A MILLION LITTLE THINGS

epidemiology basics for non-epidemiologists

Lindsay R. Townes, MPH
Child & Adolescent Health Epidemiologist
Michigan Department of Health & Human Services

1

About me

- BA in Anthropology from Wayne State University; MPH in Epidemiology from the University of Michigan School of Public Health
  - Finished doctoral studies in Epidemiological Sciences at UM-SPH as ABD 😎
- Taught several classes to graduate and undergraduate students including:
  - Introductory statistics and biostatistics (so, so many times)
  - Epidemiology for non-majors
  - Introductory SAS coding
- Currently work as the Child and Adolescent Health Epidemiologist at MDHHS where I provide epidemiological and statistical support to programs including:
  - Child and Adolescent Health Centers
  - School Hearing and Vision
  - Taking Pride in Prevention
  - Michigan Adolescent Pregnancy & Parenting Program
  - Title V

"And it was so typically brilliant of you to have invited an epidemiologist."

https://www.art.com/products/p15063320605-sa-i6845061/william-hamilton-and-it-was-so-typically-brilliant-of-you-to-have-invited-an-epidemiologist-new-yorker-cartoon.htm

2

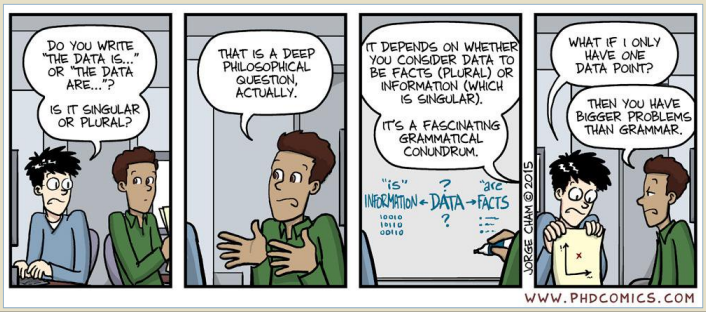
Agenda

- Why work with data?
- Data sources
- Basic statistical concepts and terms
- Basic epidemiological concepts and terms
- Using data in your work

3

BUT THE VERY FIRST THING!!

- 'Data' is the plural of datum
- If you wish to impress your epidemiologist and statistician friends, say 'These data **are/show**…' and <u>not</u> 'The data is/shows…'
  - We have had it brutally trained into us and cannot change



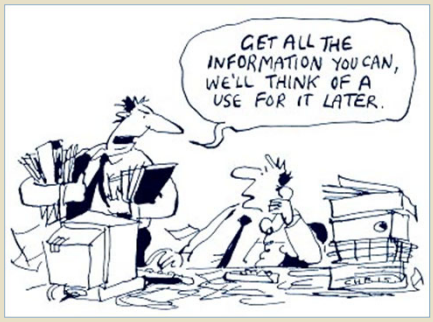
4

WHY WORK WITH DATA?

5

Working with Data as Non-Data Professionals

- What are typical interactions with data as a non-data professional?
  - Collection
    - Routinely-collected administrative data
    - Surveillance data
    - Clinical data
  - Reporting
    - Summaries to funding agencies, government, or fiduciaries
    - Internal use
  - Review
    - For clinical decision-making
    - For quality improvement

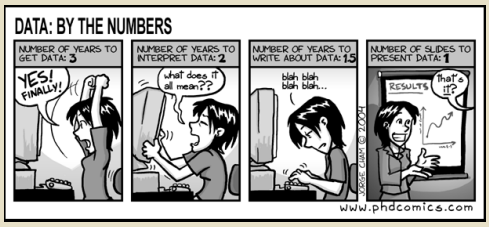


https://www.slamcartoon.com/media/fb2a1807-2d58-46e0-9351-fea6eb899fa4-get-all-the-info-you-can

6

Why work with data: Research

- Research:
  - Seeks to answer specific, bounded questions
    - NOT 'how do we prevent pregnancy in adolescents?'
    - BUT 'does offering no-cost hormonal contraception on-site increase hormonal contraception use among facility users compared to the previous fiscal year?' is more workable
  - Can be done with existing data sources
    - I.e., it is not always necessary to do primary data collection to conduct research



7

The work of research

- Generate a research question – the more specific, the better
- Come up with a sound methodology designed to answer that question
  - What data do you need to directly answer your question?
  - What data do you need to make sure you're not accidentally measuring something else?
  - How to analyze those data
  - Your friendly neighborhood epidemiologists can help with all this!
- Analyze data and report results
  - Does not have to be complicated!
    - Excel, for example, has a number of basic analytic functions
  - Does not need to be published in a journal!
    - Intra-office use, poster presentations, white papers – all valid ways of reporting research results

8

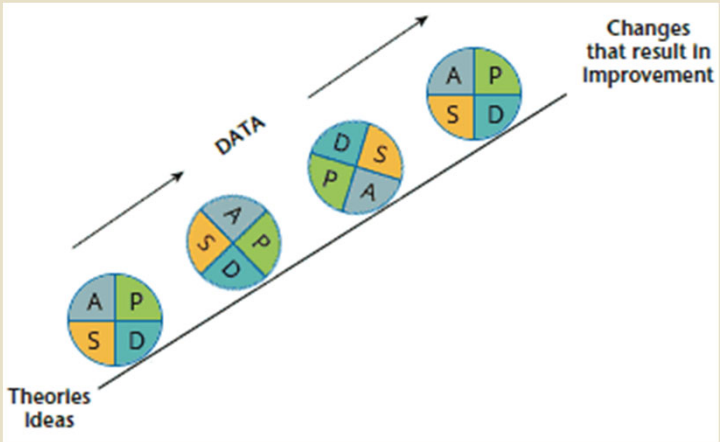
Why work with data: Evaluation

- Evaluating whether a program or intervention is effectively and efficiently doing what it's meant to is crucial for ensuring participants are well-served by the program
- Heavily data-driven process – where does data come in at each of these stages?
  - Needs assessment: what does the target population actually need?
  - Theory assessment: are our goals feasible, well-defined, and responsive to the population's needs?
  - Implementation assessment: are we reaching the people that need reaching with services they need?
  - Impact/effectiveness assessment: has our target population changed in any way as a result of our program?
  - Efficiency assessment: how's that cost-benefit ratio looking?
- Evaluation should be a holistic practice, with each step relying & building on work done in previous steps
  - What data are needed and how to collect them should be developed long before it's time to evaluate effectiveness and efficiency

http://allsformativeassessment.blogspot.com/2013/01/look-at-cartoon-linked-here-and-explain.html

9

Why work with data: CQI



PDSA cycle. Boston, Massachusetts: National Institute for Children's Health Quality; (2015). (Available on www.NICHQ.org)

- Continuous quality improvement (CQI) is the iterative practice of evaluating current practices for potential sites of improvement, with continuous monitoring and re-evaluation
- As with evaluation and research, data play a role in each step:
  - **Plan:** Assess needs and outcomes from previous cycles
  - **Do:** Implement changes then observe, document, and analyze outcomes
  - **Study:** Report analysis results and determine whether changes had the intended effects
  - **Act:** Make modifications and begin the cycle again

10

10

Why work with data: Takeaways

- Working with data provides information that allows us to better serve our populations of interest
- Develop data collection and analysis plans from the beginning and integrate throughout the life cycle of a project
  - It's not hard to do, but does require some thoughtful planning
- The next sections will talk about some things to consider as you work with data
  - What type of data to collect or use
  - Some basic analyses
  - How to interpret results

11

DATA SOURCES

12

Data sources

- The four main types of data sources we'll be discussing today are:
  - Research
  - Surveillance
  - Publicly-available
  - Routinely-collected
- NOT mutually exclusive categories – data can be routinely-collected via surveillance and made publicly-available
- Two other important data terms:
  - **Raw data:** typically unprocessed individual-level data
    - May have issues that require data cleaning or suppression of private information
  - **Aggregate data:** summarizes individual-level data into groups of interest, e.g., all Michigan residents, males or females, persons above or below a certain age, etc
    - May also require suppression depending on organizational rules

13

Surveillance data collection

- Public health surveillance is the **continuous**, **systematic** collection, analysis, and interpretation of health-related data needed for the **planning**, **implementation**, and **evaluation** of public health practice (WHO 2019)
  - Typically mandated by law or regulation
- Two varieties: **passive** and **active**
  - **Passive**: typically reported to public health agencies by health care providers
    - Pros: simple, inexpensive
    - Cons: incomplete data of uncertain quality
  - **Active**: health agencies seek out data from health care providers and the general public
    - Pros: more complete and standardized reporting
    - Cons: time-consuming, expensive
- Irrespective of how data are collected, should always lead to action from public health agencies
  - Descriptive reporting
  - Monitoring trends and patterns in disease and risk factors
  - Provide data to programs and policy-makers
  - Evaluate prevention and control efforts

https://www.ebr.com/i-love-ya-but-youre-strange-that-time-spider-man-fought-the-measles/

Centers for Disease Control and Prevention (CDC). Introduction to Public Health. In: Public Health 101 Series. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2014. Available at: https://www.cdc.gov/publichealth101/surveillance.html.

14

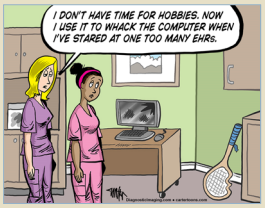
Surveillance data

- Usually collected by government agencies – CDC, MDHHS, etc
  - Can be collected by universities or non-profits, usually in collaboration with government
- Can be broadly representative of the underlying population of interest assuming:
  - Little-to-no regional, racial/ethnic, gender, etc disparities in surveillance of conditions of interest
  - Active surveillance has a large enough dragnet
- NOT typically tailored to answer a specific question
  - For example, certain infectious disease diagnoses are reported in order to detect potential outbreaks, not to determine whether some specific thing is putting people at increased risk
    - That typically comes a bit later in an outbreak investigation!
- Often made publicly-available in aggregated form
  - Individual-level data *may* be available on request but typically requires a DUA and institutional permissions

15

Administrative data

- Generated at the time of provision of health care services, hospital admission or discharge
  - Contain information about primary and secondary diagnoses, procedures performed, medications received
  - Often linked to demographic data (eg, age, sex, race)
  - Do not typically include information on laboratory test results, radiology or other imaging, or clinical measures
  - Typically made available from insurers, including private payers, Medicaid, Medicare, and the VA
- Pros: low-cost to obtain, potentially widespread population coverage
- Cons: questionable accuracy, limited information available, not generated to answer research questions
- Administrative data are used in health initiatives such as:
  - Healthy People 2020
    - Data from National Hospital Discharge Survey
    - Useful because helps to highlight geographical and racial differences in the provision of health care
  - Health Plan Employer Data and Information Set (HEDIS)
    - Quality assurance for HMOs



https://www.diagnosticimaging.com/article/radiology-comic-backhanding-ehr

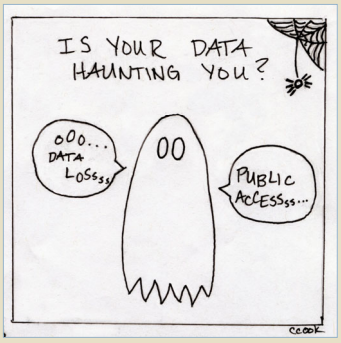
16

Research data

- Data collected via research surveys, observational studies, and experiments
- Often publicly-available only in aggregated format
  - Tables and charts in publications
- However, there is a growing movement among researchers to make data available
  - May be able to share cleaned, suppressed data either publicly or with interested researchers, members of the public, stakeholders, etc
    - Depends on a whole bunch of alphabet soup issues: IRB, DUA, etc
- Data are often tailored to answer specific questions about specific populations and may not be generalizable to your population of interest
  - For example, a sexual health education program that was fairly successful in rural Greece may not be as useful in Detroit Public Schools
- Data are only as good as the design of the study
  - poorly done study=poor quality data

17

Publicly-available data



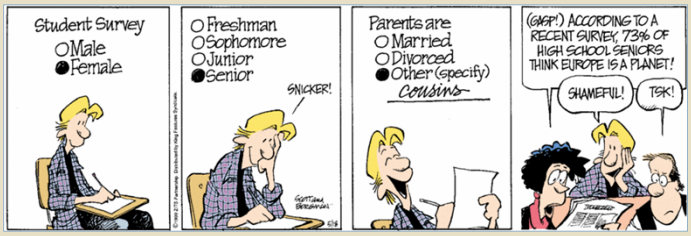
http://researchdata.wisc.edu/news/happy-halloween-from-rds/

- Most often from government organizations
- Publicly-available data are typically:
  - Aggregated (although not always)
  - Clean (although not always)
  - A few years out of date (basically always)
- Pros:
  - Free, simple to access, downloadable in many formats, sometimes comes with snazzy online tools to run basic analyses
- Cons:
  - Older, may not provide precisely the information needed (eg, you want data on 15-19-year-olds, they have data on 13-19-year-olds), often suppressed for low population groups/areas

18

A brief aside about survey data

- Surveys are an important tool for just about every type of data collection
- HOWEVER and ALAS survey response rates have plummeted over the past decades – why?
  - Privacy concerns
  - Feeling over-surveyed
  - Caller ID



https://www.comicskingdom.com/zits/

- Survey researchers employ a wide variety of tactics to improve rates and/or appropriately weight respondents to better represent the population of interest BUT
  - Beware of surveys with small number of respondents or high margins of error!
  - Beware of survey items that got low response rates in otherwise high-response surveys!
  - Think carefully about who was surveyed and how
- This is prologue to my next slides which are basically all about surveys!

19

Useful and good publicly-available data sources: ACS and US Census

- **American Community Surveys/US Census:** https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml & https://data.census.gov/cedsci/?intcmp=aff_cedsci_banner
  - Contains data aggregated at a variety of geographic levels for a wide array of socioeconomic, demographic, and other indicators
  - Note: small population areas/groups may need to use 5-year ACS or older Census data due to suppression rules
  - Data can be downloaded, analyzed, mapped – or simply reported as needed (eg, for a needs assessment)

https://peanuts.fandom.com/wiki/April_1980_comic_strips

20

Useful and good publicly-available data sources: YRBS

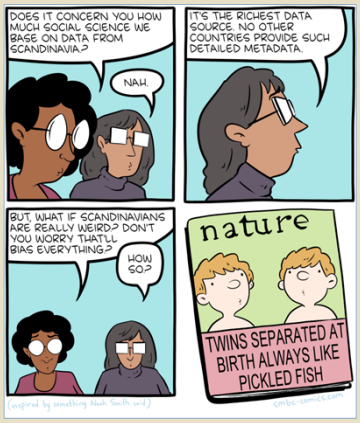
- **Youth Risk Behavior Surveys (YRBS):** https://www.cdc.gov/healthyyouth/data/yrbs/data.htm
  - Provides individual responses at the state-level (and aggregated for Detroit) to survey items from samples of adolescent students
    - Includes items on sexual behavior, diet and exercise, drug/alcohol/tobacco use, family life, bullying, etc
  - Non-data people: use Youth Online analysis tool: https://nccd.cdc.gov/youthonline/App/Default.aspx
    - Provides state-level results for survey items by year, race, grade, sex, sexual orientation/identity (if desired)
  - If you want analyses based on something other than race and sex, requires some statistical coding knowledge
    - For example, you'd like to see if respondents who ride in cars without seatbelts are different from those who always use seatbelts when it comes to sexual behaviors
    - This is not something Youth Online can do for you, but your neighborhood friendly epidemiologist might be able to!

#SpideyYOLO

21

Useful and good publicly-available data sources: MDHHS Vital Records

- **MDHHS Birth Data:** https://www.mdch.state.mi.us/pha/osr/chi/births14/frameBxChar.html
  - Aggregated data available by year at city, county, local health department and state levels
  - Wide variety of variables including:
    - Demographic: (race/ethnicity, age, marriage status)
    - Socioeconomic (non-US born, Medicaid, education)
    - Health risk factor (tobacco use, prenatal care, diabetes, hypertension)
    - Outcomes (preterm birth, low birthweight birth, C-section)
  - Can look at different groups based on race, age group, prenatal care, birthweight, weeks gestation
  - Unlike YRBS, individual-level birth data are not publicly-available
    - Analyses that cannot be done through Vital Records public-facing sites require making data requests to your friendly neighborhood epidemiologist



22

BASIC STATISTICAL CONCEPTS

23

What is statistics, exactly?

Statistics is a science that deals with the **collection**, **analysis**, **interpretation**, and **presentation** of data

In public health, we usually seek to understand a **population**, the set of all people who meet a given criterion

Since it is hard to get data on literally everyone, we typically rely on **samples**, a portion or subset of the population of interest, to make **inferences** about that population

A **statistic** is a metric derived from sample data and is an estimate of the population **parameter**

Example:

Population of interest: every Michigan resident

Sample: residents in 10 Michigan counties

Parameter: average age of all Michigan residents

Statistic: average age of residents in 10 sampled counties

24

Two major types of data

Categorical (Qualitative)

- Includes characteristics that cannot be defined numerically such as:
  - Gender
  - Marital status
  - City of residence
  - Ethnicity
- Has limitations for what kinds of analysis can be done
- Often used to **stratify** numeric data into different groups
  - For example, breastfeeding initiation rates by different ethnic groups

Numerical (Quantitative)

- Includes characteristics that do have meaningful numeric definitions such as:
  - Height and weight
  - Age
  - Income
- Two major forms: **discrete** and **continuous**
  - **Discrete:** counts (no decimal values possible)
    - # of live births is a discrete variable
  - **Continuous:** measurements (decimals or fractional values possible)
    - Weight is a continuous variable

*A special case is **ordinal data** which are an ordered series of relationships or ranks*

- *Likert scales are a common source of ordinal data ("On a scale of 1 to 5…")*

25

Some important kinds of statistics

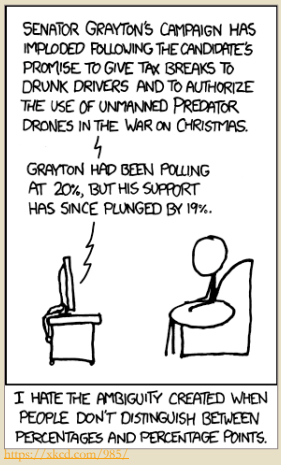
The following are all considered **descriptive** statistics:

- **Measures of central tendency** are ways of summarizing data sets into a single metric such as:
  - **Mean (AKA average):** sum of all measurements divided by the number of observations
    - Can only be used with numeric or ordinal data
    - Sensitive to the presence of **outliers**, values that are much higher or lower than the other values in a data set
  - **Median**: middle value that separates the higher and lower half of the data set
    - Not sensitive to the presence of outliers
  - **Mode:** most frequent value in the data set
    - Can be used to summarize categorical data
- **Dispersion** are measures of the spread of values in a data set
  - **Range:** the difference between the largest and smallest values
  - **Variance:** measures the average squared difference of values from the mean of the set
  - **Standard deviation:** the average distance that values vary from their mean

26

Relative vs. absolute percentage change

- **Absolute percentage change:** the simple difference in percentage between two groups or two times for an indicator
  - Unit: **percentage points**
- **Relative percentage change:** expresses the absolute change as a percentage of the value for an indicator at Time 1 (or for Group 2 relative to Group 1)
  - Unit: **percentage**
  - Calculation: (Time 2 – Time 1)/Time 1 or (Group 2 – Group 1)/Group 1
- Typically more useful to report the relative percentage change
  - For example, an increase from 5% to 10% and 85% to 90%
    - Both are a 5-point increase, but one is a more significant increase, which can be captured by calculating the relative percentage change
      - 100% increase vs. 5.9% increase



https://xkcd.com/985/

27

Quick & Easy Calculation: Absolute v Relative Percentage Change

	2016	2017
% of births to mothers ages 15-19 that were repeat births, Michigan	17.7%	15.9%

- Calculate first the absolute percentage change between 2016 & 2017 for repeat births to mothers ages 15-19.
  - 15.9 – 17.7 = **1.8 percentage point decrease**
  - **Interpretation:** "The percentage of births that were repeat births to mothers ages 15-19 dropped by 1.8 points from 2016 to 2017"
- Now calculate the relative percentage change for the same period:
  - (15.9 – 17.7)/17.7 = **10.2% relative decrease**
  - **Interpretation:** "The percentage of births that were repeat births to mothers ages 15-19 dropped by 10.2% from 2016 to 2017"

https://xkcd.com/1252/

28

Statistical hypothesis testing

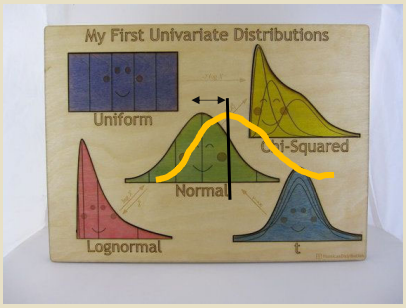
- In **statistical hypothesis testing**, our final aim is to determine whether the difference between two or more variables is greater than what we would expect from chance alone
  - These are **inferential statistics**, to be contrasted with the descriptive statistics discussed earlier
  - The **null hypothesis** is what we try to disprove – often a statement of no difference or no change between exposed and unexposed groups (or among groups over time)
  - The **alternative hypothesis** is most frequently a statement of what the null hypothesis does not equal
    - For example, if the null hypothesis is that a coin has a 50% chance of flipping heads, the alternative would be that the chance of flipping heads does not equal 50%
- A result is **statistically significant** when we reject the null hypothesis because the distribution of our data is sufficiently different than the distribution of expected values if the null hypothesis were true
  - I am 100% not going into the details of this, I swear, I just needed to give you the context for the next bit

https://xkcd.com/892/

29

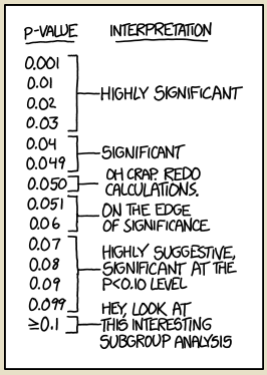
An extremely brief nod to data and probability distributions

- **Inferential statistics** is based on the comparison between the distribution of actual data points and mathematical probability curves (**probability distributions**)
  - Each probability distribution has its own set of assumptions about the nature of the comparison data
  - Using the wrong distribution for your data will give you wrong results
    - Another opportunity to reach out to your friendly neighborhood epidemiologist!
  - Allows data scientists to determine whether and to what extent our actual data represents something different from what would be expected by random chance



30

Statistical significance – what it means, what it does not



- When someone reports a statistically significant difference/change/etc, they are typically relying on the **p-value**, which is calculated after comparing our data to the null distribution, producing a test statistic, and oh, I swore I wouldn't get into this!
- For reasons as silly as they are arbitrary, the agreed-upon cutoff for statistical significance is $p<0.05$
  - In reality what this represents is a less than 1 in 20 chance that our data are consistent with the null hypothesis, given that the null hypothesis is actually true
  - The smaller the p-value, the lower the probability that the null hypothesis is actually true
- A low p-value does not, in itself, prove that an association between exposure and outcome is real or clinically important
  - Issues with study design, sample size, and effect size are important to consider
    - P-values are dependent on sample size, for example, and in very large samples, very small differences in effect size (e.g., 1 mmHg in blood pressure) will come up as significant at $p<0.05$

31

Assessing trend

- **Trend** is an upward or downward shift in data over time
- **Trend analysis** quantifies and explains trends and patterns over time
  - I do not approve of trend analysis with <5 years of data
- While the eyeball test can suggest whether data are trending in one direction or another, formal statistical tests are needed to determine whether the trend is significant
  - This typically requires the use of a statistical software package, such as SPSS, R, Stata, or SAS
  - Tests can include:
    - Cochran-Armitage test for trend
    - Kendall's Tau correlation
    - Mann-Kendall trend test
  - While Excel can produce a trendline (TREND function), it does not include information about statistical significance
  - If your office does not have a statistical programmer or a statistical software package, feel free to reach out to your friendly neighborhood epidemiologist!

32

BASIC EPIDEMIOLOGY CONCEPTS

33

Basic epidemiology

- **Basic terminology:**
  - **Outcome:** broad term representing the disease, state of health, condition, etc. of interest
  - **Exposure:** a thing that may affect the risk of developing the outcome
  - **At-risk**: any person who does not currently have a disease, but could potentially develop it is at-risk
- **Measures of occurrence:**
  - **Prevalence:** the proportion of people with a condition at a single point in time
    - ONLY tells us proportion, does not get at risk of developing the disease
  - **Incidence:** measures the occurrence of new cases of a disease over time
    - Can describe risk of developing a disease, but much harder to get these data than prevalence

https://www.youtube.com/watch?v=1jzZe3ORdd8

34

Counts vs. ratios vs. rates

- **Counts** are the most basic unit in epidemiology
  - Number of cases of a disease or occurrence of other health events (eg, live birth)
  - Counts are also useful for demonstrating the **magnitude** of a public health issue
  - By themselves, counts cannot describe **risk** of a given health outcome (need a denominator)
- **Ratios** are the relative magnitude of two values
  - Calculated by dividing one count by another
  - **Proportions** are a special class of ratios where the numerator is included in the denominator
    - Percentage if the count is sensible over a denominator of 100 (ie, 1+case per 100 people)
    - Where there are fewer than 1 case per 100 people, it may be necessary to report larger denominators to get meaningful numbers
      - 1 case per 1,000 or 100,000 people is common because people do not readily grok 0.001%
    - Used to summarize **prevalence** data
- **Rates** are the frequency that an event occurs in a defined population over a specified time
  - Counts of new diagnoses or events are divided by the population at-risk for those conditions over a set time

35

Consider the following!

- "100 cases of flu occurred in pregnant people"
  - Is this statement informative – why or why not?
- **It is not!**
  - We don't know what the time frame is – 1 year? 1 week? Today? 10 years ago?
  - We don't know what the population is – Ann Arbor? Michigan? The US? The whole Earth?
- Let's assign a time frame – 1 year (2017)
  - And a population – pregnant people in Ann Arbor (~1000 in 2017)
  - And stipulate that the whole population is at-risk
- "Among 1000 pregnant people in Ann Arbor in 2017, 100 developed flu."
  - Or 'The incidence of flu among pregnant people in Ann Arbor in 2017 is 100 cases per 1,000 pregnant people'
    - This is a **rate**

36

Measures of association

- Epidemiologists use ratios to determine **associations** between **exposures** and **outcomes**
- Takes the form of:

  Occurrence among the exposed/Occurrence among the unexposed
  - Where occurrence can be incidence, prevalence, risk, rates, odds…
- A ratio can be interpreted as approximating the **risk** of a disease given exposure to some variable of interest
  - A ratio of 1 indicates no difference in risk between the exposed and unexposed groups
  - A ratio >1 indicates higher risk among the exposed than the unexposed
  - A ratio <1 indicates lower risk among the exposed than the unexposed
    - Can you think of a scenario that might produce a ratio that is <1?

37

Example: measures of association

- We have data on teen births for the state of Michigan for 2017, and have been asked whether race is associated with teen birth rate.
- Here's a table with the relevant data:

Race	Population, females, ages 15-19	Live births to females ages 15-19	Birth rate per 1,000 females ages 15-19
White	469,924	2573	**11.3**
Black	112,012	1757	**31.9**

- Dividing the Black rate by the White rate, we find that Black females ages 15-19 were 2.8 times as likely to deliver a live birth as White females ages 15-19 in Michigan in 2017
  - This value is known as the **incidence rate ratio**
  - Incidentally, this ratio was statistically significant at $p<0.01$

38

Why do epidemiologists write like that?

- In normal human conversation, it is commonplace to hear words like 'risk' and 'odds' bandied about willy-nilly, but alas, epidemiologists are no longer normal humans
- A good epidemiologist writing about a measure of association will include:
  - The population(s) of interest
  - The measure of occurrence
  - The time period of interest
  - And end up with a monstrous sentence like
    "Among females ages 15-19 who delivered a live birth in Michigan in 2017, non-Hispanic Black mothers had 1.43 times the risk of delivering a low birthweight baby than non-Hispanic White mothers ($p<0.05$)"
- Why do we do this?
  - Well, it's important to be clear exactly who, what, where, and when the populations of interest are
  - That number may not apply to older mothers, different years, different states, and all of those conditions need to be included

39

USING DATA IN YOUR WORK

40

Holistic data collection and management

- Now that we are comfortable with types of data, where to get them, how to work with them, and what they mean, we can start thinking about integrating that knowledge in practice
- Things to think about – where can you get data to learn:
  - Who are you trying to serve?
  - What are some of the broad issues your community is facing?
  - What services do they need?
  - What services of yours are they actually using?
  - What improvements are you seeing in your patients, participants, etc?
  - What improvements does that lead to in the larger community?
- The earlier in the process you can determine what to measure and how, the easier every other step in the process becomes
  - Spare yourself the pain of frantically running last-second surveys because you forgot to ask for something crucial from Day 1! (ASK ME HOW I KNOW)

41

Practice good data hygiene

- Data hygiene and clean data are terms used to describe taking a systematic approach to maintaining reliable, accessible, functional data sets
  - A clean data set should have no errors or duplicates, be properly formatted for analysis, address outlying and missing values, and so on
- Excel is not an adequate data management platform
  - Changes are invisible – if a column or row of data is deleted, we don't know why or how it happened
  - Preferable to use Access, SAS, or R, which maintain records of how data were changed from their original forms
- No matter what system you use, the most important things to remember:
  - Always keep an unchanged version of the original data set to check against later edits
  - Maintain a record of how changes were made

42

Simple analyses can tell you a lot

- Compare the mean values between two groups (ethnicity, gender, age, etc)
  - Identifies possible health disparities
  - Suggests intervention points in a community
  - Significance testing: Student's t-test (=T.TEST in Excel)
- Run a correlation between two variables (age and outcome, year and outcome, etc)
  - While not identifying causal factors, a correlation provides evidence that two variables are associated
  - =(CORREL in Excel)
- Plot data over time to visualize trend
  - Statistical significance testing requires some sort of statistical software programming
- The next slide shares some other handy Excel analytic formulas

43

Basic analyses in Excel

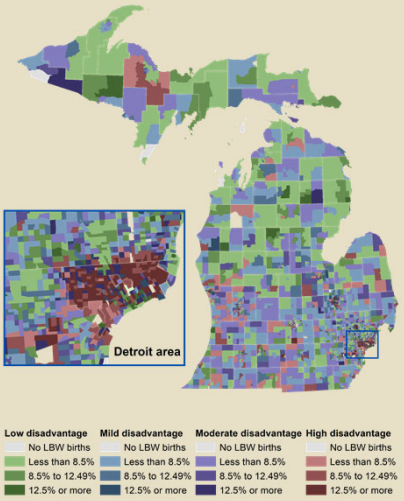
- Average: =AVERAGE(A2:A))
  - Sum of all values divided by number of values
- Median: = MEDIAN(A2:A))
  - All values are ordered low to high, then middle value (or mean of middle 2 if an even number of values) reported
  - Not affected by outlier values in the same way that mean/average is
- Correlation: =CORREL(A2:A, B2:B)
  - Measures the degree to which two variables fluctuate together,
    - Correlation between -1 and 1; 0=no correlation
  - Positive correlation: as one variable increases in value, so does the other
  - Negative correlation: as one variable increases in value, the other decreases
- More complex analyses available with Analysis ToolPak, including:
  - Linear regression
  - Statistical significance testing
  - ANOVA
    - Compares means from 2 or more groups in a sample

https://xkcd.com/1906/

44

Getting more from your data

- Augment your routinely collected data with publicly-available data!
- Here's an example using a combination of data from MDHHS (Vital Records) and the ACS
  - 5 variables from ACS 5-year Census tract data to create the Concentrated Disadvantage Index
  - Low birth weight 5-year rate by Census tract
- ACS data: **identify high need areas**
- Vital Records data: **identify areas with high rates of adverse health outcomes**
- This kind of work can be done at the county and local health department scale



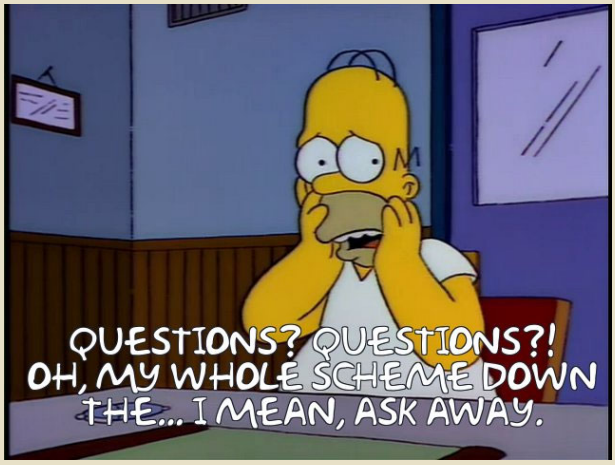
45

Here to help: your friendly neighborhood epidemiologist

- You **can** request data and analyses from MDHHS
  - We are a public service, and our aim is to improve the whole population's health
  - Data requests can include statistical analyses, summary tables and charts, and geographic mapping of exposures and outcomes
- For child and adolescent health data requests, you can contact me at townesl@michigan.gov
  - If it's not in that wheelhouse, I may be able to direct you to the right person!
- Data requests can take some time, depending on the type of data requested, so be aware of that

46

Thank you! Any questions?



47