Random Processes, Probability, and Sampling

Ryan Miller

Grinnell College
Statistics

Introduction

- So far, we've focused on methods and models to describe the trends present in our data
  - We've ignored the role of uncertainty (random chance) in our analyses
- For example, might Joseph Lister have observed a different odds ratio if his random assignment of patients to "sterile" and "conventional" groups unfolded differently?

Random Processes

- A **random process** produces an outcome that is *probabilistic* rather than *deterministic*
  - Rolling a 6-sided die is a random process, as we don't know the outcome in advance, and different rolls of the same die can produce different results
  - Converting a temperature from Celsius to Fahrenheit is a deterministic process, as we'll get the same result every time when using the same temperature

Sample Spaces

A **sample space** refers to the collection of *all* possible outcomes of a random process:

Random Process	Sample Space
Flipping a Coin	{Heads, Tails}
Rolling a 6-sided Die	{1,2,3,4,5,6}
Patient Undergoes Surgery	{Survives, Dies}
Running a 5km race	{Positive Real Numbers}

Practice

Explain whether each of the following is a *random process* or a *deterministic process* and describe the process's sample space.

1. You take out a fixed rate mortgage on a home and assess how much interest you'll pay over the life of the loan.
2. You take an exam in your statistics class and receive a letter grade on it.

Random Variables

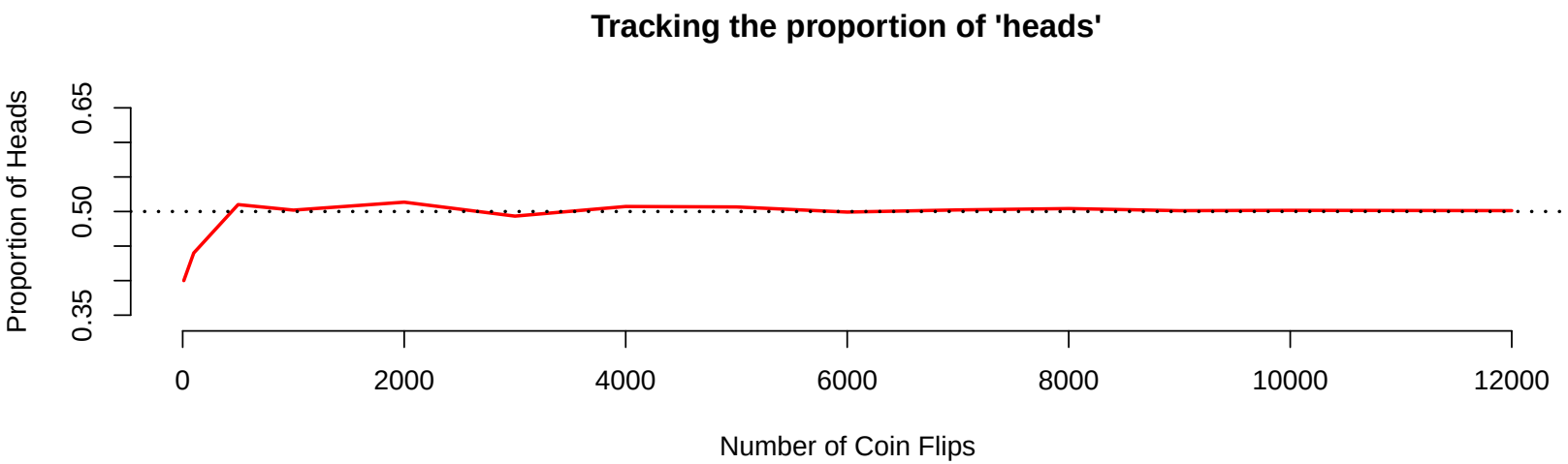
A **random variable** is a variable (ie: an entity with an unknown value) that depends upon a random process:

1. Rolling a 6-sided die is a random process, we might use the random variable X to denote the number facing up
2. Randomly assigning 75 patients to receive either sterile surgery or conventional surgery is a random process, we could use a random variable to denote the odds ratio

Note that the same random process can give rise to different random variables, as we could express the odds ratio, relative risk, or difference in proportions as random variables in Lister's experiment

Probability

- **Probability** is the *long-run relative frequency* of an outcome over increasingly many repetitions of a random process
  - This definition is based upon the **Law of Large Numbers**, which states that the proportion of times an outcome is observed converges to its probability as the number of repetitions is taken towards infinity:



Basic Probability Rules

I won't ask you to solve exercises using these rules, you should have a basic understanding of them.

1. **Addition Rule** - $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$, or the probability of one outcome *or* another outcome occurring is given by the sum of the individual probabilities minus any probability of both events occurring simultaneously
2. **Multiplication Rule** - $Pr(A \cap B) = Pr(A) * Pr(B)$, or if two events are independent (the outcome of one event doesn't impact the other), the probability of both events occurring is their product
3. **Complement Rule** - $Pr(\text{Not } A) = 1 - Pr(A)$, or the probability of an event *not occurring* is 1 minus the probability of it occurring

Probability Distributions

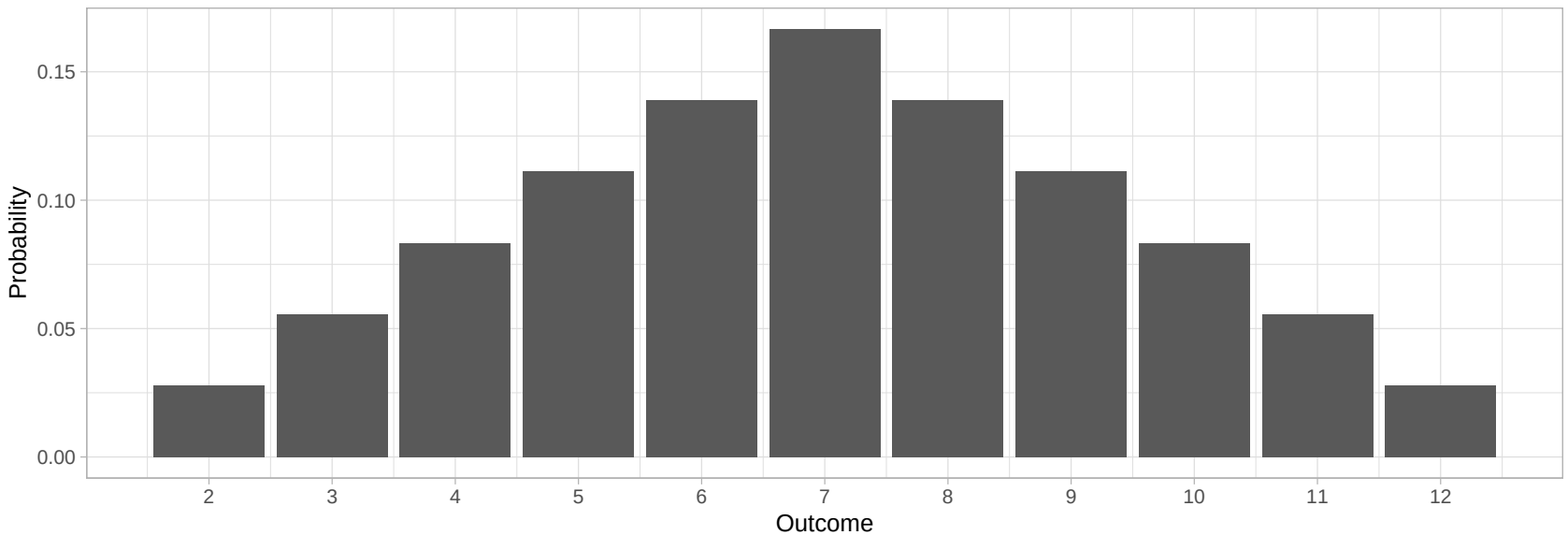
- A **Probability distribution** describes the probability of every possible value of a random variable
  - Probability distributions can be *discrete* or *continuous*, but they must include all outcomes and the involved probabilities must sum to 1
- For example, here's the probability distribution for the sum of two 6-sided dice rolls:

Event	2	3	4	5	6	7	8	9	10	11	12
Probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

If all 6 outcomes for a *single dice roll* each have a 1/6 probability, why is $\frac{2}{36}$ the probability of *two dice rolls* summing to 2?

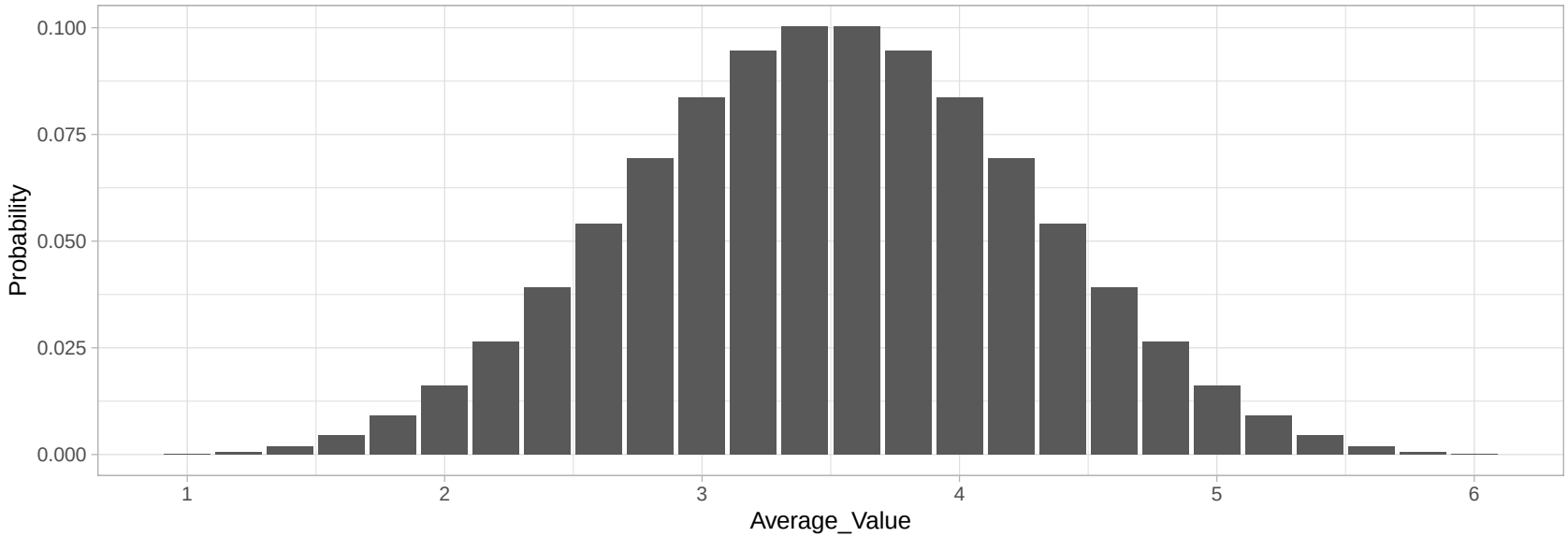
Probability Distributions

Below is a visualization of the probability distribution for the *sum* of two 6-sided dice rolls:



Probability Distributions

Below is the probability distribution for the *average* of five 6-sided dice rolls:

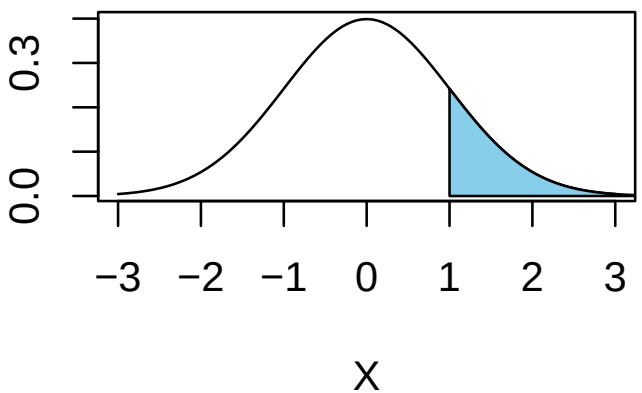


Notice that this looks a lot like a bell-curve, which is not a coincidence.

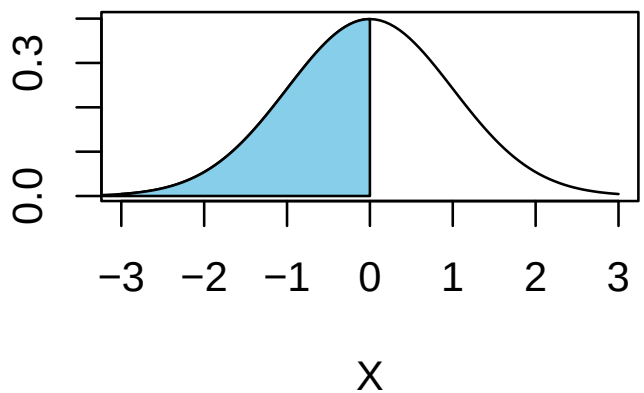
Probability Distributions

- Previous examples showed **discrete** probability distributions
  - These random variables each had a finite set of possible outcomes
- Continuous probability distributions are expressed using mathematical functions (ie: $f(X) = \ldots$)
  - These functions can be evaluated at any value of the random variable, but only produce a valid probability over an interval

Pr(X>1) = 0.16

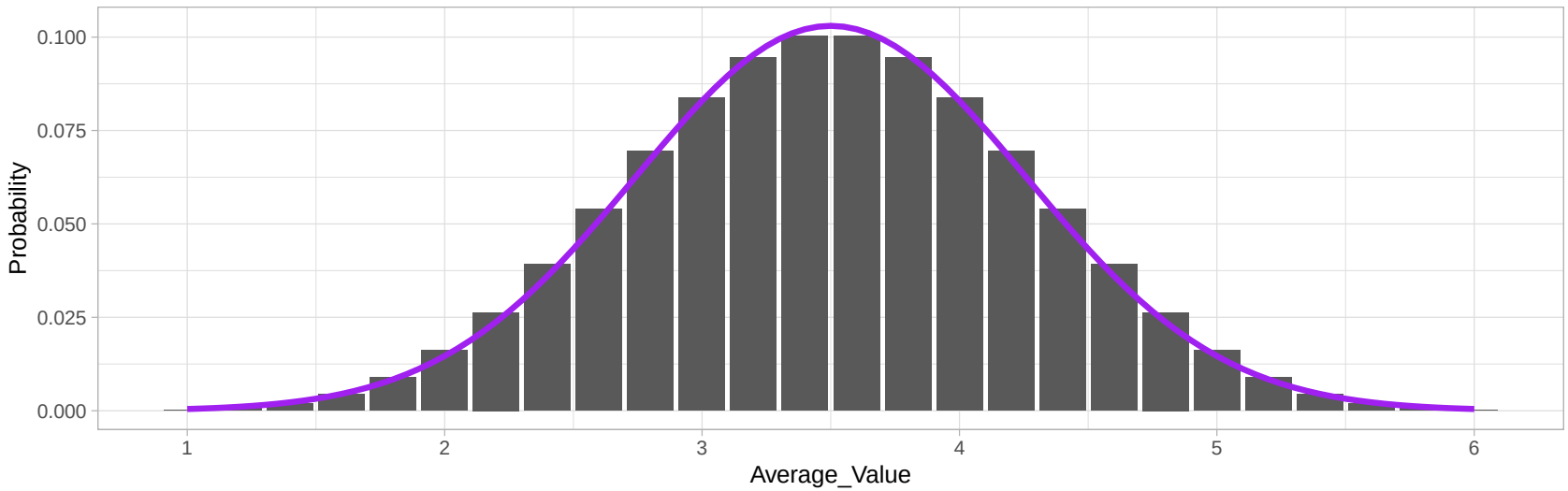


Pr(X<0) = 0.5



Approximation

Statisticians will often approximate one probability distribution (such as a discrete distribution) with another (such as a continuous distribution):



We'll soon see that this allows us to use a very similar type of *probability model* (the Normal distribution) for a wide variety of random variables.

Practice

Suppose we use a computer to randomly generate a real number between 0 and 10, letting the random variable, X, record the result.

1. Is X a discrete or continuous random variable?
2. What does the probability distribution of X look like? *Hint*: think about sketching it.
3. What is the probability that $X > 6$?

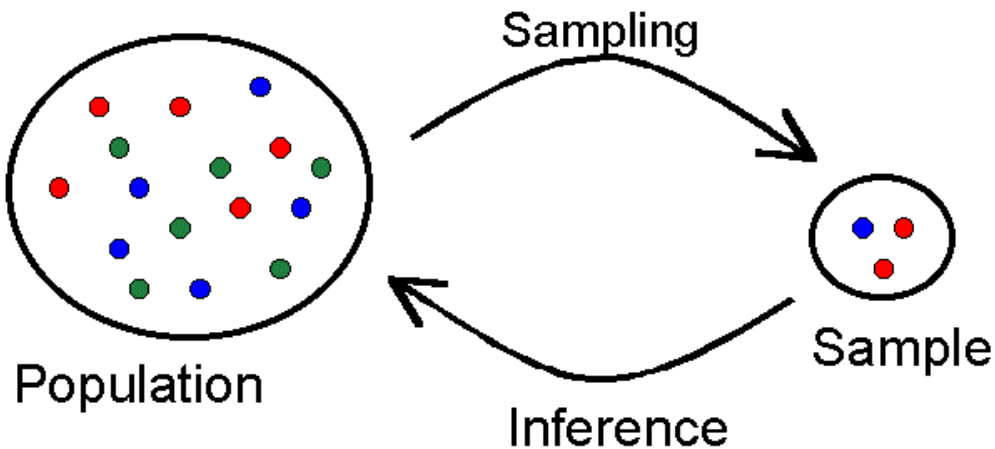
Sampling

Data collection is one of the most prominent sources of uncertainty. As an example, suppose a researcher is interested in understanding the relationship between the size of a fish and the concentration of mercury in its tissues for a particular species of fish in a certain lake.

1. Does the researcher need to collect data on *every* fish of this species residing in the lake?
2. What are the trade-offs involved in only collecting *some* fish of this species rather than all of them?
3. If the researcher decides to only collect data on some fish, what is the *random process* and a *random variable* they should be interested in?

Sampling

- In statistics, we typically want to make conclusions about a broader set of cases than those we have available to analyze
  - More specifically, we would like to use a **sample** to make conclusions about a **population**



In our previous example, the researcher might use data from *some* fish (sample) to draw conclusions about *all* of the lake's fish (population)

Notation for estimates and population parameters

Statisticians use notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample):

	Population Parameter	Estimate (from sample)
Mean	μ	$\overline{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

Sampling and Randomness

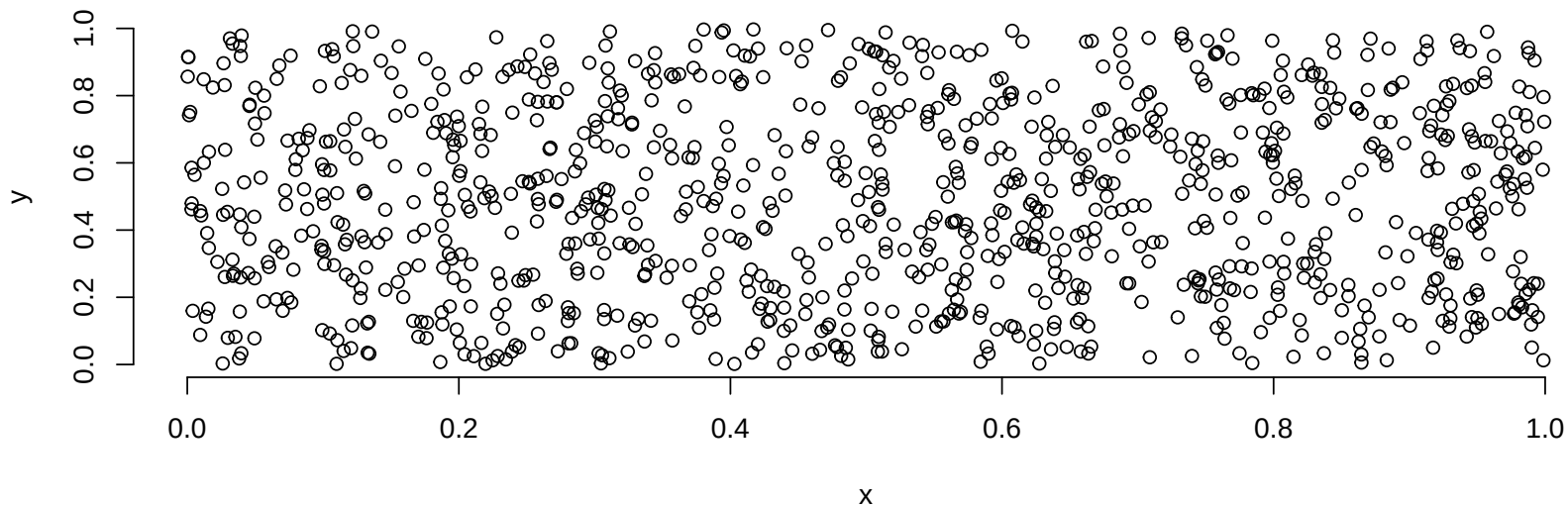
Estimates can differ from a population parameter for two reasons:

1. *Bias* - systematic issues in how the sample data were collected
2. *Variability* - chance deviations due to the underlying random process that produced the sample

For the time being we'll focus on sampling variability and we'll revisit the topic of sampling bias later this week.

Sampling Variability

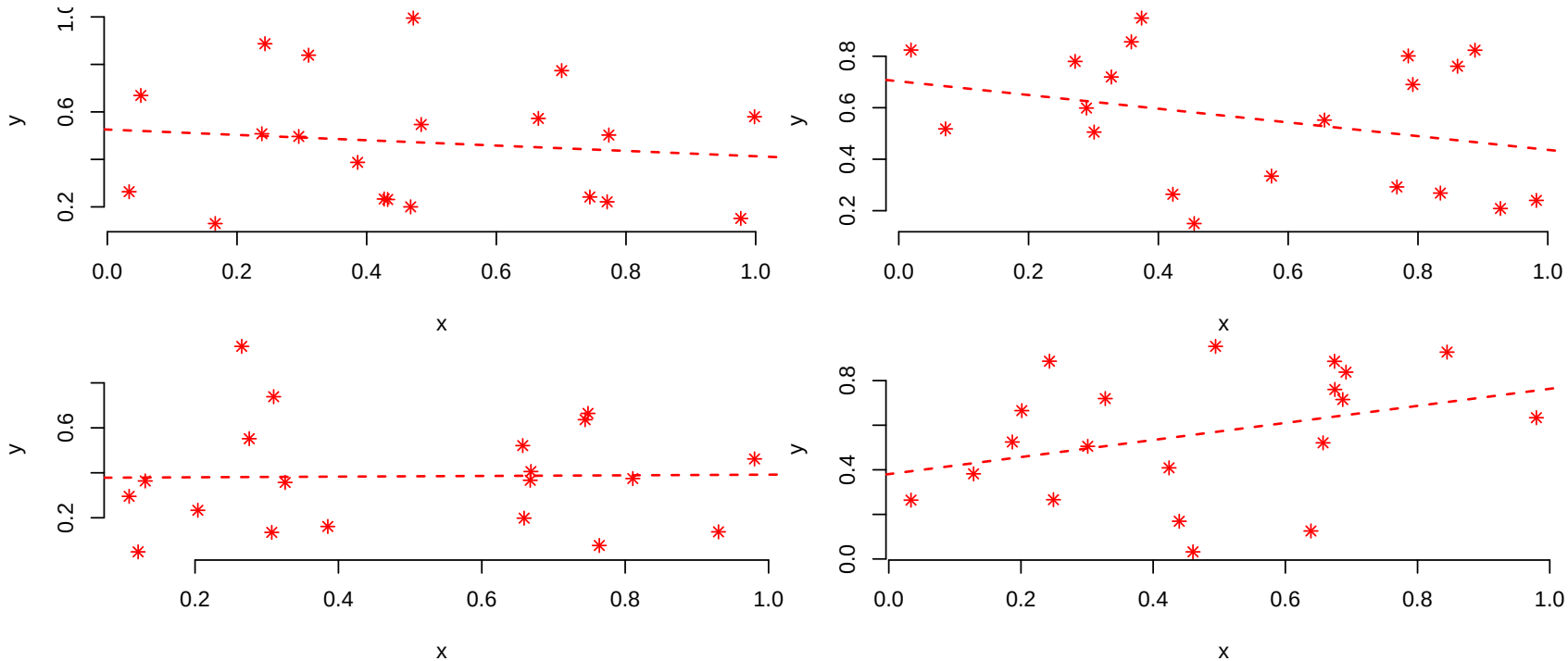
The scatter plot below depicts a *population* where the variables X and Y are *independent* (ie: $\rho = 0$):



If we randomly select a sample of 20 cases from this population, is it possible we observe a moderate correlation (say $r \approx 0.4$)? Is it likely?

Sampling Variability

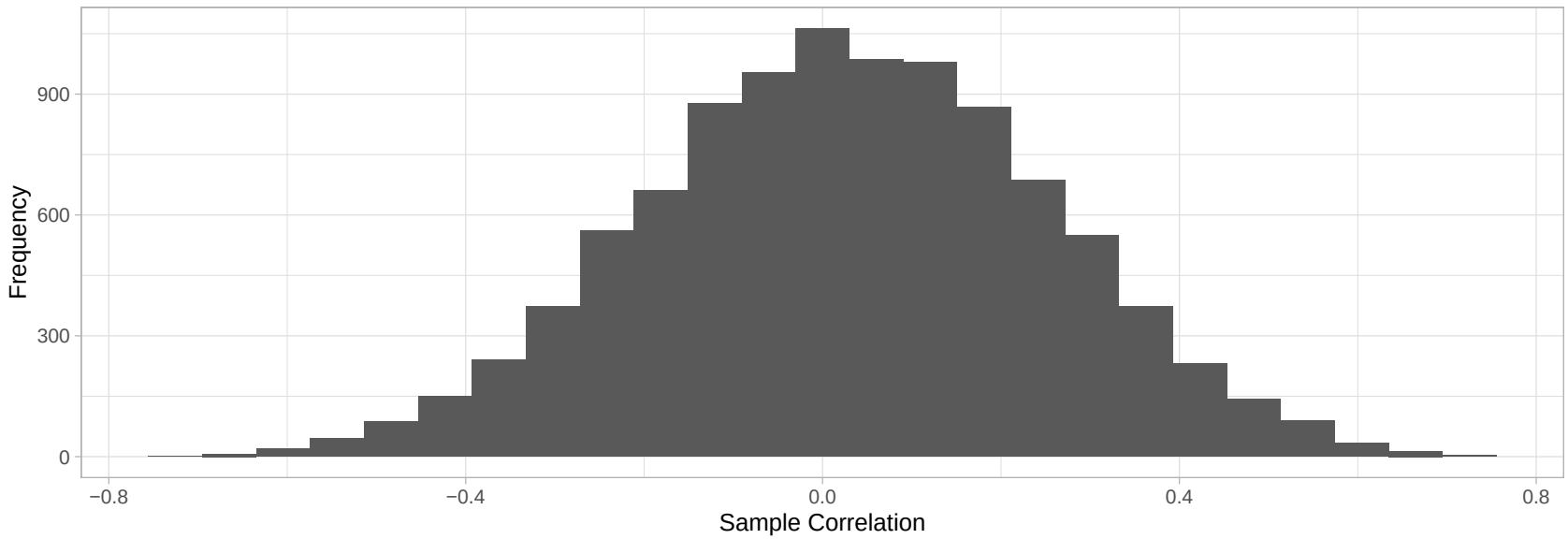
Shown below are four different random samples (each $n = 20$):



Across these samples, the observed sample correlations range from $r = -0.31$ (top right) to $r = 0.35$ (bottom right)

Sampling Variability

What if we drew 10,000 different random samples (of $n = 20$ cases) from this population and graphed the correlation coefficient found in each sample?



How might this relate to a *probability distribution*? What is the *random variable* involved in that probability distribution?

Conclusion

- Statistical analyses consider the influence of uncertainty in the observed results
  - Probability distributions provide the basis for understanding this uncertainty
- Sampling is the major source of uncertainty that we must account for during analysis
  - Sampling variability is a possible explanation for why trends in a sample of data might differ from trends in a population
  - Sampling bias is another explanation that we'll discuss later