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 find a 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 a random sample from some population. Common statistical questions are:
  (a) How much might b_j change if we had a different sample?
  (b) Is the least squares estimate of b_j small enough that it is reasonable to use an estimate $b_j = 0$? If so, we can remove this predictor and simplify 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 sample least squares coefficient b_1 and the true population coefficient is smaller than 0.1."
- We have the following goals:
  - Review the rules of probability and random variables.
  - Build the skills needed to work with probabilities for linear models.
  - Learn to use R to make probability calculations.
  - Use probability calculations to develop statistical inference procedures for linear models.

Possible outcomes and events

- A **possible outcome** of a probability model is any dataset that could be generated by the model.
- The set of all possible outcomes for model is called the **sample space** of that model.
- **Example**. The set of possible outcomes when rolling a 6-sided die can be modeled as $\{1, 2, 3, 4, 5, 6\}$. This excludes the die rolling off the table, or balancing on its edge.
- An **event** is a collection of possible outcomes.
- Formally, an event is therefore a subset of the sample space.
- An event can happen or not happen on any **realization** of the model.
- If each outcome is equally likely (e.g., a roll of a fair die) we can generate realizations of the model 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
```

A definition of probability

- The **probability** of an event is the long-run proportion of times that an event happens in a large number of realizations of the probability model.
- Probabilities are only defined in the context of a probability model. If we talk about the probability that a particular US president will be reelected, that means we have a model for it. We can draw many realizations from our model, even though we are modeling one specific election.
- For an event A, we write the probability of A in our model as $\mathcal{P}(A)$.
- We can write A in words or as a set of outcomes. Saying "A is the event that a die roll is even" is equivalent to saying $A = \{2, 4, 6\}$.
- 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*
    - *Workbook 03: Lab 2 - Probability and Random Variables*

A definition of random variables

- A **random variable** is a number associated with each possible outcome in the sample space of a probability model.
- **Example**. A probability model for 3 consecutive coin tosses has possible outcomes $\{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$. The random variable counting the number of heads takes value 2 when the outcome is HHT, HTH or THH.
- Events can be defined using random variables. Let X be the number of heads in these three coin tosses. Let A be the event that there are two heads. We can write this as $A = \{X = 2\}$ or $A = \{HHT, HTH, THH\}$.
- To talk about the probability of A, we could write $\mathcal{P}(A)$ or $\mathcal{P}(HHT, HTH, \text{ or } THH)$ or $\mathcal{P}(X = 2)$.
- Random variables, like probabilities, are only defined in the context of a model. For example, suppose that for the first three trials of a new surgical operation, two of the patients have a successful operation. It may be useful to model surgical success like the outcome of coin toss. However, data are not random variables, and random variables are not data!