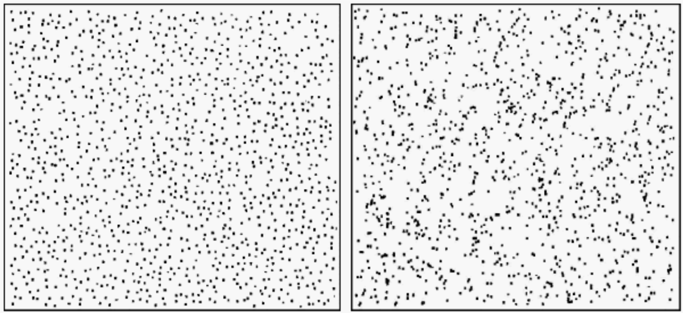
Introduction to Statistics

Ryan Miller

Grinnell College
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?)

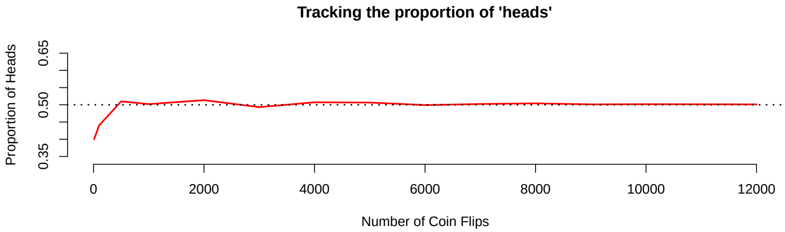
Understanding Uncertainty

Statisticians adopt the paradigm that data arise 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 50% of coin flips show 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 people will get into a crash, but those outcomes are not knowable in advance
- After a year has passed, 88 of 1000 drivers got into a crash
  - What's a reasonable estimate of the probability that a driver similar to those in this cohort crashes 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 a crash given an individual drives a truck is $Pr(\text{Accident} \mid \text{Truck}) = \frac{34}{310} = 11\%$
  - This probability is higher than the *marginal probability* of any driver crashing, which is $Pr(\text{Accident}) = \frac{88}{1000} = 8.8\%$

Simulation and Statistics

- Conditional probability provides us a tool for making inferences
  - For example, suppose we want to know whether a coin is fair or biased, but we only observed the results of 30 flips (which showed 18 heads) before losing the coin
- By the definition of probability, we know that if we tossed the coin enough times we could check whether the proportion of heads was converging to 0.5
  - But when limited to a relatively small number of flips we might not know how close to 0.5 is close enough to be convinced
  - Considering 18/30 is 0.6, do you think this coin was fair?

Simulation and Statistics (cont.)

- Using conditional probability, we might consider $Pr(\text{18 or more Heads}|\text{Coin is fair})$
  - Because it's easy to *simulate* the outcomes of fair coin flips, we could perform enough simulations to reliably estimate this conditional probability
- If $Pr(\text{18 or more Heads}|\text{Coin is fair})$ is found to be small, we might consider that evidence against our coin having been fair, because the data we observed would be very unlikely to occur if the coin were fair
  - This is the fundamental idea behind one of the most widely used tools in statistics: the *hypothesis test*
  - Today's lab will further explore this idea of using simulations and conditional probability to make statistical conclusions