4. Probability models: mean, variance and the normal distribution

- A **probability model** is an assignment of **probabilities** to **events** in the **sample space** of all **possible outcomes** for the situation being modeled.
- Probabilities are between 0 (impossible) and 1 (certain).
- The total probability of all possible outcomes is 1 (something is certain to happen!).
- Probability models are the only known way to quantify uncertainty.
- Drawing statistical conclusions from data inevitably involves uncertainty.

Probability models are important for Statistics!

Why do we need a probability model for 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 $i = 1, \ldots, n$.
- 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** y_j.
- However, there is always room for uncertainty. Maybe we sampled only a small fraction of the population of interest. There could be error in the measurements. For experimental data, the responses would be different each time we collected a dataset. Common statistical questions are:
  (a) How much might b_j change if we repeated the experiment?
  (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.
- Probability models will let us answer these questions.

Possible outcomes and events

- 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.
- 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 $A = \{2, 4, 6\}$ where the set of possible outcomes is $\{1, 2, 3, 4, 5, 6\}$.
- An event can happen or not happen on any **realization** of the model.

Random variables

- A probability model when the outcome is numeric is called a **random variable**.
- Statistics is primarily concerned with numeric quantities, so
- For the linear model, the response takes a numeric value. Therefore, a probability model for a data vector $\mathbf{y}$ uses random variables to model how the data were generated.
- Outcomes that are not numeric can be made numeric.
- Outcomes of *heads* and *tails* for a coin can be made numeric by assigning value 1 for *heads* and 0 for *tails*.
- Outcomes of "What is your favorite movie?" can be made numeric by enumerating all movies.

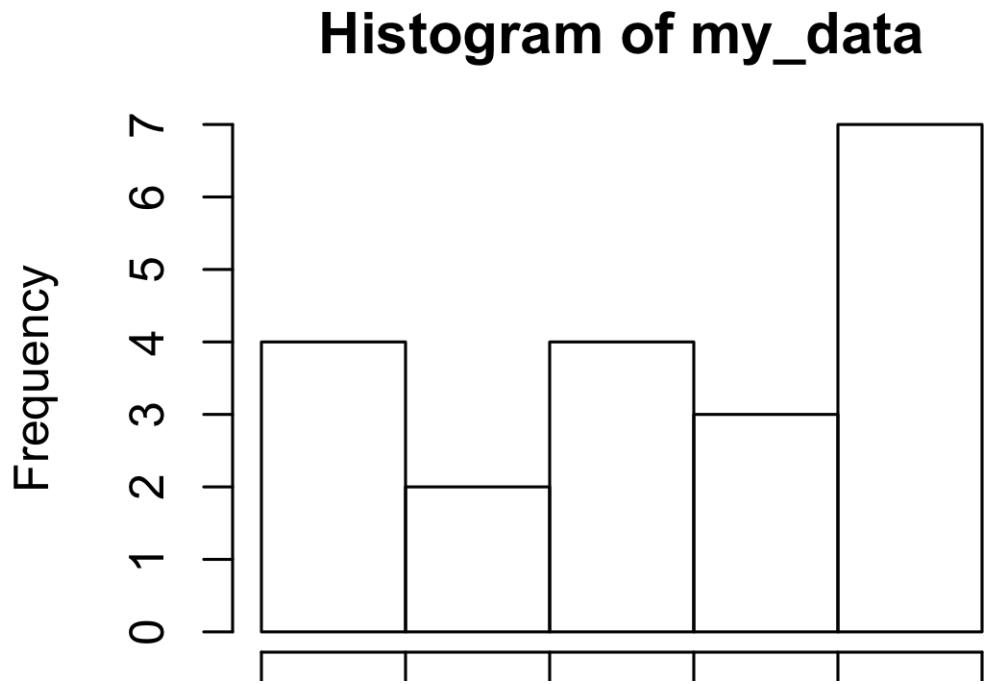
Our goals for learning about probability models

- 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.

Generating random variables with a computer, in R

- 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 20 draws with replacement from {1,2,3,4,5,6}
## This models 20 realizations of rolling a fair die
## We can plot a histogram of the simulated dice rolls
my_data <- sample(1:6,size=20,replace=TRUE)
hist(my_data)
```



A definition of probability for repeatable experiments

- 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 senator will be reelected in November, 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 $\mathrm{P}(A)$.
- We will review the material on probability 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*

Bivariate and vector-valued random variables

- Our data usually have more than one number, so a probability model for the data needs to generate a collection of random variables.
- A probability model generating n numeric values is a **vector-valued random variable** of length n.
- In the special case $n = 2$, the pair of numeric values is a **bivariate random variable**.
- We have already seen one example of a vector-valued random variable. Above, we made 20 dice rolls to generate a vector of 20 outcomes.

Notation for random variables

- We use upper case letters for random variables, and lower case letters for the possible values of the random variables. $\{X = x\}$ is common notation for the event that the random variable X takes the specific value x.
- Bivariate random variables are often called X and Y.
- Vector-valued random variables are often called $X_1, \ldots, X_n$.
- For linear models, the data are often called $y_1, \ldots, y_n$ so we call the probability model $Y_1, \ldots, Y_n$. Here, $\mathbb{X}$ is reserved for the matrix of covariates. In the usual linear model, entries of $\mathbb{X}$ are fixed, not random.

Events corresponding to the outcomes of random variables

- Events can be specified as a range of outcomes of random variables.
- Let X be the outcome of rolling a fair die. Let A be the event that the die lands on 5 or 6. We can write A as $\{X \geq 5\}$.
- To talk about the probability of A, we could write $\mathrm{P}(A)$ or $\mathrm{P}(X \geq 5)$.
- For vector-valued random variables, an event may involve many or all the random numbers.
- Let X_1, X_2, X_3 be three rolls of a die. Let $B = \{X_1 + X_2 + X_3 \geq 12\}$. B is the event that the sum of three dice is at least 12.

Question 4.1. Define a probability model for fifty dice rolls and use summation notation to write the event that the sum of fifty dice exceeds thirty.