Statistics Workshop

Ivan Corneillet

Lead Data Science Instructor, Principal Data Scientist

Why this Statistics Workshop?

As part of the application process for Galvanize's Data Science Immersive program, you'll receive a take-home exercise and have two interviews with Galvanize faculty, one to assess your skills in Python, and another to evaluate your skills in statistics, math, and probability.

This workshop surveys the concepts in probability and statistics that are covered in the second interview.

Statistics is the science of making effective use of numerical data.

It deals with all aspects of this, including the:

- collection,
- analysis, and
- interpretation of data.

"We must become more comfortable with probability and uncertainty." – Nate Silver

Agenda

Day 1: **Basics of Probability**

- Introduction
- Combinatorics
- Probability Basics
- Conditional Probability
- Independence
- Bayes' Formula

Day 2: **Basics of Statistics** and **Basics of Machine Learning**

- Random Variables
- Common Distributions
- Linear Regression
- Logistic Regression

Day 1: Basics of Probability

Day 1: Basics of Probability

- Introduction
- Combinatorics
- Probability Basics
- Conditional Probability
- Independence
- Bayes' Formula

Introduction

Introduction

- What's Probability?
- What's Statistics?
- What's the difference?

What's Probability?

- **Probability** refers to the study of patterns in a random process in which that **all** basic features of the random process are **known**.
- Our goal is to **discover** other **deeper features** of the random process.

E.g., it is a problem in probability to determine when presented with a coin **known** to be fair, how often it will never land on heads over ten consecutive flips.

What's Statistics?

- **Statistics** refers to the study of a random process in which **some** basic features of the random process are **unknown**.
- Our goal is to **infer** from observations basic, **hidden features** of the random process.

E.g., it is a problem in statistics to determine when presented with a coin which has landed tails ten consecutive times, whether one should continue to **believe** it is fair.

Combinatorics

Combinatorics

The basic problem solving skill you need to solve problems in probability is **counting** (yep, really...).

Combinatorics

E.g.,

- How many ways are there to arrange four letters of the alphabet?
- How many ways are there to arrange four different letters of the alphabet?
- How many ways are there to arrange 25 books on a bookshelf?
- How many hands are full houses?
- How many hands have three-of-a-kind?
- How many hands have three-of-a-kind and are not also full houses?

Basic Counting Principle

If a task can be accomplished as a series of steps, then the number of outcomes of the task is the **product** of the number of outcomes for each individual step.

E.g., how many ways are there to arrange four letters of the alphabet?

E.g., how many ways are there to arrange four letters of the alphabet?

(how can we accomplish this task as a step by step process?)

- Pick the first letter, write it down:
  - 26
- Pick the second letter, write it down:
  - $\times 26$
- Pick the third letter, write it down:
  - $\times 26$
- Pick the fourth letter, write it down:
  - $\times 26$

$$26 \times 26 \times 26 \times 26 = 456,976$$

How would you change your answer if we could not re-use a letter?

How would you change your answer if we could not re-use a letter?

(how can we accomplish this task as a step-by-step process?)

- Pick the first letter, write it down:
  - 26
- Pick the second letter, write it down:
  - $\times 25$
- Pick the third letter, write it down:
  - $\times 24$
- Pick the fourth letter, write it down:
  - $\times 23 = 358,800$

$$26 \times 25 \times 24 \times 23 = 358,800$$

How would you change your answer if we could not re-use a letter?

This example is very common: we are pulling from a pool of objects and we **cannot** re-use an object once selected.

Ordered selections without replacement

The number of **ordered** selections of k objects **without replacement** from a population for n objects is called the **number of permutations of k objects taken from n**.

$$P(n,k) = \underbrace{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!}$$

E.g., you have 25 books on a bookshelf; how many ways are there to arrange these books in any order?

E.g., you have 25 books on a bookshelf; how many ways are there to arrange these books in any order?

$$25 \times 24 \times 23 \times \ldots \times 2 \times 1 = 25! = 15,511,210,043,330,985,984,000,000$$

What if we have a procedure in which the order of choices does not matter?

E.g., how many hands are possible when drawing from a standard (52-card) deck?

(a hand is an **unordered** collection of five cards.)

Note on (Poker) Hands

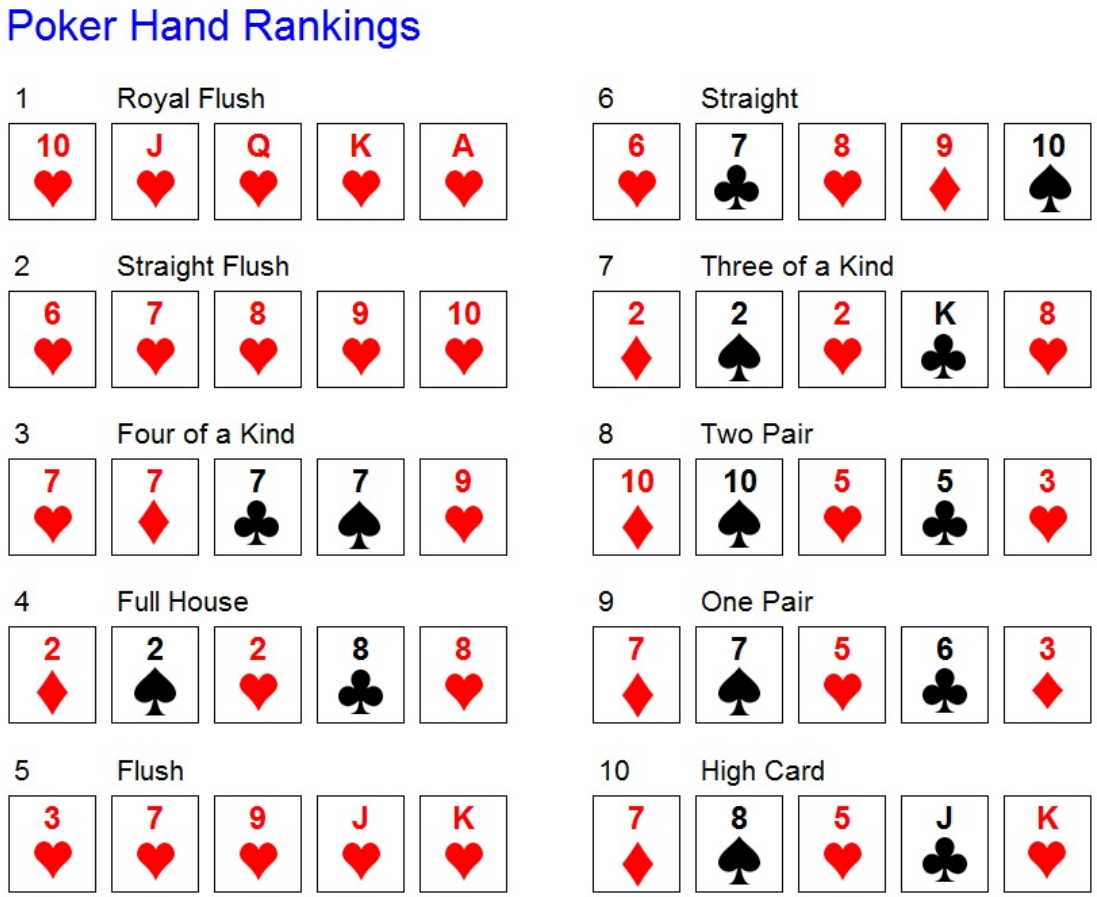

Figure 1:Poker Hands (Source: All Math Considered)

E.g., how many hands are possible when drawing from a standard deck?

E.g., how many hands are possible when drawing from a standard deck?

(to choose an ordered list of five cards, first chose the five cards you want to use, then choose a way to order them.)

$$\overbrace{\#\text{ of ordered hands}}^{52\times51\times50\times49\times48} =$$

$$\#\text{ of unordered hands} \times \underbrace{\#\text{ of ways to order a hand}}_{5\times4\times3\times2\times1}$$

E.g., how many hands are possible when drawing from a standard deck?

$$\#\text{ of unordered hands} = \frac{\overbrace{\#\text{ of ordered hands}}^{52\times51\times50\times49\times48}}{\underbrace{\#\text{ of ways to order a hand}}_{5\times4\times3\times2\times1}}$$

Unordered selections without replacement

$$\#\text{ of unordered selections} = \frac{\#\text{ of ordered selections}}{\#\text{ of ways to order a single selection}}$$

E.g.,

$$\#\text{ of unordered hands} = \frac{\#\text{ of ordered hands}}{\#\text{ of ways to order a hand}}$$

Unordered selections without replacement

The number of **unordered** selections of k objects **without replacement** from a population for n objects is called the **number of combinations of k objects taken from n**.

$$C(n,k) = \frac{P(n,k)}{P(k,k)} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

How many hands are full houses? (a full house is a hand with a pair of the same value and a three-of-a-kind of the same value.)

How many hands are full houses?

(how can we accomplish this task as a step-by-step process?)

- Pick the rank of the three-of-a-kind
  - $C(13, 1)$
- Pick three suits of that rank
  - $\times C(4, 3)$
- Pick the rank of the pair (from the remaining 12 ranks)
  - $\times C(12, 1)$
- Pick two suits of that rank
  - $\times C(4, 2)$

$$C(13, 1) \times C(4, 3) \times C(12, 1) \times C(4, 2) = 13 \times 4 \times 12 \times 6 = 3,744$$

How many hands are there that contain a three-of-kind?

How many hands are there that contain a three-of-kind that are not full houses?

Probability Basics

Probability Basics

- An **outcome** is a single thing that can happen.

E.g., when thinking about poker hands, an outcome is a single hand.

Probability Basics

- An **event** is a collection of **desired** outcomes.

E.g., the collection of all full-houses, three-of-a-kinds, etc. are events.

Probability Basics

The **probability** of an event is:

$$P(\text{event}) = \frac{\#\text{ of desired outcomes}}{\#\text{ of total outcomes}}$$

So to compute basic probabilities, we use our knowledge from combinatorics.

What's the probability of drawing a full-house?

What's the probability of drawing a full-house?

$$P(\text{full house}) = \frac{\#\text{ of full houses}}{\#\text{ of hands}}$$

$$\#\text{ of full houses} = \binom{13}{1} \times \binom{4}{3} \times \binom{12}{1} \times \binom{4}{2} = 3,744$$

$$\#\text{ of hands} = \binom{52}{5} = 2,598,960$$

$$P(\text{full house}) = \frac{3,744}{2,598,960} = .0014 = .14\%$$

What's the probability of drawing a hand containing a three-of-a-kind?

What's the probability of drawing a hand containing a three-of-a-kind that is not a full house?

What's the probability of drawing a hand containing a pair, that does not contain a three-of-a-kind or four-of-a-kind?

Conditional Probability

Conditional Probability

Suppose we know that one event B has already happened or will happen (the condition), and we want to know the probability of different event A.

Then the **conditional probability of A given B** is defined by:

$$P(A \mid B) = \frac{\#\text{ of desired outcomes for } A \text{ and } B}{\#\text{ of desired outcomes for } B}$$

E.g., what's the conditional probability that you draw a hand containing a three-of-a-kind, given that you draw a hand containing a pair?

E.g., what's the conditional probability that you draw a hand containing a three-of-a-kind, given that you draw a hand containing a pair?

$$P(\text{w/ three-of-a-kind} \mid \text{w/ pair}) = \frac{\#\text{ of w/ three-of-a-kinds}}{\#\text{ of w/ pairs}}$$

(note that a hand containing a three-of-a-kind automatically contains a pair!)

$$\#\text{ of w/ three-of-a-kinds} = \binom{13}{1} \times \binom{4}{3} \times \binom{49}{2} = 61,152$$

$$\#\text{ of w/ pairs} = \binom{13}{1} \times \binom{4}{2} \times \binom{50}{3} = 1,528,800$$

$$P(\text{w/ three-of-a-kind} \mid \text{w/ pair}) = \frac{61,152}{1,528,800} = .04 = 4\%$$

Another way to look at conditional probabilities

$$P(A \mid B) = \frac{\#\text{ of desired outcomes for } A \text{ and } B}{\#\text{ of desired outcomes for } B}$$

$$= \frac{\frac{\#\text{ of desired outcomes for } A \text{ and } B}{\#\text{ of total outcomes}}}{\frac{\#\text{ of desired outcomes for } B}{\#\text{ of total outcomes}}}$$

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

Another way to look at conditional probabilities

So we could have instead computed:

$$P(\text{w/ three-of-a-kind} \mid \text{w/ pair}) = \frac{P(\text{w/ three-of-a-kind})}{P(\text{w/ pair})}$$

$$= \frac{\frac{61{,}152}{2{,}598{,}960}}{\frac{1{,}528{,}800}{2{,}598{,}960}} = .04$$

What's the conditional probability that you draw a full house, given that you know you have at least a three-of-a-kind?

What's the conditional probability that you draw a full house, given that you know you have at least a three-of-a-kind?

$$P(\text{full house} \mid \text{w/ three-of-a-kind}) = \frac{\#\text{ of full houses}}{\#\text{ of w/ three-of-a-kinds}}$$

$$= \frac{\binom{13}{1} \times \binom{4}{3} \times \binom{12}{1} \times \binom{4}{2}}{\binom{13}{1} \times \binom{4}{3} \times \binom{49}{2}} = \frac{3,744}{61,152} = .061 = 6.1\%$$

What's the conditional probability that you draw a full house, given that you have drawn a four-of-a-kind?

What's the conditional probability that you draw a four-of-a-kind, given that you know you already have a pair?

Independence

Independence

Two events A and B are called **independent** when

$$P(A \mid B) = P(A)$$

This means that **knowledge that B has or will occur does not change our knowledge about whether A will occur**.

Independence

Remember that the definition of conditional probability is:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Combining this with our definition of independence (so, below that A and B are independent):

$$\frac{P(A \text{ and } B)}{P(B)} = P(A \mid B) = P(A)$$

Or, rearranging things:

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

(this equation is sometimes used as the definition of independence.)

Common applications of independence generally go like this:

- We deduce from context that some event bears no influence on another.
- We conclude that the two events are independent.
- We use either of the two equations defining independence to do calculations.

You roll a six-sided die six times, what's the probability that you roll all the possible numbers in decreasing order?

You roll a six-sided die six times, what's the probability that you roll all the possible numbers in decreasing order?

Let's call the values of the rolls R_1, R_2, R_3, R_4, R_5, and R_6. We are looking for:

$$P(R_1 = 6, R_2 = 5, R_3 = 4, R_4 = 3, R_5 = 2, R_6 = 1)$$

$$= P(R_1 = 6)P(R_2 = 5)P(R_3 = 4)P(R_4 = 3)P(R_5 = 2)P(R_6 = 1)$$

(all six rolls are independent: can break up and multiply.)

$$= \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{46,656}$$

(each individual roll of the die has six possible outcomes and only one of them is the number we are looking for.)

Suppose you have a bucket with 5 red and 5 yellow balls in it, which you draw in sequence, without replacement. Are the events “You draw a red ball first” and “You draw a yellow ball second” independent?

Bayes' Formula

Bayes' Formula

Remember our definition of conditional probability:

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

Setting these equal to one another leads us to **Bayes' Formula:**

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

This is probably the most important simple formula in both probability and statistics.

The Disease Screening Problem

Suppose we have developed a test for a certain disease:

- Only 1% of people have the disease.
- If a person has the disease, the test will be positive 99.9% of the time.
- If a person does not have the disease, the test will be negative 98% of the time.

You get tested for the disease, and the test is positive. What's the probability that you actually have the disease?

The Disease Screening Problem

We are looking for the following conditional probability:

$$P(\text{have disease} \mid \text{test is positive})$$

And we **know** that:

- $P(\text{have disease}) = .01$
- $P(\text{test is positive} \mid \text{have disease}) = .999$
- $P(\text{test is positive} \mid \text{don't have disease}) = .02$

The Disease Screening Problem

Using Bayes':

$$P(\text{have disease} \mid \text{test is positive})$$

$$= \frac{P(\text{test is positive} \mid \text{have disease})P(\text{have disease})}{P(\text{test is positive})}$$

We know all the things appearing in the previous formula except $P(\text{test is positive})$.

We can calculate the last piece by breaking it down

$$P(\text{test is positive})$$

$$= P[\text{test is positive and } \underbrace{(\text{have disease or don't have disease})}_{\text{everything}}]$$

$$= P[(\text{test is positive and have disease})$$

$$\text{or } (\text{test is positive and don't have disease})]$$

$$= P(\text{test is positive and have disease})$$

$$+P(\text{test is positive and don't have disease})$$

(since both events are disjoints.)

We can calculate the last piece by breaking it down

$$= P(\text{test is positive} \mid \text{have disease})P(\text{have disease})$$

$$+P(\text{test is positive} \mid \text{don't have disease})P(\text{don't have disease})$$

(apply conditional probability formula twice.)

$$= .02 \times .99 + .999 \times .01$$

The Disease Screening Problem

$$P(\text{have disease} \mid \text{test is positive})$$

$$= \frac{P(\text{test is positive} \mid \text{have disease})P(\text{have disease})}{P(\text{test is positive})}$$

$$= \frac{.999 \times .01}{.999 \times .01 + .02 \times .99}$$

$$= .34$$

$$= 34\%$$

The Base Rate Fallacy

The probability we have the disease is only 34%, even though we received a positive test.

- This kind of result is unintuitive to almost all humans, a mental bias called the **base rate fallacy**.

Pretty much everyone's intuition says that it should be much more likely that the person does have the disease after a test comes back positive.

- Pretty much everyone undervalues the prior information that

$$P(\text{have disease}) = .01$$

It takes a lot of evidence to make an unlikely situation likely.

Here is a way to think about this:

Suppose there are 1,000 people (in the universe). Then:

- $.99 \times 1,000 = 10$ actually have the disease;
- of those, $10 \times .999 \approx 10$ will test positive;
- of the remaining that do not have the disease, about $990 \times .02 \approx 20$ will also test positive!

So if all you know is that you have received a positive test, you can only conclude that you are one of the thirty, of which only ten actually have the disease!

There is some terminology that is often used to understand these relationships. These ideas form the basis of **Bayesian statistics**.

P(have disease) is called the **prior probability**.

- It is what we know **before collecting evidence/data**.

P(test is positive | have disease) is called the **likelihood**.

- It is the **strength of the evidence/data we collected**.

P(have disease | test is positive) is called the **posterior**.

- It is what we know, **after collecting evidence/data**.

Suppose you get a second test, which also comes out positive. What's the posterior probability that you actually have the disease?

Suppose you get a second test, which comes back negative. What's the posterior probability that you actually have the disease?