4. Developing a population version of the linear model

- We now know how to set up a linear model explaining a response variable $\mathbf{y}$ using a matrix of explanatory variables $\mathbb{X}$. We write $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ and use least squares to estimate the coefficient vector, $\mathbf{b}$. We understand that this is a compact way of writing $y_i = x_{i1}b_1 + x_{i2}b_2 + \cdots + x_{ip}b_p + e_i$ for $i = 1, \ldots, n$.
- Generically, we call y_i the response for **individual** i. We think of an individual as a row in the dataset. In some situations, this terminology is counter-intuitive, for example in HW3 when we consider a dataset where there are four rows of data for each kicker.
- A positive value of b_j for j in $\{1, \ldots, p\}$ means that larger values of the jth predictor variable are associated with larger values of the response.
- Suppose the individuals are randomly drawn from some sample. Common statistical questions are:

  (a) How much would our estimate of b_j change if we had a different sample?

  (b) Is there convincing evidence that $b_j \neq 0$? If not, we could remove the jth predictor from the model.

Population inference needs a probability model

- A probability model for a data vector $\mathbf{y}$ uses **random variables** to model how the data were generated.
- Probability models let us assign probabilities to events that may or may not occur, such as "What is the probability that the difference between the estimated coefficient b_1 and the true coefficient is smaller than 0.1."
- We have the following goals:
  - Review the rules of probability.
  - Build the skills needed to work with probabilities for linear models.
  - Learn to check our probability calculations using R.
  - Use probability calculations to develop statistical inference procedures for linear models.

Review of probability

- The **probability** of an event is the long-run proportion of times that an event would happen in a large number of datasets drawn from the probability model.
- We will review the material on random variables from STATS 250 at `open.umich.edu/find/open-educational-resources/statistics`. See, in particular, *Interactive Lecture Notes 04: Probability*, *Interactive Lecture Notes 05: Random Variables*, and *Workbook 03: Lab 2 - Probability and Random Variables*.

Potential outcomes, events, and their probability

- A **potential outcome** of a probability model is any dataset that could be generated by the model.
- **Example**. The set of potential outcomes when rolling a 6-sided die can be modeled as $\{1, 2, 3, 4, 5, 6\}$.
- An **event** is a collection of potential outcomes.
- An event A can either happen or not happen on any **realization** of the model.
- $\mathcal{P}(A)$ is the **probability** that A happens according to the model.
- If each outcome is equally likely (e.g., a roll of a fair die) we can generate realizations of the random variable in R using `sample()`

```
## Make 10 draws with replacement from {1,2,3,4,5,6}
## This models 10 realizations of rolling a fair die
sample(1:6,size=10,replace=TRUE)


##  [1] 4 2 2 5 5 3 6 6 6 6
```