THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Simulation 1

Lecture 22

STA 371G

What is simulation?

- When we have data, we can use regression models to predict an outcome based on input data that we know we will have available.

What is simulation?

- When we have data, we can use regression models to predict an outcome based on input data that we know we will have available.
- However, sometimes we don't have the data we need to build a model, or we might want to understand the full range of possible outcomes rather than predicting a single outcome.

Examples of simulation

- What will the value of my retirement portfolio look like in 30 years, and how does that depend on market conditions over that time span?
- What profit should I expect to see from a business selling team hats during the NBA playoffs, and how does that vary depending on who wins the playoffs and what kinds of hats I order?
- What sorts of returns are possible if I decide to drill for oil in a newly-discovered field, and how does that depend on my assumptions about drilling costs and the likelihood of success?

Simulation framework

1. Define an quantity of interest that represents the outcome.
2. Articulate a set of assumptions about the problem.
3. Simulate the random processes that are inherent in determining the outcome.
4. Compute the quantity of interest.
5. Repeat steps 3-4 a large number of times, and examine the long-term distribution of the quantity.

Simulation framework

1. Define an quantity of interest that represents the outcome.
2. Articulate a set of assumptions about the problem.
3. Simulate the random processes that are inherent in determining the outcome.
4. Compute the quantity of interest.
5. Repeat steps 3-4 a large number of times, and examine the long-term distribution of the quantity.

We call this whole process the simulation, and steps 3-4 a run of the simulation.

Example 1: Coin flipping

Suppose we flip a coin 10 times. What will the distribution of the number of heads look like?

Example 1: Coin flipping

Suppose we flip a coin 10 times. What will the distribution of the number of heads look like?

- Let's define the random variable H to be the number of heads.

Example 1: Coin flipping

Suppose we flip a coin 10 times. What will the distribution of the number of heads look like?

- Let's define the random variable H to be the number of heads.
- We assume that the probability of heads or tails is equal for each flip, and that each flip is independent.

Random sampling in R

The `sample` function in R lets us randomly select from among several alternatives with equal probability. (We used this last week when we tested for ESP!)

To sample from the set $\{0, 1\}$ 10 times, with replacement:

```
sample(c(0, 1), 10, replace=T)

 [1] 0 0 0 1 1 0 0 0 1 1
```

Replicating a run over and over

Every time we execute this run, we'll get a different answer. What we want to know is how often we'll get each answer—e.g., how often do we get 5 versus 6 versus 7?

Replicating a run over and over

Every time we execute this run, we'll get a different answer. What we want to know is how often we'll get each answer—e.g., how often do we get 5 versus 6 versus 7?

The `replicate` command lets us run a block of code repeatedly. Inside the code block, we run the `return` function to tell R what the quantity of interest is. For example, let's run our coin flipping run 15 times:

```
replicate(15, {
  flips <- sample(c(0, 1), 10, replace=T)
  return(sum(flips))
})

 [1] 6 3 5 6 4 5 4 3 6 8 2 4 6 8 6
```

Replicating a run over and over

Why limit ourselves to 15 times? How about 100 times?

Replicating a run over and over

Why limit ourselves to 15 times? How about 100 times?

```
replicate(100, {
  flips <- sample(c(0, 1), 10, replace=T)
  return(sum(flips))
})
```

```
  [1] 3 1 6 4 7 6 6 8 6 7 5 5 5 6 7 1 6 5 6 5 5 4 3 2
 [25] 6 6 8 7 6 2 6 6 8 3 6 7 6 5 8 1 5 3 4 3 8 6 7 5
 [49] 3 5 4 6 3 4 5 5 6 5 3 5 6 5 6 4 5 7 4 3 5 4 8 6
 [73] 5 5 2 6 4 6 6 6 4 4 6 6 4 8 4 7 7 4 5 8 6 2 2 4
 [97] 7 5 7 5
```

Examining the results of a simulation

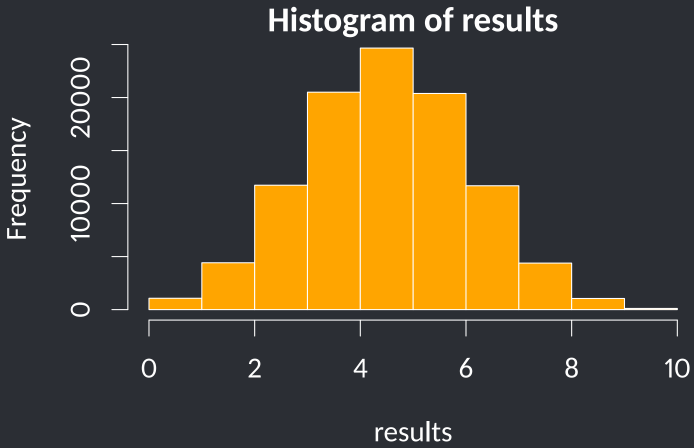
Since computers are so fast, we'll often do a very large number of runs (e.g., 100,000 times) to even out the idiosyncrasies of any individual run due to randomness.

Examining the results of a simulation

Since computers are so fast, we'll often do a very large number of runs (e.g., 100,000 times) to even out the idiosyncrasies of any individual run due to randomness.

But then we'll have a list of 100,000 numbers representing the result of 100,000 different runs, so we need a way of summarizing those results. To do this, we can look at a histogram of the results, or examine a summary statistic.

```r
results <- replicate(100000, {
  flips <- sample(c(0, 1), 10, replace=T)
  sum(flips)
})
hist(results, breaks=10, col='orange')
```



Theory tells us that H, the outcome of flipping a fair coin 10 times and counting heads should follow a Binomial distribution with

$$E(H) = np = 10 \cdot 0.5 = 5$$

and

$$\mathrm{SD}(H) = \sqrt{np(1-p)} = \sqrt{10 \cdot 0.5 \cdot 0.5} \approx 1.58.$$

Theory tells us that H, the outcome of flipping a fair coin 10 times and counting heads should follow a Binomial distribution with

$$E(H) = np = 10 \cdot 0.5 = 5$$

and

$$\mathrm{SD}(H) = \sqrt{np(1-p)} = \sqrt{10 \cdot 0.5 \cdot 0.5} \approx 1.58.$$

Let's compare our simulated results against the theoretical results:

```
mean(results)

[1] 5.00079

sd(results)

[1] 1.585098
```

Simulating random variables

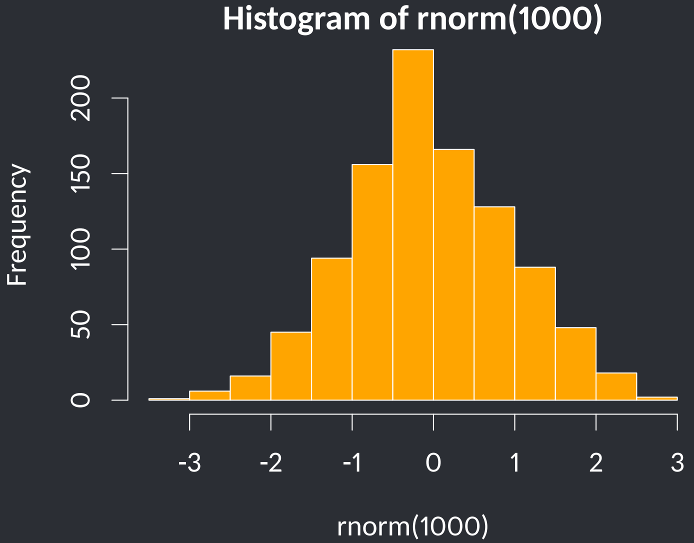
We have seen how the `sample` command can be used to draw from a set of alternatives with equal probability (e.g., flipping a coin). The `rnorm` command can be used to draw randomly from a normal distribution. Let's create 10 random heights, with mean 68 (inches) and SD 4.

```
rnorm(10, 68, 4)
```

```
 [1] 61.75752 74.42290 67.61342 66.50313 69.74625
 [6] 67.66187 70.59517 70.58163 65.89903 66.39669
```

Let's check that `rnorm` works as advertised!

```
hist(rnorm(1000), col="orange")
```



Example 2: Will I get an A?

- Let's say I'm taking a class with two midterms (each 25% weight in final grade) and a final exam (50% weight in final grade).

Example 2: Will I get an A?

- Let's say I'm taking a class with two midterms (each 25% weight in final grade) and a final exam (50% weight in final grade).
- I just got my score on the first midterm (75%).

Example 2: Will I get an A?

- Let's say I'm taking a class with two midterms (each 25% weight in final grade) and a final exam (50% weight in final grade).
- I just got my score on the first midterm (75%).
- I want to know how likely it is that I can get 90% or above on my final grade.

Example 2: Will I get an A?

- Let's say I'm taking a class with two midterms (each 25% weight in final grade) and a final exam (50% weight in final grade).
- I just got my score on the first midterm (75%).
- I want to know how likely it is that I can get 90% or above on my final grade.
- This is hard!

Example 2: Will I get an A?

Let’s start by making some assumptions:

- I think I can improve on the second midterm, and then even more on the final.

Example 2: Will I get an A?

Let's start by making some assumptions:

- I think I can improve on the second midterm, and then even more on the final.
- I'll model my Midterm 2 grade as a normal distribution.

Example 2: Will I get an A?

Let's start by making some assumptions:

- I think I can improve on the second midterm, and then even more on the final.
- I'll model my Midterm 2 grade as a normal distribution.
- My best guess is that I'll get an 80% on Midterm 2, and I'm 95% sure it will be between 70% and 90%.

Example 2: Will I get an A?

Let's start by making some assumptions:

- I think I can improve on the second midterm, and then even more on the final.
- I'll model my Midterm 2 grade as a normal distribution.
- My best guess is that I'll get an 80% on Midterm 2, and I'm 95% sure it will be between 70% and 90%.
- So my Midterm 2 grade should be simulated as a normal distribution with mean 80 and SD 5 (since 95% of a normal distribution is roughly ± 2 SD from the mean).

Example 2: Will I get an A?

Let's start by making some assumptions:

- I think I can improve on the second midterm, and then even more on the final.
- I'll model my Midterm 2 grade as a normal distribution.
- My best guess is that I'll get an 80% on Midterm 2, and I'm 95% sure it will be between 70% and 90%.
- So my Midterm 2 grade should be simulated as a normal distribution with mean 80 and SD 5 (since 95% of a normal distribution is roughly ± 2 SD from the mean).
- I think I can improve more on the final; my best guess is that I'll get a 90%, and I'm 95% sure I'll get between 80% and 100%.

Example 2: Will I get an A?

- For each run, we will:

Example 2: Will I get an A?

- For each run, we will:
    1. Randomly draw a Midterm 2 score from its normal distribution, and a Final Exam score from its normal distribution.

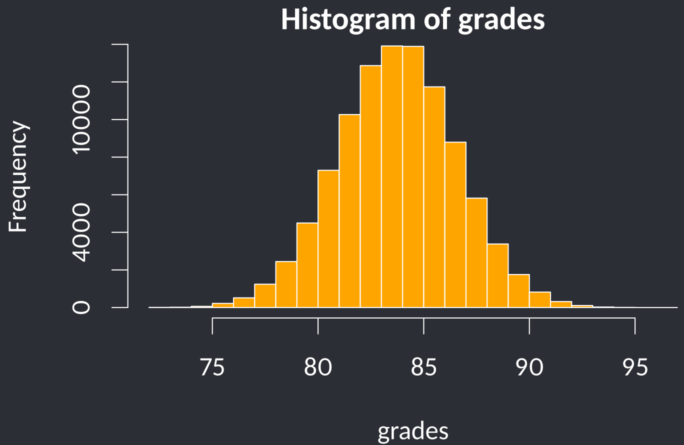
Example 2: Will I get an A?

- For each run, we will:
    1. Randomly draw a Midterm 2 score from its normal distribution, and a Final Exam score from its normal distribution.
    2. Calculate a final score for the course, and see if it's over 90%.

Example 2: Will I get an A?

- For each run, we will:
  1. Randomly draw a Midterm 2 score from its normal distribution, and a Final Exam score from its normal distribution.
  2. Calculate a final score for the course, and see if it's over 90%.
- Then we will count the percentage of runs where we got 90%+ for the course. That will be our estimate of the probability of getting an A.

```
grades <- replicate(100000, {
  midterm1 <- 75
  midterm2 <- rnorm(1, mean=80, sd=5)
  final.exam <- rnorm(1, mean=90, sd=5)
  return(.25*midterm1 + 0.25*midterm2 + 0.5*final.exam)
})
hist(grades, col="orange")
```



```r
runs <- replicate(100000, {
  midterm1 <- 75
  midterm2 <- rnorm(1, mean=80, sd=5)
  final.exam <- rnorm(1, mean=90, sd=5)
  return(0.25*midterm1 + 0.25*midterm2 + 0.5*final.exam >= 90)
})
sum(runs) / 100000

[1] 0.01286
```

```r
runs <- replicate(100000, {
  midterm1 <- 75
  midterm2 <- rnorm(1, mean=80, sd=5)
  final.exam <- rnorm(1, mean=90, sd=5)
  return(0.25*midterm1 + 0.25*midterm2 + 0.5*final.exam >= 90)
})
sum(runs) / 100000

[1] 0.01286
```

There's only about a 1.29% chance that I'll get an A.

What if our assumptions are wrong?

- The problem with simulations is that the results can be very sensitive to our assumptions.

What if our assumptions are wrong?

- The problem with simulations is that the results can be very sensitive to our assumptions.
- To the extent that our assumptions are incorrect, we can't trust the results of the simulation. ("Garbage in, garbage out")

What if our assumptions are wrong?

- The problem with simulations is that the results can be very sensitive to our assumptions.
- To the extent that our assumptions are incorrect, we can't trust the results of the simulation. ("Garbage in, garbage out")
- But the great thing about simulations is that it's easy to change the assumptions and see what effect it has on the results.

What if our assumptions are wrong?

- The problem with simulations is that the results can be very sensitive to our assumptions.
- To the extent that our assumptions are incorrect, we can't trust the results of the simulation. ("Garbage in, garbage out")
- But the great thing about simulations is that it's easy to change the assumptions and see what effect it has on the results.
- For example, what if Midterm 2 is drawn from a normal distribution with mean 90, instead of mean 80?

```r
runs <- replicate(100000, {
  midterm1 <- 75
  midterm2 <- rnorm(1, mean=90, sd=5)
  final.exam <- rnorm(1, mean=90, sd=5)
  return(0.25*midterm1 + 0.25*midterm2 + 0.5*final.exam >= 90)
})
sum(runs) / 100000

[1] 0.0881
```

```r
runs <- replicate(100000, {
  midterm1 <- 75
  midterm2 <- rnorm(1, mean=90, sd=5)
  final.exam <- rnorm(1, mean=90, sd=5)
  return(0.25*midterm1 + 0.25*midterm2 + 0.5*final.exam >= 90)
})
sum(runs) / 100000

[1] 0.0881
```

Under this new assumption, there's an 8.81% chance that I'll get an A.