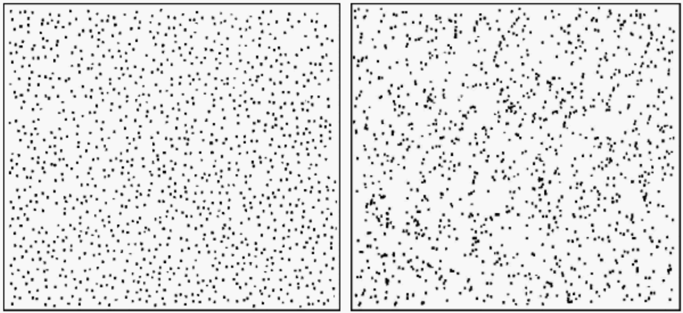
Introduction to Statistics

Ryan Miller

Statistics

One panel displays *randomly positioned* dots, the other shows dots whose positions reflect *meaningful patterns* (ie: biological/physical truths, etc.)



Which panel is which?

Statistics

A fundamental goal of “statistics” is to identify and understand meaningful patterns that exist within data under the *presence of uncertainty*. To do this we need:

1. Ways of expressing or describing patterns (descriptive statistics, visualizations, and models)
2. Contextual understanding (how were the data collected, what types of patterns are practically meaningful)
3. Methods to judge the role uncertainty in what we observed (is the pattern real, or could it be explained by chance?)

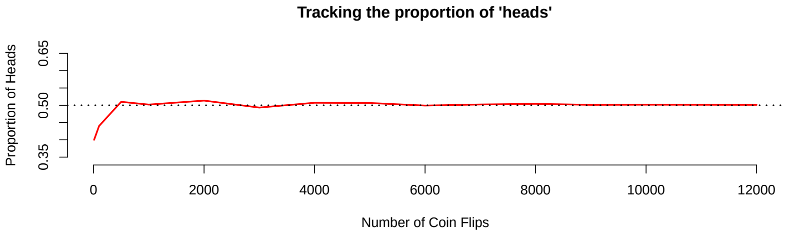
Understanding Uncertainty

Statisticians adopt the paradigm that data arises from a **random process**, meaning the observed outcomes/values are not knowable until after the random process has unfolded

- Example: The number showing after rolling a 6-sided die
  - We do not know the number until the die has been rolled
  - The result will vary if the random process is repeated
- Non-example: Converting a temperature from Celsius to Fahrenheit
  - There's no variability, the result will be the same every time

Probability

We will define **probability** as the *long-run relative frequency* (proportion) of an outcome over increasingly many repetitions of a random process:



The probability that "heads" is showing when a fair coin is flipped is 0.5 because we observe 50% of coin flips result in heads over a large number of repetitions.

Probability (cont.)

- Suppose we observe a large cohort of individuals during their first year of driving
    - Some number of these individuals will get into an accident, but these outcomes are not knowable in advance
- After a year has passed, 88 of 1000 drivers got into an accident
    - What's a reasonable estimate of the probability that a driver similar to those in this cohort gets into an accident during their first year of driving?

Probability (cont.)

- Probabilities sometimes differ depending upon the presence/absence of some factor
  - We call these **conditional probabilities**, and they can be estimated using *conditional proportions*

	Accident	No Accident
Other Vehicle	54	636
Truck	34	276

- The *conditional probability* of an accident given an individual drives a truck is $Pr(\text{Accident}|\text{Truck}) = \frac{34}{310} = 11\%$
  - This probability is higher than the *marginal probability* of any driver getting into an accident $Pr(\text{Accident}) = \frac{88}{1000} = 8.8\%$

Simulation and Statistics

- Suppose we want to know whether a certain coin is fair or biased
  - If we tossed the coin enough times we could assess whether $Pr(\text{Heads})$ was converging to 0.5
  - But what if after the 30th toss the coin rolled into a sewer and was lost?

Simulation and Statistics (cont.)

- Suppose we had recorded 18 “heads” across the 30 tosses we observed before losing the coin
  - We might approach the biased vs. fair question by considering $Pr(\geq 18 \text{ Heads} | \text{Coin is fair})$
- This probability relates to the amount of *evidence* our data provide against the presumption that the coin is fair
  - Importantly, we can calculate this probability as we can easily simulate the flipping of a fair coin
  - Today’s “lab” will further explore this idea of using conditional probability and simulations to make statistical conclusions