

Lesson 4/Week 4

Sources of Data, Data type, and Geographic Visualization

- ✓ Medical Geographic Data Sources
- ✓ Analytical Methods
- ✓ Types of Maps
- ✓ Rates and Ratios
- ✓ Age Standardization

Data sources, USA

Two types of data

- ✓ Primary data
- ✓ Secondary data (inc. vital registration)

Secondary data

National - example

<http://www.cdc.gov/nchs/data/hus/hus10.pdf#001>

<http://www.cdc.gov/mmwr/PDF/wk/mm5754.pdf>

State Health Departments

County Health Departments (death registration and vital registration + reportable illnesses)

Global - United Nations, World Health Organization

Primary data

Original research (you create your own data)

Primary data:

Population-based (covering all residents of a city, county, etc)

Disadvantages

Takes time

Costs more

Advantages: Covers entire population

Sample-based (a subsection of the population is covered)

Advantages

Takes less time

Costs a lot less

Disadvantages

Only a fraction of the population covered but a well-designed study can reduce errors and produce results nearly as good as population-based studies.

Sampling:

Type of sampling – simple random, stratified, purposive

Hypothesis testing: is the sample a true representation of the entire population?

Significance tests, confidence intervals.

What do medical geographers do with disease data?

Trend-analysis: Are morbidity/mortality rates increasing or decreasing?

Spatial patterns: Maps are what geographers do best.

Dot maps (Graduated symbols and pie chart)

Choropleth maps

Isolines (also known as isopleths or isarithms). See map below.

Lines used in **flow maps** to show volumes and directions

Buffers (GIS): Example₁: Areas/population within 10 miles of the source of an epidemic.

Example₂: Areas/populations within half a mile of major freeways (exposed to pollution from exhaust emissions)

Relationships: Relating morbidity/mortality observations/data to socio-economic/environmental variables.

Correlation

Regression

Spatial autocorrelation

Question? Can we apply remote sensing/digital image interpretations?

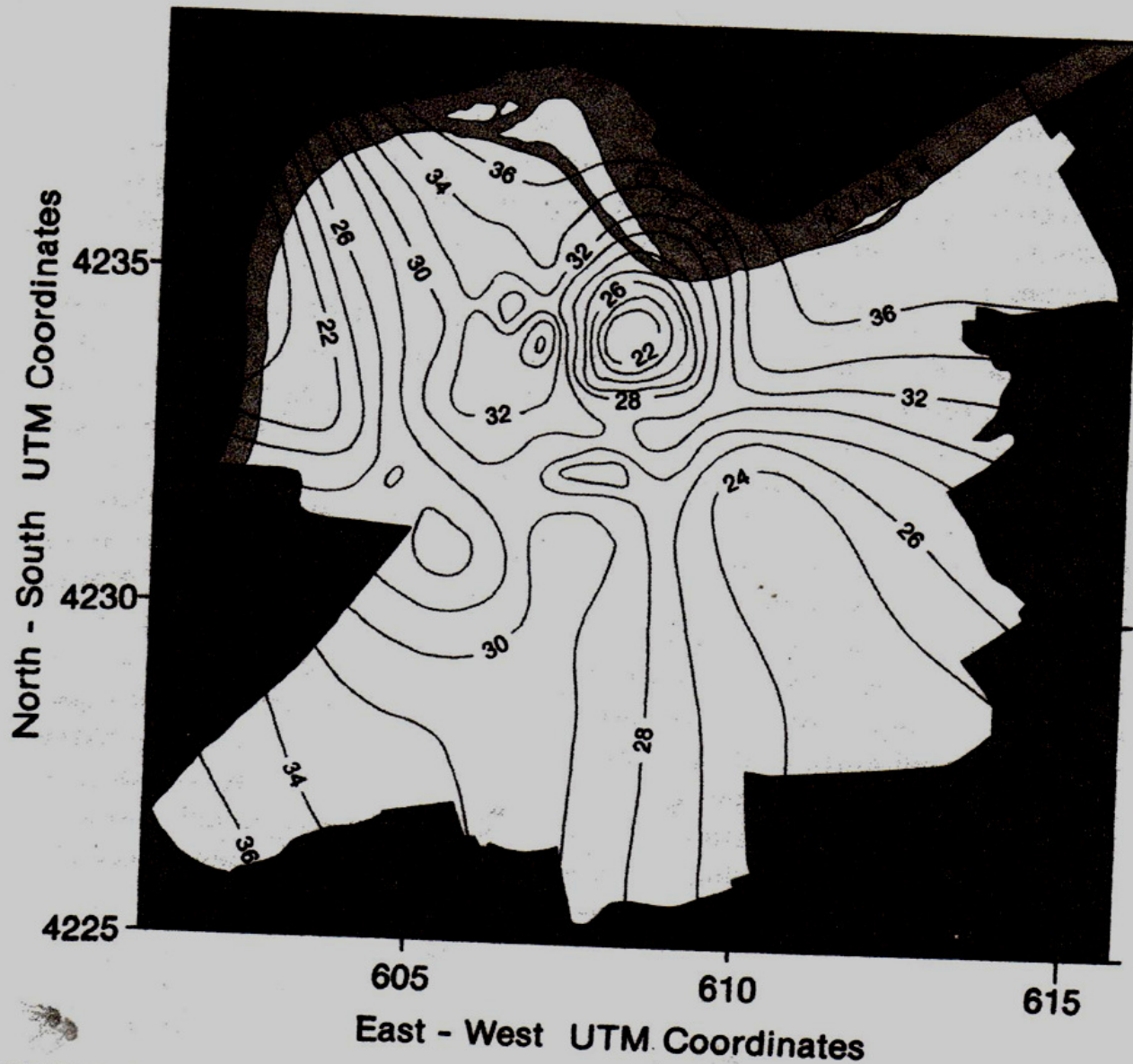


FIGURE 13-4. Lines of equal pediatric blood lead levels, Louisville, Kentucky, 1979. Values are micrograms per deciliter of blood.

More on Maps:

Scale:

Small scale - 1:250,000 - 1: 1,000,000,000

Medium scale - 1: 25,000 - 1 : 250,000

Large scale - 1: 200 - 1 : 25,000

Critical Internals

“A map made well can communicate a lot of information. This information, however, can be distorted if the intervals are not appropriate to the data distribution”. Example: too many or too few intervals. **See text p.452-453**

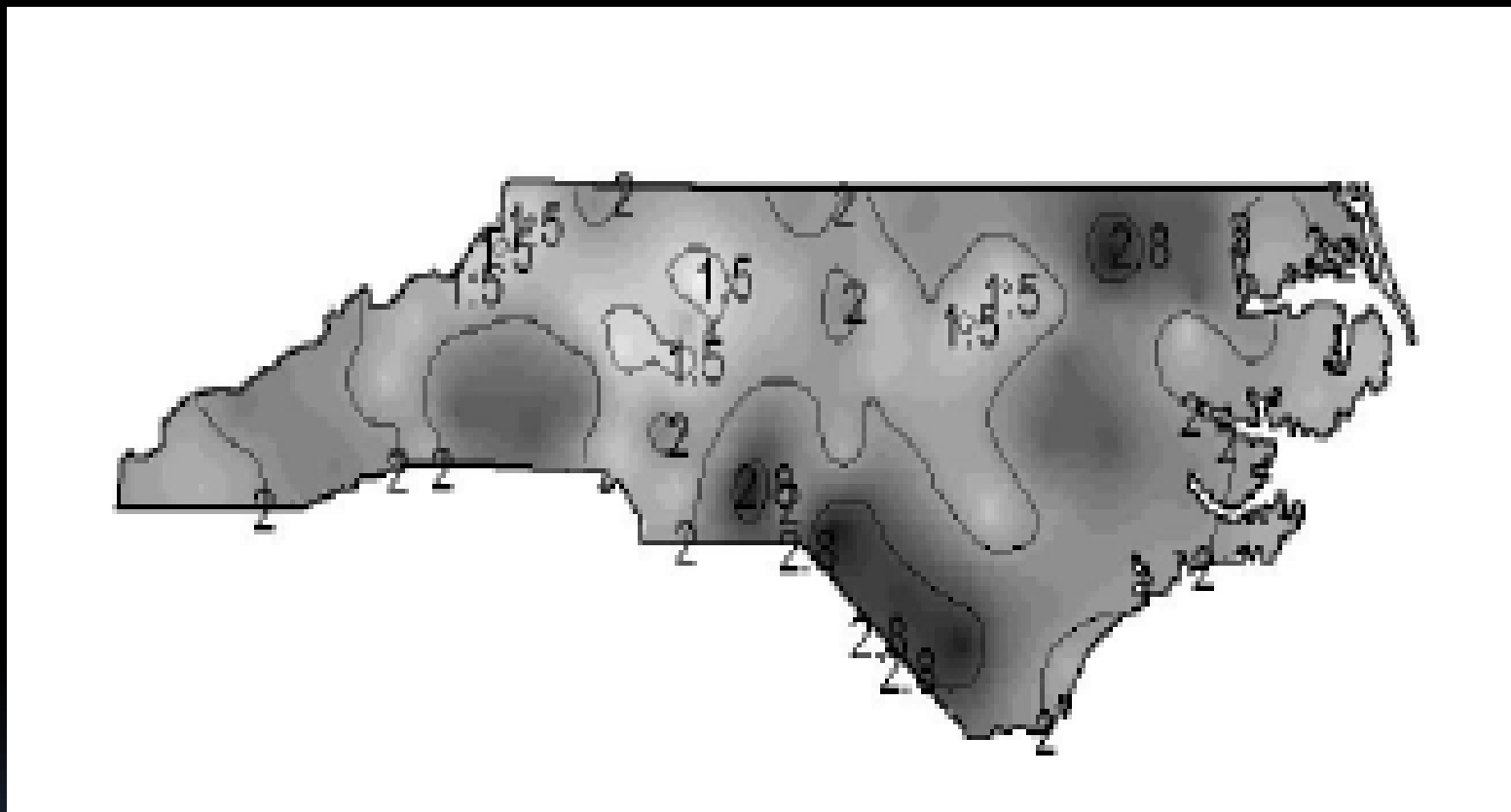
“The mind can only differentiate pattern from six or seven categories in black and white, and fewer in color”.

Trend-surface analysis: Regression over space

A set of X and Y coordinates for location in space and a Z axis of data values, which (like elevation), create a surface of highs and lows based on distance from the origin (0.0). Most appropriate for data that are continuous over space or can be validly interpreted to be so.

Trend-surface analysis maps

<http://www.ij-healthgeographics.com/content/pdf/1476-072X-3-18.pdf>



Infant mortality. North Carolina

Use of Graphs and Statistics

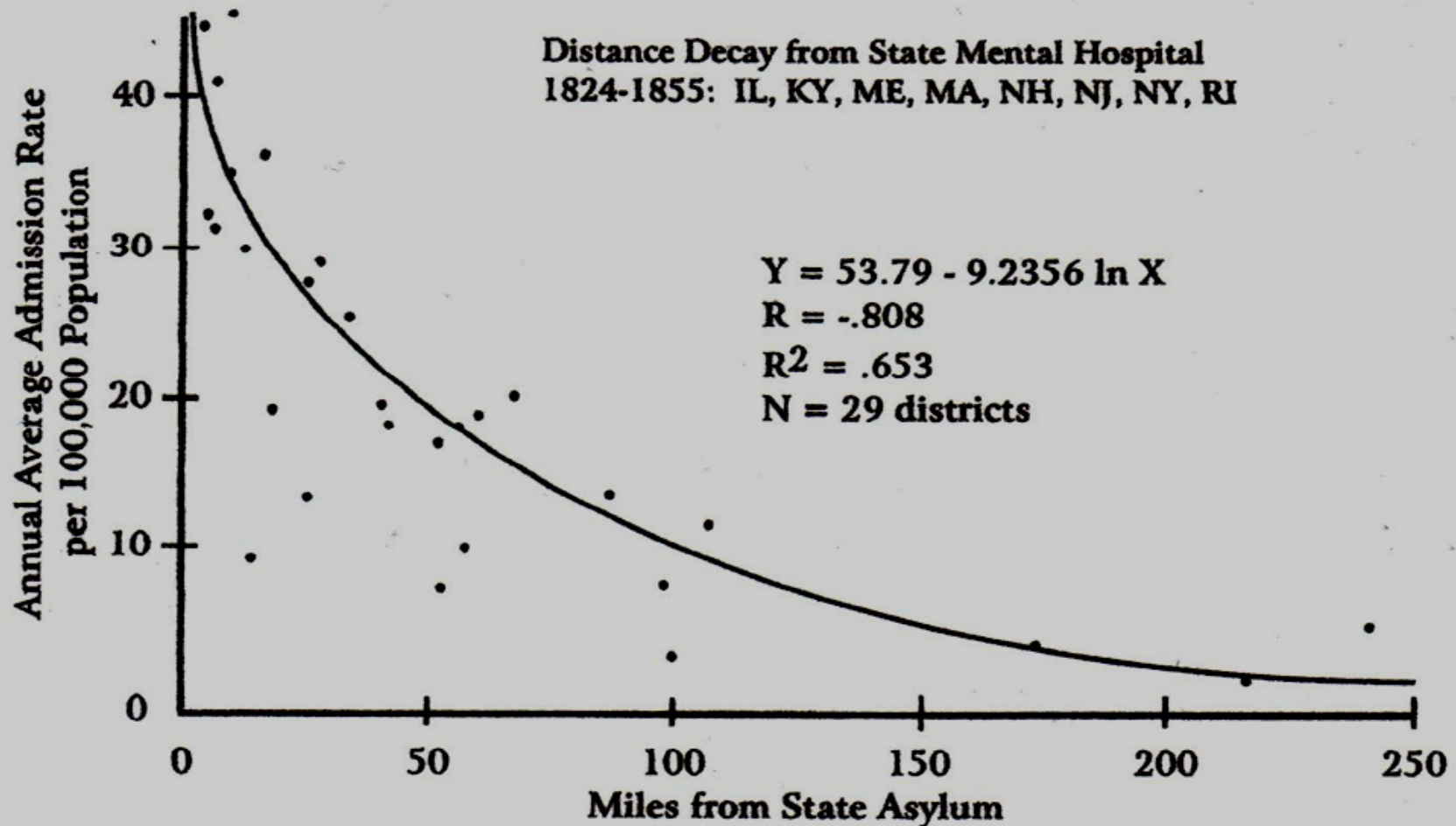


FIGURE 11-1. A distance decay curve of admissions to state mental hospitals in selected states in the 19th century. From Hunter, Shannon, and Sambrook (1986, p. 1046). Copyright 1986 by Elsevier Science. Reprinted by permission.

Heart disease rate twice higher near freeways

Watch TV news clip (February 15, 2010)

http://abclocal.go.com/kabc/story?section=news/health/your_health&id=7279121

“We've known for a long time that automobile pollution can cause respiratory problems. A new study from the University of Southern California and University of California-Berkeley points to more life-threatening issues. They've concluded that living too close to a freeway -- within 100 yards -- can cause strokes and heart disease. ”

More on data.....

“Although many health scientists speak about studying and promoting health, it is usually diseases and the risk of disease that gets measured, discussed, and portrayed in charts and graphs”

Rates and Ratios

No. of disease events for a specified time and place

Rate

Total population exposed during the specified time in that place

Example : Crude death rate ($D/P \times 1000$)
 Age-specific death rate ($D_a/P_a \times 1000$)
 Infant mortality rate ($D_{0-1}/B \times 1000$)
 Case mortality (fatality) rate ($D_{\text{cause}}/P \times 1000$)

Ratio:

Population per physician ratio

Population per bed

Relative Risk: The incidence of disease in one group divided by the incidence of disease in another group.

Relative risk

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

Disease is breast cancer (BC). A woman is considered to be exposed if she gave birth at or after the age of 25.

| exposure or factor | disease or condition | | total |
|--------------------|----------------------|-----------------|-------|
| | BC ⁺ | BC ⁻ | |
| Birth ≥ 25 | 31 | 1597 | 1628 |
| Birth < 25 | 65 | 4475 | 4540 |
| total | 96 | 6072 | 6168 |

Odds-ratio

$$OR = \frac{P(\text{disease}|\text{exposed})/(1 - P(\text{disease}|\text{exposed}))}{P(\text{disease}|\text{unexposed})/(1 - P(\text{disease}|\text{unexposed}))}$$

$$\begin{aligned} RR &= \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})} \\ &= \frac{(A)/(A + B)}{(C)/(C + D)} \\ &= \frac{31/1628}{65/4540} \\ &= 1.33 \end{aligned}$$

<http://www.biostat.wisc.edu/~kendzior/STAT541/lc4.short.pdf>

Odds Ratio and Relative Risk

Relative Risk - RR

In [statistics](#) and mathematical [epidemiology](#), **relative risk (RR)** is the risk of an event (or of developing a disease) relative to exposure. Relative risk is a [ratio](#) of the [probability](#) of the event occurring in the exposed group versus a non-exposed group.

$$RR = \frac{P_{\text{exposed}}}{P_{\text{non-exposed}}}$$

Consider an example where the [probability](#) of developing lung cancer among smokers was 20% and among non-smokers 1%. Here, $a = 20(\%)$, $b = 80$, $c = 1$, and $d = 99$. Then the relative risk of cancer associated with smoking would be

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{20/100}{1/100} = 20.$$

Another term for **relative risk** is **risk ratio** because it is the ratio of the risk in the exposed, divided by the risk in the unexposed.

In this example, smokers would be twenty times as likely as non-smokers to develop lung cancer.

Odds Ratio - is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

$$\frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 / q_1}{p_2 / q_2}$$

| | Total | Cancer Yes | Cancer No |
|---|----------------------|-------------------------|-------------------------------|
| Example: Smokers/non-smokers | | | |
| | Smokers | 100 (a) | 20 (b) (a - b) 80 |
| | Non-smokers | 100 (c) | 1 (d) (c - d) 99 |
| | | 0.2 (b/a) | 0.8 (a - b/a) |
| | | 0.01 (d/c) | 0.99 (c - d/c) |
| | Relative risk | 20 (0.2/0.01) | 0.25 (0.2/0.8) |
| | | | 0.0101 (0.01/0.99) |
| | Odds Ratio | | 24.75 (0.25/0.0101) |

An odds ratio of 1 indicates that the condition or event under study is equally likely to occur in both groups.



An odds ratio greater than 1 indicates that the condition or event is more likely to occur in the first group.

And an odds ratio less than 1 indicates that the condition or event is less likely to occur in the first group.

Similarities and differences:

“Since relative risk is a more intuitive measure of effectiveness, the distinction is important especially in cases of medium to high probabilities. If action A carries a risk of 99.9% and action B a risk of 99.0% then the relative risk is just over 1, while the odds associated with action A are almost 10 times higher than the odds with B.”

See calculation
below

| | | Total | Dead | Alive |
|---------------|--|-------|-------|-------|
| Age 105-109 | | 1000 | 999 | 1 |
| Age 100-104 | | 1000 | 990 | 10 |
| | | 0.999 | 0.001 | |
| | | 0.99 | 0.01 | |
| Relative risk |  | 1.009 | 999 | |
| | | | 99 | |
| Odds Ratio |  | | 10.09 | |
| | | | | |
| | | | | |

Example 2. Survivors of the Titanic:

<http://www.childrens-mercy.org/stats/journal/oddsratio.asp>

“ There were 462 female passengers: **308 survived** and **154 died**. There were 851 male passengers: **142 survived** and **709 died** (see table below).

Clearly, a male passenger on the Titanic was more likely to die than a female passenger. But how much more likely?

You can compute the odds ratio or the relative risk to answer this question. The odds ratio compares the relative odds of death in each group. For females, the odds were exactly **2 to 1 against dying ($154/308=0.5$)**. For males, the odds were almost **5 to 1 in favor of death ($709/142=4.993$)**. The odds ratio is **9.986 ($4.993/0.5$)**. There is a **ten-fold greater odds of death for males than for females**.

The relative risk (sometimes called the risk ratio) compares the probability of death in each group rather than the odds. For females, the probability of death is **33% ($154/462=0.3333$)**. For males, the probability is **83% ($709/851=0.8331$)**.

The relative risk of death is **2.5 ($0.8331/0.3333$)**. There is a **2.5 greater probability of death for males than for females**.

| | Alive | Dead | Total |
|--------|-------|------|-------|
| Female | 308 | 154 | 462 |
| Male | 142 | 709 | 851 |
| Total | 450 | 863 | 1,313 |

Example 3: Calculate the odds ratio

Disease is breast cancer (BC). A woman is considered to be exposed if she gave birth at or after the age of 25.

| exposure or factor | disease or condition | | total |
|--------------------|----------------------|--------|-------|
| | BC^+ | BC^- | |
| $Birth \geq 25$ | 31 | 1597 | 1628 |
| $Birth < 25$ | 65 | 4475 | 4540 |
| total | 96 | 6072 | 6168 |

$$\begin{aligned} RR &= \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})} \\ &= \frac{(A)/(A + B)}{(C)/(C + D)} \\ &= \frac{31/1628}{65/4540} \\ &= 1.33 \end{aligned}$$

Odds ratio calculation below

| | | | | |
|---------------------------|----------------|--|---------------|---------------|
| P(Diseased/Exposed) | 31/1628 | | 0.019 | |
| 1 - P(Diseased/Exposed) | 1-0.019 OR | | 1597/1628 | 0.981 |
| | | | | |
| P(Diseased/Unexposed) | 65/4540 | | 0.0143 | |
| 1 - P(Diseased/Unexposed) | 1-0.0143 OR | | 4475/4540 | 0.9857 |
| | | | | |
| | | | | |
| Numerator | 0.0194 OR | | 0.019/0.981 | |
| Denominator | 0.0145 OR | | 0.0143/0.9857 | |
| | | | | |
| Odds Ratio (OR) | | | | 1.338 |
| | 0.0194/0.0145 | | | |

Data Problems:

Age as a confounding variable.

“Age is one of the most common and important confounding factors in health studies. Age can confound comparisons when the groups being compared have different age distributions and age is related to the outcome of interest (e.g. death or the prevalence of disease).”

http://www.cdc.gov/nchs/tutorials/nhanes/NHANESAnalyses/AgeStandardization/age_standardization_intro.htm

Crude death Rate (PRB: 2010): USA=8 Mexico= 5 Nicaragua = 4

Solution: Age Standardization

“**Age standardization** is a method that allows you to take away the confounding effect of age in order to allow you to make fair comparisons.” **Age-standardized mortality rates** are used to compare the mortality rates of places without being skewed by the difference in age distributions from place to place.

Standardized rates are favored over Crude Rates because they take age groups into consideration, e.g. one should not compare a population of over-70s to a younger population in another region.

The use of a standard population is needed when comparing the mortality rates of differing population groups to discount the effect of age on mortality. Without this standardization it would be unclear if differing mortality rates were due to age or other factors.

TABLE 12-1. Age-Specific Death Rates for Three Hypothetical Counties

| Age | Population | Deaths (per 1,000) | Age-specific death rate |
|-----------------|------------|--------------------|-------------------------|
| County A | | | |
| 0-1 | 300 | 15 | 50.0 |
| 1-9 | 2,250 | 11 | 4.8 |
| 10-19 | 1,700 | 5 | 2.9 |
| 20-29 | 1,400 | 2 | 1.4 |
| 30-39 | 1,350 | 1 | 0.7 |
| 40-49 | 1,200 | 1 | 0.8 |
| 50-59 | 1,000 | 2 | 2.0 |
| 60+ | 800 | 6 | 7.5 |
| Total | 10,000 | 43 | CDR = 4.3/1,000 |
| County B | | | |
| 0-1 | 600 | 30 | 50.0 |
| 1-9 | 3,500 | 17 | 4.8 |
| 10-19 | 1,950 | 6 | 3.0 |
| 20-29 | 1,400 | 2 | 1.4 |
| 30-39 | 1,050 | 1 | 0.9 |
| 40-49 | 700 | 1 | 1.4 |
| 50-59 | 500 | 1 | 2.0 |
| 60+ | 300 | 2 | 6.7 |
| Total | 1,000 | 60 | CDR = 6.0/1,000 |
| County C | | | |
| 0-1 | 300 | 4 | 13.3 |
| 1-9 | 2,250 | 2 | 0.9 |
| 10-19 | 1,700 | 1 | 0.6 |
| 20-29 | 1,400 | 2 | 1.4 |
| 30-39 | 1,350 | 3 | 2.2 |
| 40-49 | 1,200 | 5 | 4.2 |
| 50-59 | 1,000 | 10 | 10.0 |
| 60+ | 800 | 16 | 20.0 |
| Total | 10,000 | 43 | CDR = 4.3/1,000 |

TABLE 12-2. Direct Age Standardization for the Death Rates of Table 12-1

| Age | Standard population (A + B + C) | Population A | | Population C | |
|-------|---------------------------------|-------------------------|-----------------|-------------------------|-----------------|
| | | Age-specific death rate | Expected deaths | Age-specific death rate | Expected deaths |
| 0-1 | 1,200 | .0500 | 60.0 | .0133 | 16.0 |
| 1-9 | 8,000 | .0048 | 38.4 | .0009 | 7.1 |
| 10-19 | 5,350 | .0029 | 15.5 | .0006 | 3.2 |
| 20-29 | 4,200 | .0014 | 5.9 | .0014 | 5.9 |
| 30-39 | 3,750 | .0007 | 2.6 | .0022 | 8.2 |
| 40-49 | 3,100 | .0008 | 2.5 | .0042 | 13.0 |
| 50-59 | 2,500 | .0020 | 5.0 | .0100 | 25.0 |
| 60+ | 1,900 | .0075 | 14.3 | .0200 | 38.0 |
| Total | 30,000 | | 144. | | 116. |

Note. Total population crude death rate per 1,000 = $146/30,000 = 4.87$.

Standard death rate per 1,000: population A = $144/30,000 = 4.8$.
 population B = same ASDR = 4.8.
 population C = $116/30,000 = 3.8$.

Standard rate ratio
 population A = $4.8/4.87 = 1$.
 population B = $4.8/4.87 = 1$.
 population C = $3.87/4.87 = 0.8$.

| Age | Population County B | ASDR County C | Observed D in CC | Expected D in CC |
|---|---------------------|---------------|------------------|------------------|
| 0-1 | 600 | 0.0133 | 4 | 7.98 |
| 1-9 | 3,500 | 0.0009 | 2 | 3.15 |
| 10-19 | 1,950 | 0.0006 | 1 | 1.17 |
| 20-29 | 1,400 | 0.0014 | 2 | 1.96 |
| 30-39 | 1,050 | 0.0022 | 3 | 2.31 |
| 40-49 | 700 | 0.0042 | 5 | 2.94 |
| 50-59 | 500 | 0.01 | 10 | 5 |
| 60+ | 300 | 0.02 | 16 | 6 |
| Total | 10,000 | | 43 | 30.51 |
| If county C had the youthful age structure of county B | | | | |
| its total number of death (for the entire population of 10,000) | | | | |
| would be 30.51. In other words, if its age structure can be "younged" (not a real word) | | | | |
| (or made youthful), its overall mortality will drop from 43 down to 30.51 | | | | |
| (a decline of 30%). | | | | |

Note: The text (p. 415) shows a difference of only 20% for all counties but in this case we are talking about the gain all three counties will register in overall mortality reduction if the combined populations of the three counties had the lower age specific death rates of county C.

More on data problems:

Numerator problems: The numerator of a rate is composed of events such as the number of deaths or cases of illnesses from specific causes. Numerator errors could include:

- ✓ Reporting errors (respondents withholding information or giving a wrong answer)
- ✓ Incompleteness of coverage (esp. in LDC)
- ✓ Diagnosis varies over time and space (change of disease classification over time)
- ✓ Type of laboratory test, equipment, lab tech or physician. etc.
- ✓ The degree to which laws mandate or permit autopsies
- ✓ Death from multiple causes

Denominator problems: The denominator of a rate represents the population at risk of an event.

Lack of accurate population data (LDC) – incomplete coverage, age misreporting, etc.

Small populations – sudden spike in denominator (e.g. tourists, sudden influx of migrants) or sudden jump numerator (e.g. 5 teens out of 20 residing in small village die in a single auto accident – rate 250/1000)

Data aggregation based on administrative divisions – Imagine two adjacent counties through which a river flows. Imagine further that encephalitis is epidemic to an equal extent in both counties but one county is predominantly urban (high density) and the other is predominantly rural (low density). Reported (county-level) rates would be much higher in the former than the latter even though the disease experiences of the populations along the river banks in the localities affected are identical or nearly identical.