Measures of Association

Questions from Last Class?

- Syllabus update pending, we do have class next week

Why We Care About Associations

- A huge amount of epidemiology can be accomplished given only what we covered last week
- But it is largely limited to chronicling disease, rather than being able to address it
- Associations can tell us about *who* or *how* or *when* an outcome occurred, rather than just when it occurred

Proportion Surviving
1.00
0.98
0.96
0.94
0.92
0.90
0.88
30
32
34
36
38
40
42
44
Age

Exposures

- Last week, we talked primarily about outcomes
  - Something that happens (typically health related)
- Now we also must consider exposures
  - A property of an individual or group of individuals
  - Age, occupation, geographic location, whether a proceeding event occurred (smoker vs. non-smoker), etc.
- As discussed last week, there are some parts of epidemiology that are exposure focused, rather than outcome focused

Modifiability

- Modifiability is whether someone's exposure can be intervened on, or modified
- "Vaccinated against COVID-19" is a modifiable exposure
- "Born in 1975" is not a modifiable exposure
- Is it useful to know things about non-modifiable exposures?

Where This Gets Tricky

- Some exposures are a basket of multiple, potentially complex exposures
- Classic examples:
  - Gender and Biological Sex
  - Race
  - Socioeconomic status

Socioeconomic Status

- “Low Socioeconomic Status is associated with X outcome”
  - What does that mean?
  - Is it just money
  - Is it exposure to poorer maintained housing, more polluted neighborhoods, etc.?
  - Is it different jobs?
  - Is it access to healthcare, education, etc?
  - Is it chronic stress?
  - Is it malnutrition?

Intersection with Modifiability

- Is “Race” or “Gender” modifiable?
- Can can they be broken down into components, some of which are modifiable, and some of which are not?
- Modifiable by whom, and over what time scale?

The 2x2 Table

	Y=1	Y=0	
X=1	A	B	A+B
X=0	C	D	C+D
	A+C	B+D	A+B+C+D

Note that many of the measures discussed last week can be obtained by looking only across the lower margin

Stratification

- Looking across each row of X, you can get the measures of prevalence or incidence *among that value of X*
- Incidence proportion in the exposed:
  - A/(A+B)
- Incidence odds in the unexposed:
  - (C/(C+D)/(1-(C/(C+D)
- Note: These formulas assume the cells are in a specific order. If you re-orient the 2x2 table, you have to rethink the resulting calculations

Example

- 2000 individuals in Pullman, some of whom are exposed to a contaminated water source on Pioneer Hill, and resulting GI disease

	Y=1	Y=0	
X=1	85	915	A+B
X=0	40	960	C+D
	A+C	B+D	A+B+C+D

Risk (i.e. Incidence Proportion)

- Exposed: 85/1000= 0.085
- Unexposed: 40/1000 = 0.040

Risk Difference

- Risk in the Exposed – Risk in the Unexposed
- 0.085 – 0.040 = 0.045
- The risk difference between those exposed to the Pioneer Hill water supply vs. those who were not was 0.045.
- Sometimes written as "…was 4.5%", as all of these can be validly presented as percentages
- There is an ambiguity here, because risk ratios (coming up next!) are *also* expressed as percentages, so as general practice, I would avoid this

Risk Ratio

- Risk in the Exposed / Risk in the Unexposed
- 0.085 / 0.040 = 2.125
- The risk of GI illness in those exposed to the Pioneer Hill water supply was 2.125 that of those who were not.
- "…was 1.125 times higher than that…"
- "…was 212.5% that…"
- "…was 112.5% higher…"

RD vs. RR

- Both measures are valuable and say something
- RD gives you a notion of the absolute difference
- RR can sometimes exaggerate the effect of a rare vs. very rare outcome
- 0.0001 vs. 0.00001
  - RR = 10
  - RD = 0.00009
- Note the “null” (no difference) is different. RD = 0, RR = 1.
- Ranges are also different. RD can go from 1 to -1, RR can go from 0 to infinity.
  - This can be tricky. Which is a bigger deal, a RR of 3, or an RR of 0.25?

Risk Differences are Awesome!*

- Risk differences seem great, so why are they so rare?
- Mathematics (and thus statistics) does not conform to our desires
- Risk differences, while relatively intuitive, are much more difficult to estimate using regression models
- Risk ratios may be more meaningful when talking about an individual

Odds Ratios

	Y=1	Y=0	
X=1	85	915	
X=0	40	960	C+D
	A+C	B+D	A+B+C+D

- Odds in the Exposed: (85/1000)/(1-(85/1000) = 0.093
- Odds in the Unexposed: (40/1000)/(1-(40/1000)=0.042
- Odds Ratio: 0.093/0.042 = 2.21

RR vs. OR

- RR = 2.125
- OR = 2.21
- Explain why these are similar?

Why Would You Ever Use an OR?

- *Almost no one* thinks an odds ratio is intuitive, so why would we ever use it?
- Again, the answer is “the math is easier”. Logistic regression, which estimates ORs, is considerably easier than binomial regression, which estimates RRs.
- Also, for case-control studies, where you cannot observe the full cohort, ORs become necessary because you cannot directly calculate risk

Incidence Rates

- We need to add one more column, for person years. Let's say people were enrolled over *approximately* a two-year period

	Y=1	Y=0	Person-years	Incidence Rate per 10 person-years
X=1	85	915	1894	0.45
X=0	40	960	1913	0.2

Incidence Rate Ratio (IRR) = 0.45/0.02 = 2.25

DO NOT CALL THIS ANYTHING ABBREDIATED "RR"

Relative Times and Survival

- Time to first calving for two herds of Bantha

- Long term risk is clearly not useful – by age 12, all animals have calved
- We can calculate a risk difference at a time
- Or we can calculate a time difference
- Or the ratio versions

- Risk at 8 Years:
  - Blue: 0.63
  - Red: 0.43
  - RD: 0.2
  - RR:1.47
  - Where did these numbers come from?
- Median Time
  - Blue: 7 years
  - Red: ~9 years
  - Time Difference: -2
  - Time Ratio: 0.78
  - Why are these flipped?

Hazard Ratios

- Hazard: The instantaneous probability of an event occurring, given it did not occur at an earlier time point
- Hazard ratio: A ratio of hazards
- Hazard ratios are… hazardous, and unintuitive, and should be approached with caution. We'll discuss them in the survival analysis section
- I take a somewhat less dire view of them than some in epi, because they are extreme useful in modeling

Break Time!

Associations vs. Causes

- "X is associated with Y" means that for some level of X, Y occurs more or less often than another level of X
- Association and correlation are often used interchangeably, though correlation is often assumed to be a linear association, and associations can be *much* more complex
- Association does not imply causation
- *But* all causal relationships are inherently also associations

Predictive and Descriptive

- Descriptive Epidemiology: As with last class, primarily concerned with describing associations without necessarily assigning causal arguments to them
- Predictive Epidemiology: Diagnostics and prognosis. Causal relationships are *helpful* but not necessarily required
  - Note this can lead you badly astray
  - Rotavirus and prom dresses

The Problem with Intervening

- If you intervene on something that isn't a cause, it isn't going to matter
- Banning prom will not eradicate rotavirus
- To intervene effectively, we need a notion of *cause*

Counterfactuals

- Counterfactual questions are at the core of causal inference and underly most medical research.
- What if a patient had been given Treatment A instead of Treatment B
- What if someone had never smoked?
- These are impossible to answer – but important to articulate

Potential Outcomes

- A phrase often associated with counterfactuals
- What *can* happen
- For many health states, this is binary.
- $Y^{X=0}$ vs. $Y^{X=1}$
- If Sarah had been assigned to the treatment arm of an RCT (X=1) and *not* gotten COVID-19, that is her *factual* outcome
- Her outcome with X=0 is thus the unobservable counterfactual

“Well Defined Interventions”

- This is, essentially, the same problem as the idea of modifiability
- Currently an open debate in the literature
- Genetic epidemiology seems a clear place where causal inference is possible absent a well defined intervention
- Social epidemiology is a considerably bigger struggle

Where This Gets Tricky

- Consider whether obesity (as measured by BMI) causes heart disease
  - Leaving aside the actual question, take this as true
- Factual outcome:
  - $Y^{obese=1}$
- What is the counter-factual?
  - $Y^{obese=0}$
- What intervention does that describe?
- Ideas?

Individual Causal Effect

- This is the difference between an individual's potential outcome for X=1 and their potential outcome for X=0
- Barring a time machine, this cannot, but definition, be observed
- But it's kind of what we want, and it's *absolutely* what medicine wants

Average Causal Effects

- Aka “Population Average Causal Effects”
- The difference between two groups *at a population level*

	$Y^{X=1}=1$	$Y^{X=1}=0$
$Y^{X=0}=1$	A	B
$Y^{X=0}=0$	C	D

	$Y^{X=1}=1$	$Y^{X=1}=0$
$Y^{X=0}=1$	A	B
$Y^{X=0}=0$	C	D

Average Causal Risk Difference

- A+C/N – A+B/N
- $P(Y^{X=1}=1)-P(Y^{X=0}=1)$
- $P(Y^1)-P(Y^0)$
- The average causal risk ratio is similar
- Right now, think about this conceptually – worry about how we calculate this later

Target Population

- All of this is, ideally, focused on a population we are interested in, the “target population”
- *This is not necessarily the same as the study population*
- Consider a randomized trial – we are interested in all people who might take a drug, but the trial participants are likely a subset of those

Causal Identification Criteria

- These are things you need to have for causal effects to be capable of being estimated
- Temporality
- Consistency
- Exchangeability
- Positivity

Temporality

- Causes must precede effects in time
- Easiest to solve by making sure no one has the outcome at the time the study begins

Consistency

- Among people who were exposed, their outcome is the same as if they had been assigned that exposure and vice versa for unexposed
- "Treatment variation irrelevance"
  - Different doses of a drug – probably violate consistency
  - Different timing – maybe
  - Different shape of pills – probably not
- Dependent happenings (often called interference in the causal inference world) do a number on this

Exchangeability

- Risks are equal outside the effect of exposure
- This is where RCTs draw a lot of their claims about being the only means of estimating causal effects (they aren't)
- Conditional exchangeability: Risks are equal outside the effect of the exposure *conditional on other variables*
  - This is how you can attempt to estimate causal effects in observational studies

Positivity

- All subjects must have a non-zero probability of being exposed or unexposed
  - There are no always or never exposed study subjects
  - Note this must be true across all values of an exposure or outcome for non-binary versions
  - Note that does not mean there has to *be* anyone in that category
  - For conditional exchangability *this includes the conditional variables*
- Can anyone think of examples?

Other Conceptions of Causality

- Koch's Postulates
  - Especially important in infectious diseases
  - The microorganism must be found in abundance in all organisms suffering from the disease but should not be found in healthy organisms.
  - The microorganism must be isolated from a diseased organism and grown in pure culture
  - The cultured microorganism should cause disease when introduced into a healthy organism.
  - The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.
- Can anyone see a problem with these?

Necessary and Sufficient Causes

- Necessary Causes:
  - If X is a necessary cause of Y, the presence of Y implies the occurrence of X. X does not however imply that Y will occur.
- Sufficient Causes:
  - If X is a sufficient cause of Y, then the presence of X will cause Y. Y however does not imply X occurred.

Causal Pies and Causal Components

- Introduced by Rothman
- Useful for understanding multi-cause outcomes
- Probably the only good use for pie charts
- Death from Stroke
- If either thing hadn't happened, the outcome would not have occurred
- Mostly used as a conceptual framework

Bradford Hill's Causal Criteria

- Lecture given in 1965 to the Royal Society of Medicine
- Nine Criteria
- "We can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis, and none can be required as a *sine qua non*."
  - "Essential condition"

Hill's Criteria

- Strength
- Consistency
- Specificity
- Temporality
- Biological Gradient
- Plausibility
- Coherence
- Experiment
- Analogy

Strength

- The larger an association, the more likely an effect is causal
  - Infectious diseases gets lots of big, double-and-triple digit relative risks because of clean causation

Consistency

- Consistent findings over different samples in different places strengthens the case that an effect is causal
- Aka Reproducibility

Specificity

- Causation is likely if an exposure leads to a specific outcome, not multiple outcomes
- You can get overly cheeky with this – “Arsenic specifically causes arsenic poisoning”

Temporality

- The cause occurs before the effect

Biological Gradient

- There is a dose-response relationship for continuous exposures – greater exposure should lead to greater effect (or at least, changed effect)

Plausibility

- A plausible mechanism between cause and effect is helpful
- This is limited, but is often useful for ruling out causal arguments

Coherence

- Agreement between laboratory findings and epidemiological studies
- This criterion is often abused by certain industries demanding further and different sources of evidence

Experiment

- If you can show something in an experiment, you're likely on the right track
- Again, this is where RCTs draw a lot of their argument from

Analogy

- It is helpful if you can leverage evidence that a similar exposure causes a similar outcome

Questions?