Probability (part 1)

Ryan Miller

Introduction

- *Statistical inference* (generalizing something from a sample to population) inherently involves uncertainty
  - Uncertainty in which cases are sampled
  - Uncertainty in who will be assigned to the treatment/control groups

Introduction

- *Statistical inference* (generalizing something from a sample to population) inherently involves uncertainty
  - Uncertainty in which cases are sampled
  - Uncertainty in who will be assigned to the treatment/control groups
- In able to quantify this uncertainty, statisticians need to study probability
  - However, in this course we'll only go as far into probability as is necessary to understand basic methods of statistical inference
  - If you want a more thorough look at probability, take MATH-311 - Probability Theory

Basic Definitions

- A **random process** describes any phenomenon whose outcome cannot be predicted with 100% certainty
- A **sample space** refers to the collection of possible outcomes from a random process
- An **event** is an outcome of a random process

Process	Space	Outcome
Flipping a Coin	{H,T}	Seeing H
Rolling a 6-sided Die	{1,2,3,4,5,6}	Seeing an odd number
Person takes Vaccine	{Disease, No Disease}	No Disease

Probability Distributions

- **Probability distributions** are mappings of *disjoint* events to probabilities
- Below is an example probability distribution for the sum of rolling two dice:

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

Probability Distributions

- **Probability distributions** are mappings of *disjoint* events to probabilities
- Below is an example probability distribution for the sum of rolling two dice:

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

- A *valid* probability distribution must satisfy the following:
  - outcomes must be *disjoint* (ie: it's impossible for the dice to sum to both 3 and 4)
  - each outcome must have a probability ≥ 0 (ie: nothing in the sample space is unobservable)
  - the entire sample space has a probability of 1 (ie: the probability row sums to 1)

Frequentist Probability

- How are these probabilities determined? Everyone agrees the probability of rolling a 1 on a fair die is $1/6$, but why?

Frequentist Probability

- How are these probabilities determined? Everyone agrees the probability of rolling a 1 on a fair die is $1/6$, but why?
- The most widely adopted viewpoint defines the probability of an event as its *long run frequency*
  - Statisticians who adopt this perspective are known as **frequentist** statisticians

Law of Large Numbers

- The frequentist view of probability is rooted in the *law of large numbers*
- Let $\hat{p}_n$ denote the observed proportion of an outcome after n repetitions of a random process
  - For example, the number of "heads" observed in $n = 10$ coin flips
- Given p is the true probability of that outcome, the *law of large numbers* suggests that $\hat{p}_n \rightarrow p$ as n increases towards infinity
  - Thus, as a coin is flipped more and more times, the observed proportion of heads converges to a constant ($p = 0.5$ for a fair coin)

Example

- The infection status of COVID-19 vaccine recipients can be viewed as a random process where each recipient either remains healthy or contracts the virus
  - Suppose the true probability of a vaccinated individual developing Coronavirus is 0.001 or 0.1% (note that this is typically unknown to us!)

1) Consider a sample of 2 vaccinated individuals, what is the sample space?
2) Consider samples of 1,000 and 100,000 individuals, which sample is more likely to have an infection rate closer to a 0.1%? Why?

Example (solution)

1) For each individual, the sample space is $\{$Disease, No Disease$\}$. So for two individuals, the collection of possible outcomes is: $\{(\text{D, D}), (\text{D, ND}), (\text{ND, D}), (\text{ND, ND})\}$
2) The sample of 100,000 individuals by the law of large numbers

Probability for Non-Repeatable Events

- Not every application of probability involves repeatable events
  - What is the probability of life on Mars?
  - What is the probability that Donald Trump is re-elected in 2020?
  - Can we still apply probability to these processes and outcomes?

Probability for Non-Repeatable Events

- Not every application of probability involves repeatable events
  - What is the probability of life on Mars?
  - What is the probability that Donald Trump is re-elected in 2020?
  - Can we still apply probability to these processes and outcomes?
- Most agree the answer is yes, but these situations make more sense under different perspective on probability
  - *Bayesian* statisticians view probability as a combination of subjective prior information and previously observed outcomes
  - We won't work with Bayesian probabilities in this course, but it's an interesting area of statistics that I encourage you to look into

Probability Rules

When working with probability, we often want to consider multiple outcomes from random processes. There are *three* major probability rules involved in these calculations:

1) The complement rule
2) The addition rule
3) The multiplication rule

We'll briefly go through these one-by-one, often starting with the *special case* of **independent** random processes before moving to the general rule

The Complement Rule

- Let A be an event and A^C be the complement of that event
  - For example, if A is contracting COVID-19, then A^C is not contracting COVID-19
- The **complement rule** states: $P(A^C) = 1 - P(A)$
  - Or if the probability of contracting COVID-19 is 0.02, the probability of not contracting it is 0.98

The Addition Rule

- For *disjoint* or *mutually exclusive events*, the **addition rule** states: $P(A \text{ or } B) = P(A) + P(B)$

The Addition Rule

- For *disjoint* or *mutually exclusive events*, the **addition rule** states: $P(A \text{ or } B) = P(A) + P(B)$
- **Example:** Let event A be rolling a number 3 or lower, and event B be rolling an odd number
- $P(A) = .5$ and $P(B) = 0.5$, but does $P(A \text{ or } B) = 1$?

The Addition Rule

- For *disjoint* or *mutually exclusive events*, the **addition rule** states: $P(A \text{ or } B) = P(A) + P(B)$
- **Example:** Let event A be rolling a number 3 or lower, and event B be rolling an odd number
- $P(A) = .5$ and $P(B) = 0.5$, but does $P(A \text{ or } B) = 1$?
  - Clearly the answer is no, we could roll a 4 or a 6, which isn't "A or B"
  - The issue is that these are not disjoint events, so we were "double-counting" certain outcomes

The Addition Rule (general)

- More generally, $P(A \text{ or } B) = P(A) + P(B) + P(A \text{ and } B)$
  - So in our last question, $P(A) = 0.5$, $P(B) = 0.5$, and $P(A \text{ and } B) = 0.333$, so $P(A \text{ or } B) = 0.5 + 0.5 - 0.333 = 0.667$

The Addition Rule (general)

- More generally, $P(A \text{ or } B) = P(A) + P(B) + P(A \text{ and } B)$
  - So in our last question, $P(A) = 0.5$, $P(B) = 0.5$, and $P(A \text{ and } B) = 0.333$, so $P(A \text{ or } B) = 0.5 + 0.5 - 0.333 = 0.667$
  - We can verify this by considering the complement of A or B, which would be rolling a 4 or 6, a result that clearly has a probability of $1/3$

The Multiplication Rule

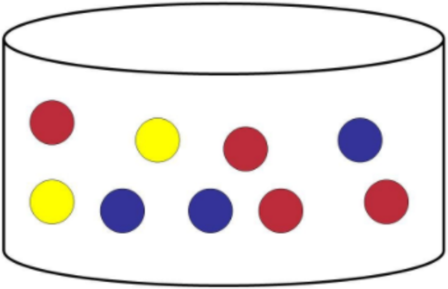
- Two random processes are deemed **independent** if the outcome of one does not influence the outcome of the other
- Suppose A and B are outcomes of independent processes, the **multiplication rule** states $P(A \text{ and } B) = P(A) * P(B)$
  - For example, if we flip a coin then roll a die, $P(\text{Heads and } 6) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$
  - These events are obviously independent, the coin and die share no relationship

Dependent Events

- Very often the probability of an event will depend on other events, a notion known as **conditional probability**

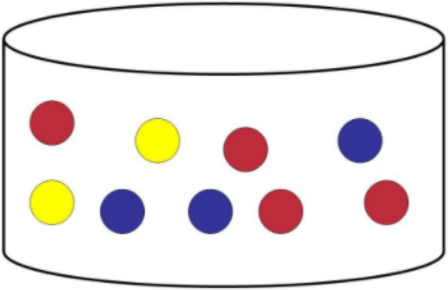
Dependent Events

- Very often the probability of an event will depend on other events, a notion known as **conditional probability**
- To understand this concept, let's consider sampling from a small population
  - let A_1 denote randomly drawing a yellow ball, clearly $P(A_1) = 2/9$



Sampling from a Small Population

- Now consider A_i to represent the i^{th} draw from the urn being a yellow ball
  - What is $P(A_1 \text{ and } A_2)$?
  - What about $P(A_1 \text{ and } A_2 \text{ and } A_3)$?



The Incorrect Approach

- The multiplication rule suggests:
  - $P(A_1 \text{ and } A_2) = P(A_1) * P(A_2) = \frac{2}{9} * \frac{2}{9} \approx 0.05$
  - $P(A_1 \text{ and } A_2 \text{ and } A_3) = P(A_1) * P(A_2) * P(A_3) = (\frac{2}{9})^3 \approx 0.01$

The Incorrect Approach

- The multiplication rule suggests:
  - $P(A_1 \text{ and } A_2) = P(A_1) * P(A_2) = \frac{2}{9} * \frac{2}{9} \approx 0.05$
  - $P(A_1 \text{ and } A_2 \text{ and } A_3) = P(A_1) * P(A_2) * P(A_3) = (\frac{2}{9})^3 \approx 0.01$
- Unfortunately, the multiplication rule assumes we are working with *independent* random processes
  - In this example, getting a yellow ball on the first draw changes our chances of getting a yellow ball on the second draw

Multiplication Rule (general)

- It's quite easy to see $P(A_1) = 2/9$ and $P(A_2|A_1) = 1/8$
  - Notice $P(A_3|A_1, A_2) = 0$
  - These are examples of **conditional probability**, and they lead us to a *general multiplication rule*:

Multiplication Rule (general)

- It's quite easy to see $P(A_1) = 2/9$ and $P(A_2|A_1) = 1/8$
  - Notice $P(A_3|A_1, A_2) = 0$
  - These are examples of **conditional probability**, and they lead us to a *general multiplication rule*:

$$P(A \text{ and } B) = P(A) * P(B|A)$$

- Note that sometimes event B is *completely unaffected* by event A, in which case $P(B|A) = P(B)$
  - Sampling *with replacement* is an example of this

Probability and Contingency Tables

- We've previously seen *contingency tables* as a method used to relate two categorical variables
  - We saw that *row proportions* (or column proportions) were particularly useful in describing potential associations

Probability and Contingency Tables

- We've previously seen *contingency tables* as a method used to relate two categorical variables
  - We saw that *row proportions* (or column proportions) were particularly useful in describing potential associations
- Since we've defined probability as a long run frequency, it makes sense to use proportions observed in a sample as our *best estimate* of certain probabilities

Probability and Contingency Tables

- We've previously seen *contingency tables* as a method used to relate two categorical variables
  - We saw that *row proportions* (or column proportions) were particularly useful in describing potential associations
- Since we've defined probability as a long run frequency, it makes sense to use proportions observed in a sample as our *best estimate* of certain probabilities

	death	not
black	38	142
white	46	152

- In the Florida Death Penalty example, we might estimate $P(\text{Death}|\text{WhiteOffender}) = 46/152 = 0.23$

Example #1 - Spam Classification

A classic data science problem involves creating an algorithm to detect "spam" emails. Below are the results of an algorithm applied to 1265 emails:

	Spam	Real
Included in Inbox	29	978
Marked as Spam	209	49

1) Estimate the probability of an email being spam
2) Estimate the conditional probability an email being spam given it was marked as spam by the algorithm
3) Estimate the conditional probability an email being marked as spam given that it is actually spam
4) Estimate the probability of an email being spam *and* being marked as spam by the algorithm
5) Using 1-3, verify the general multiplication rule for this application

Example #1 (solution)

1) Of the 1265 emails, 238 were spam. So, $P(\text{Spam}) = 238/1265 = 0.188$
2) Of the 258 emails marked as spam, 209 were actually spam. So, $P(\text{Spam}|\text{Marked Spam}) = 209/258 = 0.810$
3) Of the 238 emails that were spam, 209 were marked as spam. So, $P(\text{Marked Spam}|\text{Spam}) = 209/238 = 0.878$
4) Of the 1265 emails, 209 were spam *and* marked as spam by the algorithm. So, $P(\text{Spam and Marked Spam}) = 209/1265 = 0.165$
5) The general multiplication rule states: $P(\text{Spam and Marked Spam}) = P(\text{Spam}) * P(\text{Marked Spam}|\text{Spam}) \implies 209/1265 = \frac{238}{1265} * \frac{209}{238}$

Example #2 - Free Throw Shooting

- Shaquille O'Neal, or Shaq, is regarded as one of the best NBA centers of all-time despite being a notoriously bad free throw shooter
- During his career, Shaq made only 52% of his free throw attempts
- Because of his poor free throw shooting, many teams used the "hack-a-shaq" strategy to force him to shoot free throws
  - But was this strategy actually helpful?

Example #2 - Free Throw Shooting

Example #2 - Free Throw Shooting

1) Suppose Shaq is fouled and gets to shoot 2 free throws, what is the sample space?
2) What is the probability distribution for *the number of points* resulting from two free throws? (based upon Shaq's 52% career FT percentage)
3) Suppose your teaming is leading by 2 points in the last possession of a close game when Shaq gets the ball in the paint, should you foul him?

Example #2 (solution)

1) Letting H denote a hit, or made free throw, and M denote a missed free throw, sample space is $\{MM, MH, HM, MM\}$
2) Possible points are $\{0, 1, 2\}$, the probabilities of these outcomes are shown below:

Points	0	1	2
Probability	$.47 * .47 = .22$	$.47 * .53 + .53 * .47 = 0.50$	$.53 * .53 = .28$

3) While Shaq has only a 0.28 chance of hitting both free throws, this decision is complex due to other areas of uncertainty (how likely was Shaq to convert the post touch, how likely is an offensive rebound following the second free through, etc.)

Conclusion

- This presentation introduced several important definitions and laws of probability
- The examples we saw were interesting, but they weren't *statistical inference*
  - In the next presentation we'll get into how probability relates to evaluating results observed in sample data