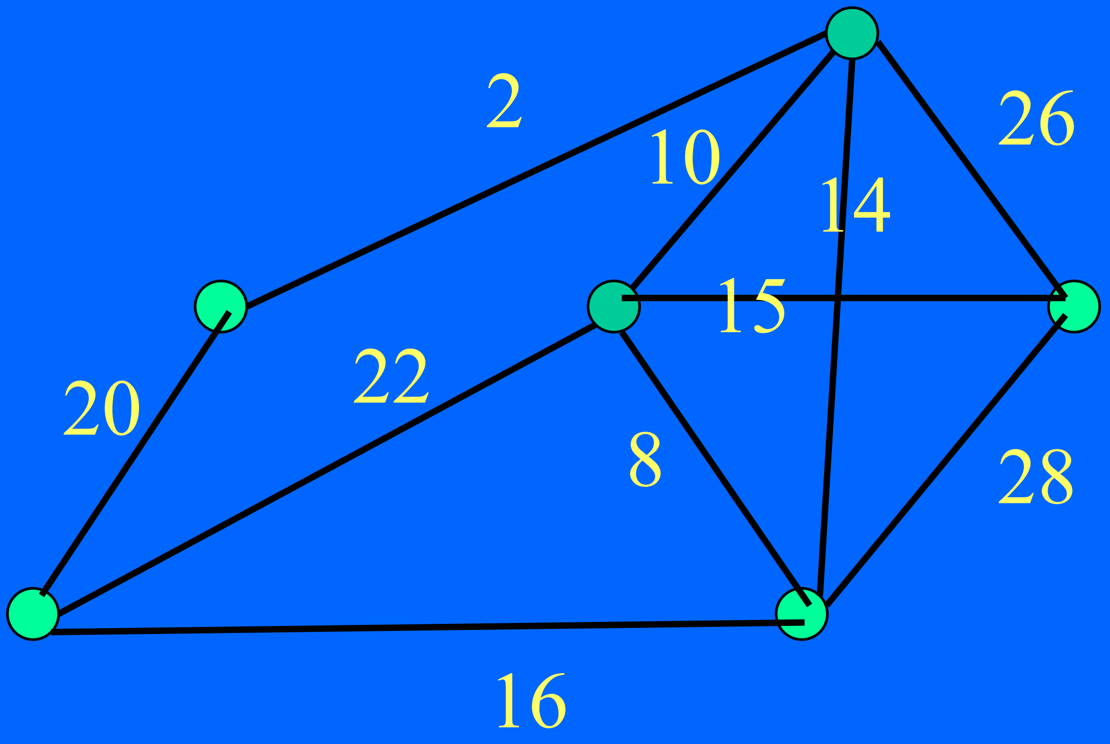


- Consider the admissible transformation $\varphi(x) = 3x + 100$.
- Now we get the above numbers on the edges.
- Now the shortest path is to go directly from x to z.
- The original conclusion was meaningless.

# 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 public health, for example to solve problems like:
  - Optimal inventories of medicines
  - Assignment of patients or doctors to clinics
  - Optimization of size of a treatment facility
  - Amount to invest in preventive treatments

# Related Example: Minimum Spanning Tree Problem

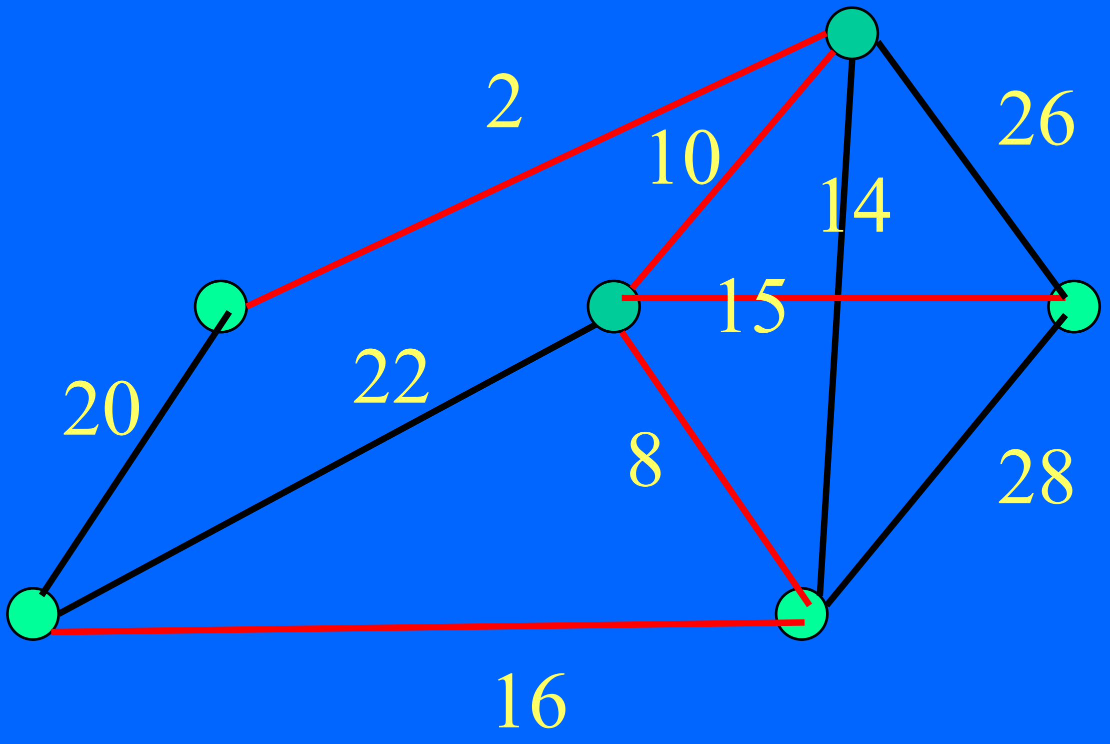


- A spanning tree is a tree using the edges of the graph and containing all of the vertices.
- It is minimum if the sum of the numbers on the edges used is as small as possible.

# Related Example: Minimum Spanning Tree Problem

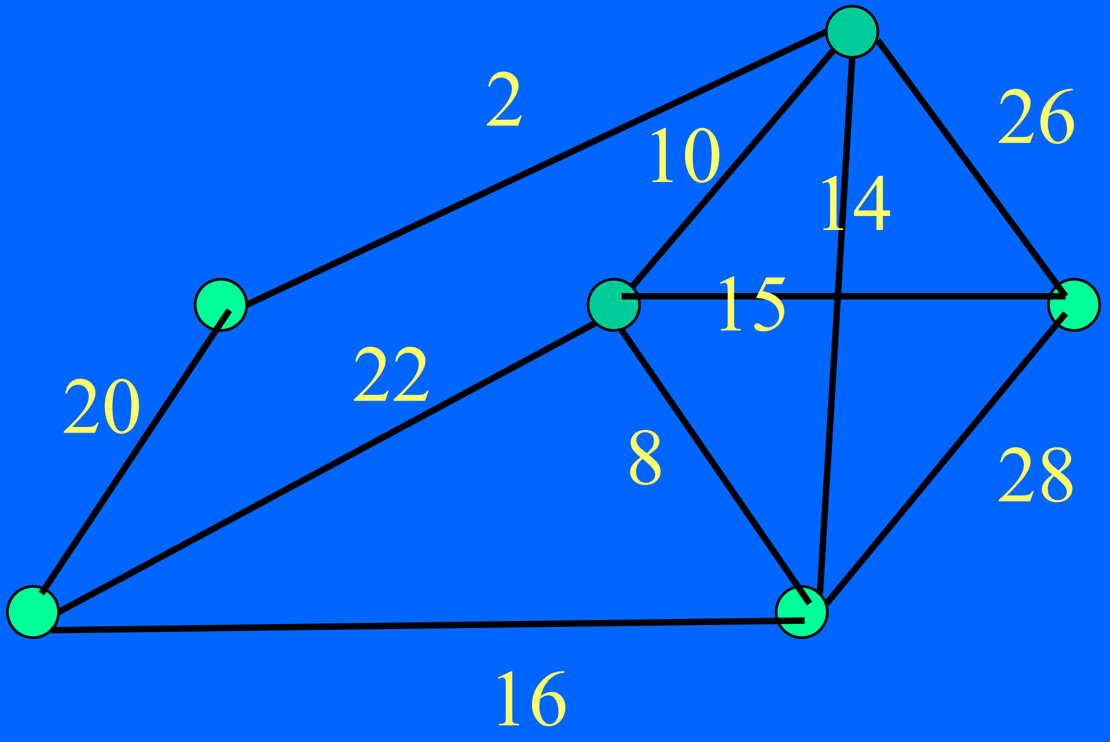
- Minimum spanning trees arise in many applications.
- One example: Given a road network, find usable roads that allow you to go from any vertex to any other vertex, minimizing the lengths of the roads used.
- This problem arises in another DIMACS Climate and Health project: Find a usable road network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

# Related Example: Minimum Spanning Tree Problem



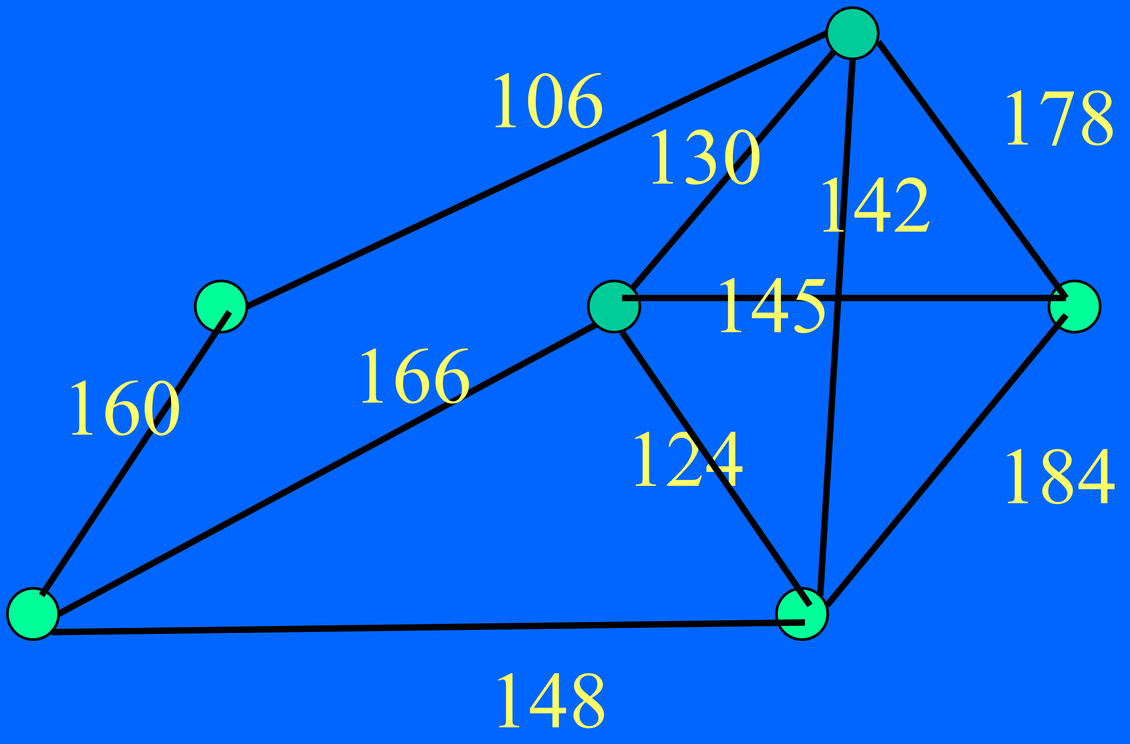
- Red edges define a minimum spanning tree.
- Is it meaningful to conclude that this is a minimum spanning tree?

# Related Example: Minimum Spanning Tree Problem



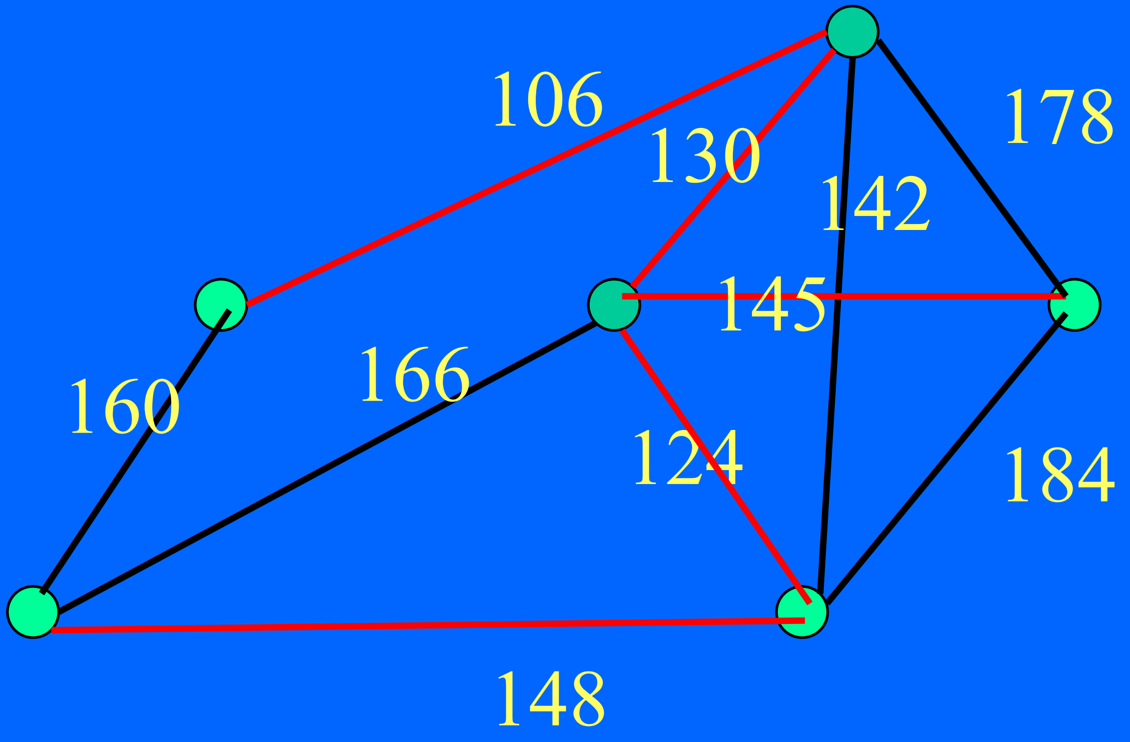
- Consider the admissible transformation $\varphi(x) = 3x + 100$.

# Related Example: Minimum Spanning Tree Problem



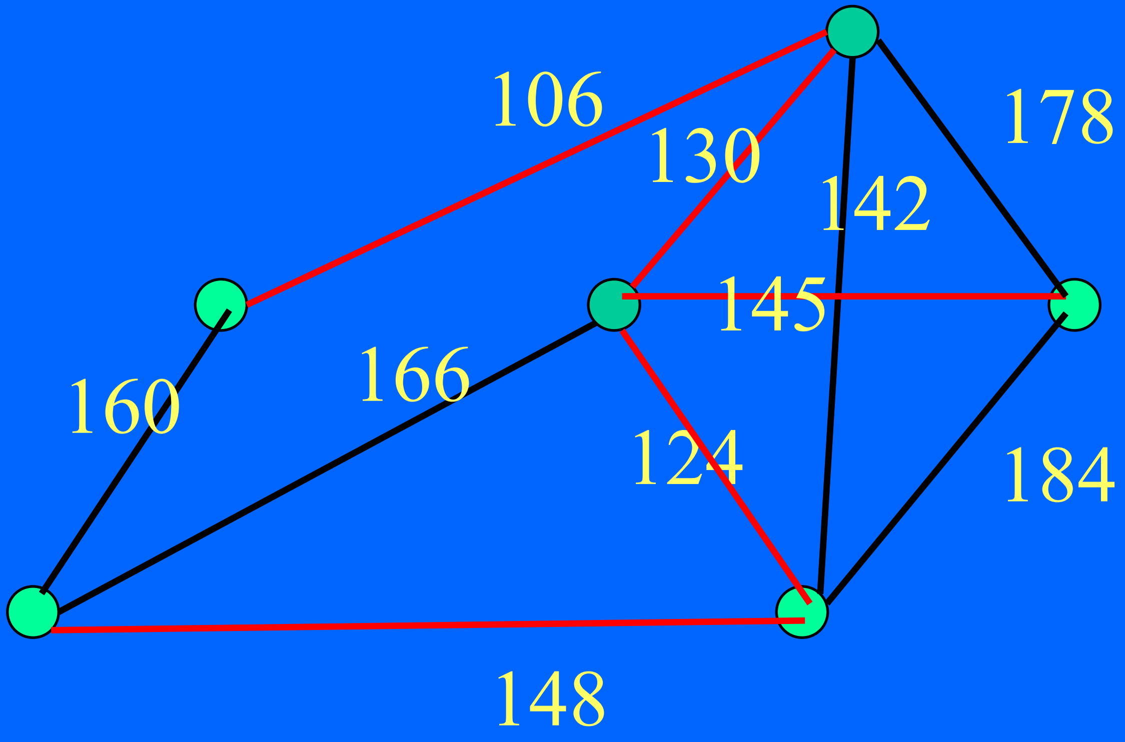
- Consider the admissible transformation $\varphi(x) = 3x + 100$.
- We now get the above numbers on edges.

# Related Example: Minimum Spanning Tree Problem



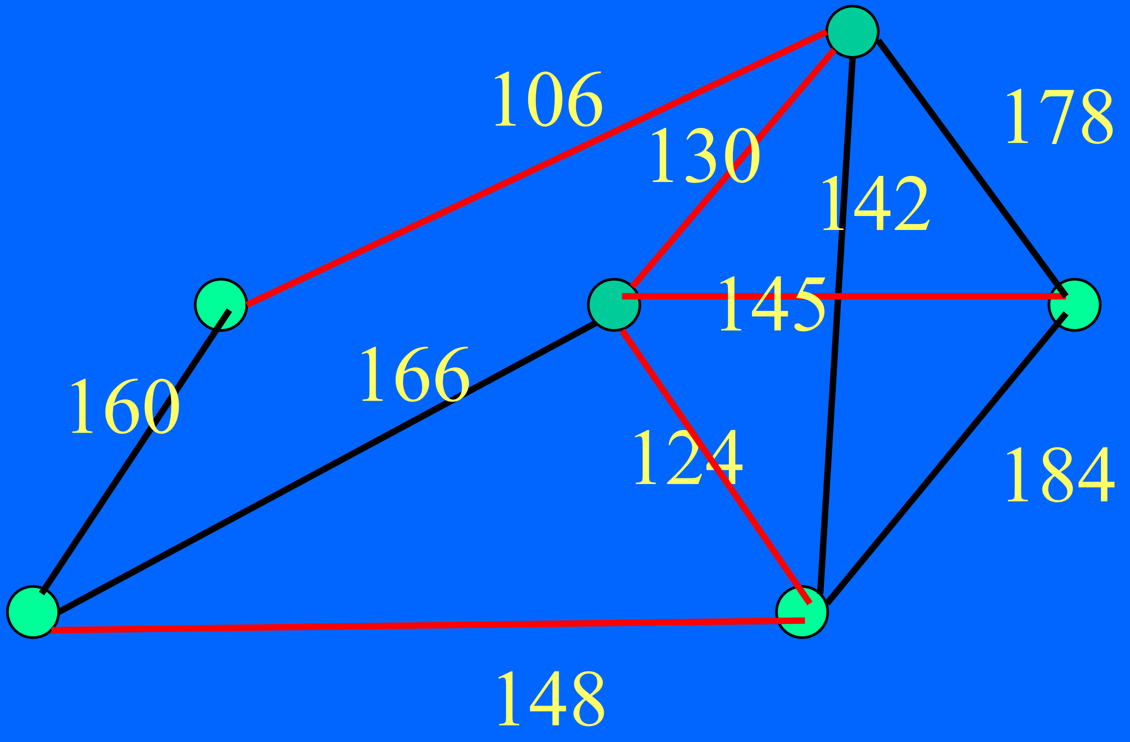
- The minimum spanning tree is the same.

# 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. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"
6. Optimization Problems in Epidemiology
7. **Meaningfulness of Statistical Tests**
8. How to Average Scores

# Meaningfulness of Statistical Tests

(joint work with Helen Marcus-Roberts)

•Biostatistics a key component of epidemiological research.

•However, biostatisticians know virtually nothing about measurement theory.

•Most have never heard about the theory of meaningfulness or limitations that meaningfulness places on conclusions from 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.

<u>Frederic Lord</u>: 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: Controversy

# 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

$H_0$: 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.

# Outline

1. Theory of Uniqueness of Scales of Measurement/Scale Types
2. Meaningful Statements
3. Averaging Judgments of Cough Severity
4. Measurement of Air Pollution
5. Evaluation of Alternative HIV Treatments: "Merging Normalized Scores"
6. Optimization Problems in Epidemiology
7. Meaningfulness of Statistical Tests
8. **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 have described will help.

•Let $a_1, a_2, \ldots, a_n$ be "scores" or ratings, e.g., scores on criteria for evaluating treatments.

•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 $F:(\Re^+)^n \to \Re^+$ has the following properties:

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

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

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

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

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

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.)

# How Should we Average Scores?

## A Functional Equations Approach

Example: $u = F(\mathrm{a})$. Assume $a$ and $u$ are ratio scales.

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

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

# 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.

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

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

$\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_n > 0$, $\alpha > 0$, $\alpha$ depends on $a_1$, $a_2$, …, $a_\mathrm{n}$.

- 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

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

$\lambda$, $c_1$, $c_2$, …, $c_n$ constants

# How Should We Average Scores?

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:

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

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

# How Should We Average Scores?

Functional Equation:

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

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

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

Solutions to (****) (Even Without Continuity Assumed)

(Aczél, Roberts, and Rosenbaum):

$$F(a_1,a_2,\ldots,a_\mathrm{n}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_\mathrm{i} a_\mathrm{i} + b$$

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

# How Should We Average Scores?

Theorem (Aczél and Roberts):

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

$$\Sigma \lambda_{\mathrm{i}} = 1, b = 0$$

(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

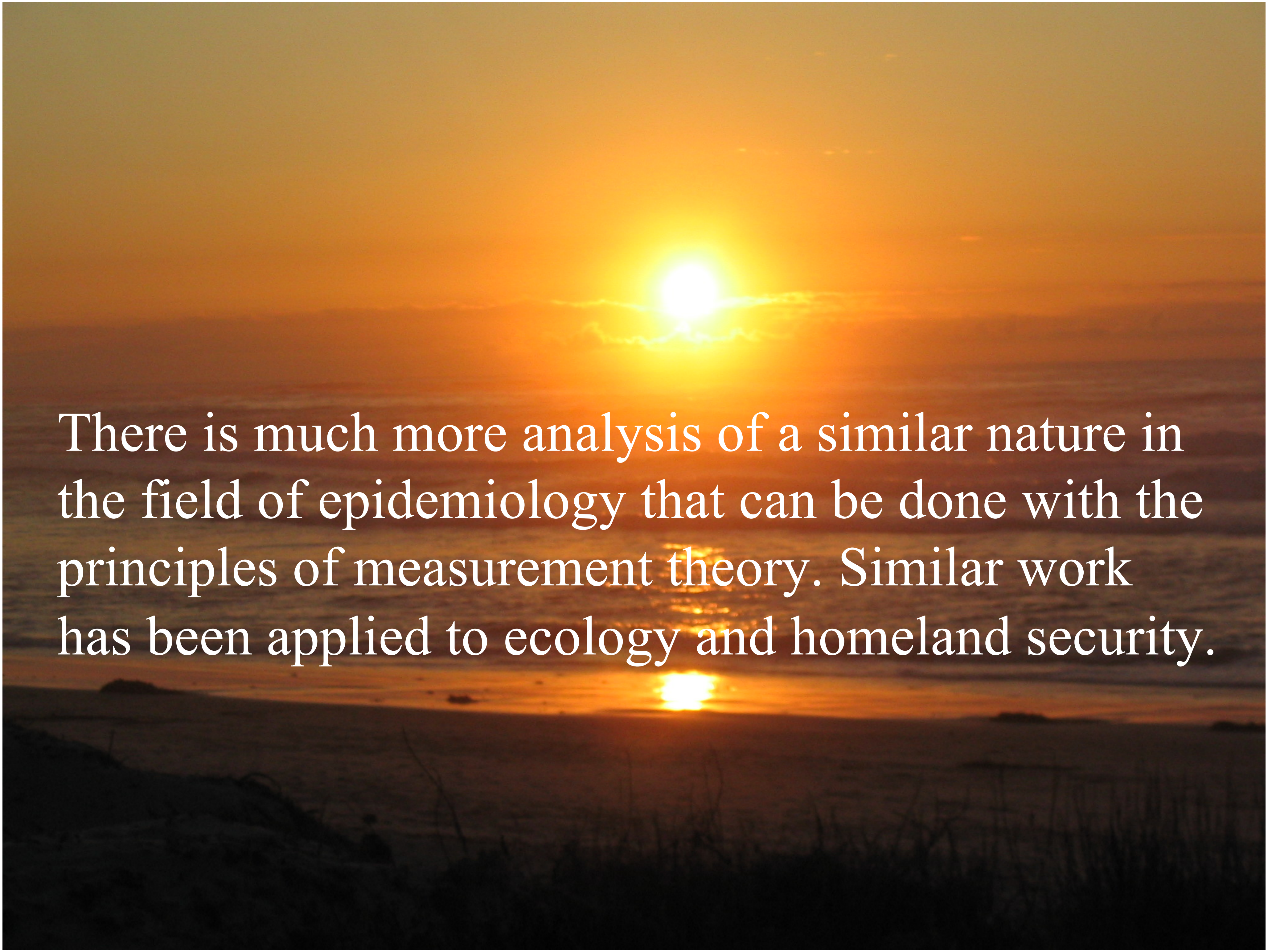
$$F(a_1,a_2, \ldots, a_n) = kF(b_1, b_2, \ldots, b_n)$$

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.



There is much more analysis of a similar nature in the field of epidemiology that can be done with the principles of measurement theory. Similar work has been applied to ecology and homeland security.