### Categorizing Continuous Data

In addition to using larger spatial designations, researchers aggregate the physical exposures themselves, simplifying continuous measurements down to a single exposure metric. While aggregate values often represent the mean of all the values recorded, weather data is typically assessed by the maximum value. Regardless, aggregating physical exposure data requires researchers to categorize continuous data, which involves choosing appropriate thresholds.

Thresholds are often used to assign exposure status to individuals or populations (often using a county as proxy for a population). For example, a county may be classified as exposed or unexposed based on local winds exceeding a threshold (e.g. gale-force winds or higher). Using the maximum wind speed at the center of a county is often the measurement used to determine exposure status. [@grabich2016measuring] classified hurricane exposure in a Florida county using maximum wind speed. Maximum wind speed is a continuous variable, but the study used binary categorizations to divide it into tropical wind speeds, classified as greater than 39 miles per hour, and hurricane wind speeds, classified as greater than 74 miles per hour. Florida counties experiencing maximum wind speeds below 39 miles per hour were considered unexposed. In this example it is noteworthy to examine that all the Florida counties in this paper likely experienced hurricane winds somewhere on this spectrum, but categorizing that continuous data made exposure much simpler and concrete.

The Saffir-Simpson scale is an example of how entire storms are often classified by their maximum wind speed. Forecasters classify hurricanes into categories on the Saffir-Simpson scale based on maximum sustained surface wind speed. This is defined as the peak one minute wind speed at a height of 10 feet over an unobstructed exposure [@taylor2010saffir]. The Saffir-Simpson scale uses five different bins to classify varying levels of wind speed and determine the severity of a storm. The first level, Category 1 is designated for hurricanes and tropical storms with maximum wind speeds of between 64 - 82 knots and is generally considered dangerous to people, livestock, and pets from the hazard of flying and falling debris [@taylor2010saffir]. On the higher end of the scale, Category 5 designates hurricanes with maximum wind speeds above 137 knots and is considered to have catastrophic effect on damage and a high probability of injury or death to people, livestock, and pets even if they are sheltering indoors [@taylor2010saffir]. An important limitation of the Saffir-Simpson scale is that it doesn't account for other hurricane-related impact variables such as storm surges, flooding, and tornadoes [@taylor2010saffir]. [@shao2017understanding] used this scale to assign wind speed categories to counties along the Gulf Coast in a study assessing perceptions of risk to tropical cyclones. [@belasen2008hurricanes] used this scale as well in a study that compared hurricane intensities to average earnings in different counties. Hurricanes with categories one, two or three were considered lower intensity, and hurricanes of counties four and five were considered high intensity.

Another scale used to categorize wind speed is the Beaufort scale, created by Admiral Sir Francis Beaufort, used to classify wind speeds both over land and sea. While the Saffir-Simpson scale is only designated for wind speeds that are already at hurricane levels (greater than 64 knots), the Beaufort scale considers the wind speeds below this. The scale ranges from Force 0 (0-1 knots and calm) to Force 12 (64 to 71 knots and hurricane). Other interesting parts of the scale include Force 3 (4-6 knots) which is a gentle breeze, and Force 8 (34-40 knots) which is considered a gale.

There are times when thresholds of human impacts are also used to assign exposure. In [@christopher2017effects], pregnancy outcomes were studied in relation to exposure to tropical cyclones and tornadoes. Birth outcomes to, others who had been pregnant during the cyclone disasters in counties exposed to the storm were compared to birth outcomes of mothers in unexposed counties. In this study, exposure was analyzed by using thresholds of fatalities (greater than ten deaths), and property damage (greater than $10 billion).

Despite several advantages to dichotomizing continuous variables that we just discussed, there are several limitations to consider. Statistical power is lost because so much information is lost when categorization occurs [@van2008leave]. This makes sense when you consider that continuous variables allow you to observe nuance in the data and perceive a dose response relationship between the predictor and response variables, should one exist. This effect is masked when researchers categorize data, and even more so when a smaller number of categorical variables are used (for example dichotomization itself at 2). Generally, if you are going to categorize continuous data, it is better to use 3 or more categories rather than just two, because this will capture more variation in the data that would otherwise be lost. An example of a paper that used three different bins was [@kinney2008autism], which explored the risk of autism after a pregnancy that included exposure to a tropical storm in the state of Louisiana. The study authors classified tropical storm exposure as severe, intermediate, and low exposure, and these exposure classifications were determined based on whether a mother lived in a Louisiana parish that had both of the exposure factors of interest: storm intensity and storm vulnerability. Storm vulnerability in this case was based on another dichotomy: whether or not the storm center passed through the parish of interest. Storm vulnerability was a measure of how vulnerable the inhabitants of the parish were to the effects of a storm (higher socioeconomic neighborhoods and parishes have more resources to withstand and recover from a tropical storm for example).

Another obvious problem with categorizing continuous data is that the cutoff points are often arbitrary. In the case of dichotomization, the median is often used, but there is typically no reason to assume that the median is a reasonable cutoff point. Because different samples will have different medians, this automatically makes many categorical bins difficult to compare across studies [@altman2006cost]. Further, choosing optimal cutoff points that give the smallest p-values can lead to spurious results [@altman2006cost].

Not surprisingly, dichotomizing continuous variables can bias results. A study by Selvin showed that the odds ratios can be significantly different depending on the chosen cutoff that is implemented in a study [@van2008leave]. Categorical variables can also put otherwise similar observations into separate bins if they are close but on opposite sides of the cutoff [@altman2006cost]. Choosing a median as a cutoff is intended to delineate bins, but if the bins are a "high" and "low" group, two individual observations that may only be a fraction different, but on either sides of the mean, will be classified as high and low respectively, and give the false impression that they are significantly different.

Although using a single exposure value can simplify analysis and interpretation, particularly over an extended temporal scale, there are some obvious drawbacks to relying on one single aggregate value. For example, the Saffir Simpson categories typically correspond only to the geographic point location where the maximum wind speed was observed [@taylor2010saffir]. Hurricane Wilma in 2005 for example, was a Category 3 hurricane when it made landfall on the southwest coast of Florida, but it created Category 1 and Category 2 conditions for the more populous Miami-Dade, Broward, and Palm Beach counties when it finally reached them [@taylor2010saffir].

Reverse Funnel

(Small)

- We've discussed the datasources, temporal and spatial scales, and methods of collecting data on human impacts and physical exposures related to tropical cyclones.

- We've talked about how different temporal and spatial scales make it challenging to immediately use this data.

- We've discussed the main methods for integrating data at different spatial/temporal scales: aggregating, interpolating, matching.

- We've discussed the implications these integration methods have, and what effect they may have on the observed associations with regard to error and bias.

(Mid-Funnel)

- Error and bias are bad because they make the results of tropical cyclone studies less generalizable.

- Less generalizable studies make it harder to predict human impacts in the future and therefore prepare for them.

- If we don't know what to expect, some communities will be very vulnerable when the next big storm hits. (Thinking about Daniel's paper. He mentioned that the short temporal record of storm tracks meant that some communities weren't thought of as having a high risk because they weren't in the historical record).

(Wide Funnel)

- As modeling becomes more advanced from physical exposure data, we need to make sure that human impacts data is collected and made available in a way that matches the physical exposure quality.