Independence
Chris Piech, CS109
Announcements

- **Sections** start today. Wahoo! Enjoy.
- **PSet #1** is due Friday 1p. Recall grace period.
Today, start with a cool program
<table>
<thead>
<tr>
<th>Index</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>False, True, False, False, True, False</td>
</tr>
<tr>
<td>2</td>
<td>True, True, False, True, True, False</td>
</tr>
<tr>
<td>3</td>
<td>True, True, False, True, True, True</td>
</tr>
<tr>
<td>4</td>
<td>False, True, False, True, True, False</td>
</tr>
<tr>
<td>5</td>
<td>False, True, False, True, True, False</td>
</tr>
<tr>
<td>6</td>
<td>True, True, False, True, True, True</td>
</tr>
<tr>
<td>7</td>
<td>True, True, False, True, True, True</td>
</tr>
<tr>
<td>8</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>9</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>10</td>
<td>False, True, False, True, False, False</td>
</tr>
<tr>
<td>11</td>
<td>False, True, False, True, False, False</td>
</tr>
<tr>
<td>12</td>
<td>False, True, False, True, False, False</td>
</tr>
<tr>
<td>13</td>
<td>False, True, False, True, False, False</td>
</tr>
<tr>
<td>14</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>15</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>16</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>17</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>18</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>19</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>20</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>21</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>22</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>23</td>
<td>True, True, False, True, True, False</td>
</tr>
<tr>
<td>24</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>25</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>26</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>27</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>28</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>29</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>30</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>31</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>32</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>33</td>
<td>True, False, True, False, True, False</td>
</tr>
<tr>
<td>34</td>
<td>True, True, False, True, True, False</td>
</tr>
<tr>
<td>35</td>
<td>True, True, False, True, True, False</td>
</tr>
<tr>
<td>36</td>
<td>False, False, True, False, True, False</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Discovered Hypothesis

\[ p(T|G_1 \text{ and } G_2) = 0.9 \]
\[ p(T|\sim G_1 \text{ or } \sim G_2) = 0.2 \]

\[ p(G_1) = 0.5 \]
\[ p(G_5) = 0.6 \]
\[ p(G_2|G_5) = 0.9 \]
\[ p(G_2|\sim G_5) = 0.2 \]

These genes don’t impact T
We’ve gotten ahead of ourselves

Source: The Hobbit
Start at the beginning

Source: The Hobbit
Review
Review: Conditional Probability

$p(AB) \text{ vs } p(A|B)$

$P(AB) = P(A|B)P(B)$
Review: Chain Rule

Definition of conditional probability:

\[ P(E|F) = \frac{P(EF)}{P(F)} \]

The Chain Rule:

\[ P(EF) = P(E|F)P(F) \]
Relationship Between Probabilities

- **$P(E \text{ and } F)$**
  - Chain rule (Product rule)
  - Definition of conditional probability

- **$P(E \mid F)$**

- **$P(E)$**

- **$P(F \mid E)$**

- **Law of Total Probability**

- **Bayes’ Theorem**
To day

OR
\[ P(E \cup F) \]

AND
\[ P(EF) \]

DeMorgan’s

Mutually Exclusive?

Just Add!

Inclusion Exclusion

Independent?

Just Multiply

Chain Rule
Today

OR
P(E ∪ F)

AND
P(EF)

DeMorgan’s

Mutually Exclusive?

Just Add!

Inclusion Exclusion

Independent?

Just Multiply

Chain Rule
Probability of “OR”
If events are mutually exclusive, probability of OR is simple:

\[ P(E \cup F) = P(E) + P(F) \]

Piech, CS109, Stanford University
If events are mutually exclusive, probability of OR is simple:

\[ P(E \cup F) = \frac{7}{50} + \frac{4}{50} = \frac{11}{50} \]
What about when they are not *Mutually exclusive*?
OR without Mutually Exclusive Events

\[ P(E \cup F) = P(E) + P(F) - P(EF) \]

Piech, CS109, Stanford University
OR *without* Mutually Exclusive Events

\[ P(E \cup F) = \frac{8}{50} + \frac{14}{50} - \frac{3}{50} = \frac{19}{50} \]
More than two sets?
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) + P(F) + P(G) . \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(EF) \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(EF) - P(EG) \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) + P(F) + P(G) \]
\[ \quad - P(EF) - P(EG) - P(FG) \]
Inclusion / Exclusion with Three Events

\[ P(E \cup F \cup G) = P(E) + P(F) + P(G) \]
\[ -P(EF) - P(EG) - P(FG) \]
\[ +P(EFG) \]
General Inclusion / Exclusion

\[ P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{r=1}^{n} (-1)^{r+1} Y_r \]

\[ Y_1 = \text{Sum of all events on their own} \]
\[ \sum_{i} P(E_i) \]

\[ Y_2 = \text{Sum of all pairs of events} \]
\[ \sum_{i,j \text{ s.t. } i \neq j} P(E_i \cap E_j) \]

\[ Y_3 = \text{Sum of all triples of events} \]
\[ \sum_{i,j,k \text{ s.t. } i \neq j, j \neq k, i \neq k} P(E_i \cap E_j \cap E_k) \]

* Where \( Y_r \) is the sum, for all combinations of \( r \) events, of the probability of the union those events.
Today

- OR
  - P(E ∪ F)
    - Mutually Exclusive?
      - Just Add!
      - Inclusion Exclusion
    - DeMorgan's
  
- AND
  - P(EF)
    - Independent?
      - Just Multiply
      - Chain Rule

- DeMorgan’s
  - "AND"
  - "OR"
Today

P(E ∪ F) = P(E) + P(F) − P(EF)

P(EF) = P(E|F)P(F)

DeMorgan’s

Mutually Exclusive?

Inclusion Exclusion

Just Add!

Just Multiply

Independent?

Chain Rule
Probability of “AND”
Independence

Two events $A$ and $B$ are called **independent** if:

$$P(A) = P(A|B)$$

Knowing that event $B$ happened, doesn't change our belief that $A$ will happen.

Otherwise, they are called **dependent** events.
Independence is reciprocal

If A is independent of B, then B is independent of A

\[ P(A) = P(A|B) \quad \quad P(B|A) = P(B) \]

Proof:

\[
P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \quad \text{Bayes’ Thm.}
\]

\[
= \frac{P(A)P(B)}{P(A)}
\]

\[
= P(B)
\]

Because A is independent of B
Alternative Definition of Independence

\[ P(A, B) = P(A) \cdot P(B|A) \]

\[ = P(A) \cdot P(B) \]

Probability of and
Since B is independent of A

If you show this is true, you have proved the two events are independent!
If events are *independent* probability of AND is easy!

*You will need to use this “trick” with high probability.

Piech, CS109, Stanford University
Dice, our misunderstood friends

Roll two 6-sided dice, yielding values $D_1$ and $D_2$
- Let $E$ be event: $D_1 = 1$
- Let $F$ be event: $D_2 = 1$

What is $P(E)$, $P(F)$, and $P(EF)$?
- $P(E) = 1/6$, $P(F) = 1/6$, $P(EF) = 1/36$
- $P(EF) = P(E)P(F) \rightarrow E$ and $F$ independent

Let $G$ be event: $D_1 + D_2 = 5 \{ (1, 4), (2, 3), (3, 2), (4, 1) \}$
What is $P(E)$, $P(G)$, and $P(EG)$?
- $P(E) = 1/6$, $P(G) = 4/36 = 1/9$, $P(EG) = 1/36$
- $P(EG) \neq P(E)P(G) \rightarrow E$ and $G$ dependent
What does independence look like?
Independence

Independence Definition 1:

\[ P(AB) = P(A)P(B) \]

\[ \frac{|AB|}{|S|} = \frac{|A|}{|S|} \times \frac{|B|}{|S|} \]
Independence

Independence Definition 1:

\[ P(AB) = P(A)P(B) \]
\[
\frac{|AB|}{|S|} = \frac{|A|}{|S|} \times \frac{|B|}{|S|}
\]

Independence Definition 2:

\[ P(A|B) = P(A) \]
\[
\frac{|AB|}{|B|} = \frac{|A|}{|S|}
\]
Independence

This ratio, $P(A)$... ... is the same as this one, $P(A|B)$
Independence

Independence Definition 1:

\[ P(AB) = P(A)P(B) \]
\[ \frac{|AB|}{|S|} = \frac{|A|}{|S|} \times \frac{|B|}{|S|} \]

Independence Definition 2:

\[ P(A|B) = P(A) \]
\[ \frac{|AB|}{|B|} = \frac{|A|}{|S|} \]
Independence Definition 1:

\[ P(AB) = P(A)P(B) \]

\[ \frac{|AB|}{|S|} = \frac{|A|}{|S|} \times \frac{|B|}{|S|} \]

Independence Definition 2:

\[ P(A|B) = P(A) \]

\[ \frac{|AB|}{|B|} = \frac{|A|}{|S|} \]
More Intuition through proofs:
Independence

Given independent events $A$ and $B$, prove that $A$ and $B^C$ are independent

We want to show that $P(AB^C) = P(A)P(B^C)$

$$P(AB^C) = P(A) - P(AB)$$  \hspace{1cm} \text{By Total Law of Prob.}

$$= P(A) - P(A)P(B)$$  \hspace{1cm} \text{By independence}

$$= P(A)[1 - P(B)]$$  \hspace{1cm} \text{Factoring}

$$= P(A)P(B^C)$$  \hspace{1cm} \text{Since } P(B) + P(B^C) = 1

So if $A$ and $B$ are independent $A$ and $B^C$ are also independent

Piech, CS109, Stanford University
Generalization
Generalized Independence

General definition of Independence:
Events $E_1$, $E_2$, ..., $E_n$ are independent if for every subset with $r$ elements (where $r \leq n$) it holds that:

$$P(E_1', E_2', E_3', ..., E_r') = P(E_1')P(E_2')P(E_3')...P(E_r')$$

Example: outcomes of $n$ separate flips of a coin are all independent of one another
- Each flip in this case is called a “trial” of the experiment
Math > Intuition
Roll two 6-sided dice, yielding values $D_1$ and $D_2$

- Let $E$ be event: $D_1 = 1$
- Let $F$ be event: $D_2 = 6$
- Are $E$ and $F$ independent? Yes!

Let $G$ be event: $D_1 + D_2 = 7$

- Are $E$ and $G$ independent? Yes!
- $P(E) = 1/6$, $P(G) = 1/6$, $P(E \cap G) = 1/36$ [roll (1, 6)]
- Are $F$ and $G$ independent? Yes!
- $P(F) = 1/6$, $P(G) = 1/6$, $P(F \cap G) = 1/36$ [roll (1, 6)]
- Are $E$, $F$ and $G$ independent? No!
- $P(E \cap F \cap G) = 1/36 \neq 1/216 = (1/6)(1/6)(1/6)$
New Ability
Properties of Pairs of Events

Mutually Exclusive

\[ P(A \text{ and } B) = 0 \]

also:

\[ P(A \text{ or } B) = P(A) + P(B) \]

Independent

\[ P(A) = P(A|B) \]

also:

\[ P(A \text{ and } B) = P(A) \cdot P(B) \]
To 

day

AND

OR

P(E ∪ F)

P(E \cap F)

Just Add!

Mutually Exclusive?

Inclusion Exclusion

Independent?

Just Multiply!

Chain Rule

DeMorgan’s
Think of the children as independent trials

Two parents both have an (A, a) gene pair.

- Each parent will pass on one of their genes (each gene equally likely) to a child.
- The probability of any single child having curly hair (the recessive trait) is 0.25, independent of other siblings.
- There are three children.

What is the probability that all three children have curly hair?

Let \( E_1, E_2, E_3 \) be the events that child 1, 2, and 3 have curly hair, respectively.

\[
P(E_1E_2E_3) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2) = P(E_1)P(E_2)P(E_3)
\]
Independence

Two events $E$ and $F$ are defined as independent if:

$$P(EF) = P(E)P(F)$$

For independent events $E$ and $F$,

- $P(E|F) = P(E)$
- $E$ and $F^C$ are independent.
Independence of complements

Statement:

If $E$ and $F$ are independent, then $E$ and $F^C$ are independent.

Proof:

\[ P(EF^C) = P(E) - P(EF) \]
\[ = P(E) - P(E)P(F) \text{  Intersection} \]
\[ = P(E)[1 - P(F)] \text{  Independence of } E \text{ and } F \]
\[ = P(E)P(F^C) \text{  Factoring} \]
\[ = P(E)P(F^C) \text{  Complement} \]

$E$ and $F^C$ are independent

Knowing whether $F$ does or doesn’t happen doesn’t change our belief about $E$ happening.
Consider the following parallel network:

- \( n \) independent routers, each with probability \( p_i \) of functioning (where \( 1 \leq i \leq n \))
- \( E = \) functional path from A to B exists.

What is \( P(E) \)?
Consider the following parallel network:

- \( n \) independent routers, each with probability \( p_i \) of functioning (where \( 1 \leq i \leq n \))
- \( E = \) functional path from A to B exists.

What is \( P(E) \)?

\[
P(E) = P(\geq 1 \text{ one router works})
\]

\[
= 1 - P(\text{all routers fail})
\]

\[
= 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_n)
\]

\[
= 1 - \prod_{i=1}^{n} (1 - p_i)
\]

\( \geq 1 \) with independent trials: take complement
The Most Important Core Probability Question

Say a coin comes up heads with probability $p$
- Flip the coin $n$ times
- Each coin flip is an *independent* trial
- What is the probability of exactly $k$ heads?
Say a coin comes up heads with probability $p$.

Flip the coin $n$ times.

Each coin flip is an independent trial.

What is the probability of exactly $k$ heads?

The Most Important Core Probability Question

Piech, CS109, Stanford University
Pedagogical Pause

**DeMorgan's**

**OR**
P(E ∪ F)

**AND**
P(EF)

**Mutually Exclusive?**

**Just Add!**

**Inclusion Exclusion**

**Independent?**

**Just Multiply!**

**Chain Rule**
Sets Review
Sets Review

Say $E$ and $F$ are events in $S$

- $S = \{1, 2, 3, 4, 5, 6\}$ die roll outcome
- $E = \{1, 2\}$
- $F = \{2, 3\}$
- $E \cup F = \{1, 2, 3\}$
Sets Review

Say E and F are events in S

Event that is in E and F

\[ E \cap F \]
Sets Review

Say $E$ and $F$ are events in $S$

Event that is not in $E$ (called complement of $E$)

$E^c \text{ or } \sim E$
Sets Review

Say E and F are subsets of S

Which of these two is it?

a) \((E \text{ or } F)^C\)  
b) \((E^C \text{ and } F^C)\)
Sets Review

Say $E$ and $F$ are subsets of $S$

Which of these two is it?

a) $(E \text{ and } F)^C$

b) $(E^C \text{ or } F^C)$
De Morgan’s Laws

De Morgan’s Law lets you alternate between AND and OR.

\[(E \cap F)^C = E^C \cup F^C\]

In probability:
\[P(E_1 E_2 \cdots E_n)\]
\[= 1 - P\left((E_1 E_2 \cdots E_n)^C\right)\]
\[= 1 - P\left(E_1^C \cup E_2^C \cup \cdots \cup E_n^C\right)\]
Great if \(E_i^C\) mutually exclusive!

\[(E \cup F)^C = E^C \cap F^C\]

In probability:
\[P(E_1 \cup E_2 \cup \cdots \cup E_n)\]
\[= 1 - P\left((E_1 \cup E_2 \cup \cdots \cup E_n)^C\right)\]
\[= 1 - P\left(E_1^C E_2^C \cdots E_n^C\right)\]
Great if \(E_i\) independent!
Augustin Demorgan

- British Mathematician who wrote the book “Formal Logic” in 1847
- Celebrity lookalike is Jason Alexander from Seinfeld.
Hash Tables. Hardest Core Probability Question

- **Keys**: John Smith, Lisa Smith, Sandra Dee
- **Hash Function**: Maps keys to buckets
- **Buckets**:
  - 00
  - 01: 521-8976
  - 02: 521-1234
  - 03
  - 13
  - 14: 521-9655
  - 15
Hash table fun

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is independent with probability $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E =$ bucket 1 has $\geq 1$ string hashed into it?

2. $E =$ at least 1 of buckets 1 to $k$ has $\geq 1$ string hashed into it?
Hash table fun

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E = \text{bucket 1 has } \geq 1 \text{ string hashed into it}$?

Define:

$S_i = \text{string } i \text{ hashes to bucket 1}$

$S_i^C = \text{string } i \text{ doesn’t hash to bucket 1}$

$P(S_i) = p_1$

$P(S_i^C) = 1 - p_1$
Hash table fun

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E =$ bucket 1 has $\geq 1$ string hashed into it?

**WTF** (not-real acronym for Want To Find):

\[
P(E) = P(S_1 \cup S_2 \cup \cdots \cup S_m) \\
= 1 - P\left((S_1 \cup S_2 \cup \cdots \cup S_m)^C\right) \\
= 1 - P\left(S_1^C S_2^C \cdots S_m^C\right) \\
= 1 - P\left(S_1^C\right)P\left(S_2^C\right)\cdots P\left(S_m^C\right) \\
= 1 - \left(P\left(S_1^C\right)\right)^m \\
= 1 - (1 - p_1)^m
\]

Define: $S_i =$ string $i$ hashes to bucket 1
\[S_i^C = \text{string } i \text{ doesn’t hash to bucket } 1\]

Complement

De Morgan’s Law

\[
P(S_i) = p_1 \\
P(S_i^C) = 1 - p_1
\]

$S_i$ independent trials
More hash table fun: Possible approach?

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E = \text{bucket 1 has } \geq 1 \text{ string hashed into it?}$
2. $E = \text{at least 1 of buckets 1 to } k \text{ has } \geq 1 \text{ string hashed into it?}$

$$P(E) =$$
More hash table fun: Possible approach?

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E = \text{bucket 1 has } \geq 1 \text{ string hashed into it}$?
2. $E = \text{at least 1 of buckets 1 to } k \text{ has } \geq 1 \text{ string hashed into it}$?

\[
P(E) = \text{Define } F_i = \text{bucket } i \text{ has at least one string in it}
\]
More hash table fun: Possible approach?

• $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
• Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if
1. $E = \text{bucket 1 has } \geq 1 \text{ string hashed into it}$?
2. $E = \text{at least 1 of buckets 1 to } k \text{ has } \geq 1 \text{ string hashed into it}$?

$$P(E) = P(F_1 \cup F_2 \cup \ldots \cup F_k)$$

Define $F_i = \text{bucket } i \text{ has at least one string in it}$

⚠️ $F_i$ bucket events are dependent! So we cannot just add.
More hash table fun: Possible approach?

- \( m \) strings are hashed (not uniformly) into a hash table with \( n \) buckets.
- Each string hash is an independent trial w.p. \( p_i \) of getting hashed into bucket \( i \).

What is \( P(E) \) if

1. \( E = \) bucket 1 has \( \geq 1 \) string hashed into it?
2. \( E = \) at least 1 of buckets 1 to \( k \) has \( \geq 1 \) string hashed into it?

\[
P(E) = P(F_1 \cup F_2 \cup \ldots \cup F_k)
= 1 - P\left((F_1 \cup F_2 \cup \ldots \cup F_k)^C\right)
= 1 - P\left(F_1^C F_2^C \ldots F_k^C\right)
\]

Define \( F_i = \) bucket \( i \) has at least one string in it

\( F_i \) bucket events are dependent! So we cannot just add.
More hash table fun: Possible approach?

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E$ = bucket 1 has $\geq 1$ string hashed into it?
2. $E$ = at least 1 of buckets 1 to $k$ has $\geq 1$ string hashed into it?

$$P(E) = P(F_1 \cup F_2 \cup \cdots \cup F_k)$$
$$= 1 - P\left( (F_1 \cup F_2 \cup \cdots \cup F_k)^c \right)$$
$$= 1 - P\left( F_1^c F_2^c \cdots F_k^c \right)$$

Define $F_i$ = bucket $i$ has at least one string in it

$= P($buckets 1 to $k$ all denied strings$)$
$= (P($each string hashes to $k + 1$ or higher $))$
$= (1 - p_1 - p_2 - \cdots - p_k)^m$

$F_i$ bucket events are dependent! So we cannot just add.

Piech, CS109, Stanford University
More hash table fun: Possible approach?

- \( m \) strings are hashed (not uniformly) into a hash table with \( n \) buckets.
- Each string hash is an independent trial w.p. \( p_i \) of getting hashed into bucket \( i \).

What is \( P(E) \) if

1. \( E = \) bucket 1 has \( \geq 1 \) string hashed into it?
2. \( E = \) at least 1 of buckets 1 to \( k \) has \( \geq 1 \) string hashed into it?

\[
P(E) = P(F_1 \cup F_2 \cup \ldots \cup F_k)
= 1 - P((F_1 \cup F_2 \cup \ldots \cup F_k)^C)
= 1 - P(F_1^C F_2^C \ldots F_k^C)
= 1 - (1 - p_1 - p_2 \ldots - p_k)^m
\]

Define \( F_i = \) bucket \( i \) has at least one string in it.

\[
= P(\text{buckets 1 to } k \text{ all denied strings})
= (P(\text{each string hashes to } k + 1 \text{ or higher}))
= (1 - p_1 - p_2 \ldots - p_k)^m
\]

\( F_i \) bucket events are dependent! So we cannot just add.

Piech, CS109, Stanford University
The fun never stops with hash tables

- $m$ strings are hashed (not uniformly) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if
1. $E = \text{bucket } 1 \text{ has } \geq 1 \text{ string hashed into it}$? ✔
2. $E = \text{at least } 1 \text{ of buckets } 1 \text{ to } k \text{ has } \geq 1 \text{ string hashed into it}$? ✔

Looking for a challenge? 😊
The fun never stops with hash tables

- $m$ strings are hashed (unequally) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E =$ bucket 1 has $\geq 1$ string hashed into it?
2. $E =$ at least 1 of buckets 1 to $k$ has $\geq 1$ string hashed into it?
3. $E =$ each of buckets 1 to $k$ has $\geq 1$ string hashed into it?
THIS IS FINE.
The fun never stops with hash tables

- $m$ strings are hashed (unequally) into a hash table with $n$ buckets.
- Each string hash is an independent trial w.p. $p_i$ of getting hashed into bucket $i$.

What is $P(E)$ if

1. $E =$ bucket 1 has $\geq 1$ string hashed into it?
2. $E =$ at least 1 of buckets 1 to $k$ has $\geq 1$ string hashed into it?
3. $E =$ each of buckets 1 to $k$ has $\geq 1$ string hashed into it?

Hint: Use Part 2’s event definition:

Define $F_i =$ bucket $i$ has at least one string in it

Hint: Try $k = 2$, then $k = 3$, then generalize.
The fun never stops with hash tables

Solution

- $F_i = \text{at least one string hashed into } i\text{-th bucket}$
- $P(E) = P(F_1 F_2 \ldots F_k) = 1 - P((F_1 F_2 \ldots F_k)^c)$
  $$= 1 - P(F_1^c \cup F_2^c \cup \ldots \cup F_k^c)$$ (DeMorgan’s Law)

where

$$P\left( \bigcup_{i=1}^{k} F_i^c \right) = 1 - \sum_{r=1}^{k} (-1)^{(r+1)} \sum_{i_1 < \ldots < i_r} P(F_{i_1}^c F_{i_2}^c \ldots F_{i_r}^c)$$

$$P(F_{i_1}^c F_{i_2}^c \ldots F_{i_r}^c) = (1 - p_{i_1} - p_{i_2} - \ldots - p_{i_r})^m$$
Here we are

Source: The Hobbit
100,000 samples

6 observations per sample

Piech, CS109, Stanford University
Discovered Pattern

Piech-2: dna piech$ python findStructure.py
size data = 100000
p(G1) = 0.500
p(G2) = 0.545
p(G3) = 0.299
p(G4) = 0.701
p(G5) = 0.600
p(T) = 0.390
p(T and G1) = 0.291, p(T)p(G1) = 0.195
p(T and G2) = 0.300, p(T)p(G2) = 0.213
p(T and G3) = 0.116, p(T)p(G3) = 0.117
p(T and G4) = 0.273, p(T)p(G4) = 0.273
p(T and G5) = 0.309, p(T)p(G5) = 0.234

...
Discovered Pattern

```
$ python findStructure.py
size data = 100000
p(G1) = 0.500
p(G2) = 0.545
p(G3) = 0.299
p(G4) = 0.701
p(G5) = 0.600
p(T) = 0.390
p(T and G1) = 0.291, P(T)p(G1) = 0.195
p(T and G2) = 0.300, P(T)p(G2) = 0.213
p(T and G3) = 0.116, P(T)p(G3) = 0.117
**p(T and G4) = 0.273, P(T)p(G4) = 0.273**
p(T and G5) = 0.309, P(T)p(G5) = 0.234
```

...  

\[
p(T \text{ and } G5 \mid G2) = 0.450
\]
\[
p(T \mid G2)p(G5 \mid G2) = 0.450
\]
Discovered Pattern

\[
\begin{align*}
\text{Piech-2: dna piech}\$ \text{ python findStructure.py} \\
\text{size data = 100000} \\
p(G1) &= 0.500 \\
p(G2) &= 0.545 \\
p(G3) &= 0.299 \\
p(G4) &= 0.701 \\
p(G5) &= 0.600 \\
p(T) &= 0.390 \\
p(T \text{ and } G1) &= 0.291, \ P(T)p(G1) = 0.195 \\
p(T \text{ and } G2) &= 0.300, \ P(T)p(G2) = 0.213 \\
p(T \text{ and } G3) &= 0.116, \ P(T)p(G3) = 0.117 \\
p(T \text{ and } G4) &= 0.273, \ P(T)p(G4) = 0.273 \\
p(T \text{ and } G5) &= 0.309, \ P(T)p(G5) = 0.234 \\
\end{align*}
\]

\[
\begin{align*}
p(T \text{ and } G5 \mid G2) &= 0.450 \\
p(T \mid G2)p(G5 \mid G2) &= 0.450
\end{align*}
\]
Discovered Pattern

Piech-2: dna piech$ python findStructure.py
size data = 100000
p(G1) = 0.500
p(G2) = 0.545
p(G3) = 0.299
p(G4) = 0.701
p(G5) = 0.600
p(T) = 0.390
p(T and G1) = 0.291 , P(T)p(G1) = 0.195
p(T and G2) = 0.309 , P(T)p(G2) = 0.213
p(T and G3) = 0.116 , P(T)p(G3) = 0.117
p(T and G4) = 0.273 , P(T)p(G4) = 0.273
p(T and G5) = 0.309 , P(T)p(G5) = 0.234

\[ p(T \text{ and } G5 \mid G2) = 0.450 \]
\[ p(T \mid G2)p(G5 \mid G2) = 0.450 \]

Piech, CS109, Stanford University
Discovered Pattern

```bash
Piech-2: dna piech$ python findStructure.py
size data = 100000
p(G1) = 0.500
p(G2) = 0.545
p(G3) = 0.299
p(G4) = 0.701
p(G5) = 0.600
p(T) = 0.390
p(T and G1) = 0.291, P(T)p(G1) = 0.195
p(T and G2) = 0.300, P(T)p(G2) = 0.213
p(T and G3) = 0.116, P(T)p(G3) = 0.117
p(T and G4) = 0.273, P(T)p(G4) = 0.273
p(T and G5) = 0.309, P(T)p(G5) = 0.234

... 
```

\[ p(T \text{ and } G5 \mid G2) = 0.450 \]
\[ p(T \mid G2)p(G5 \mid G2) = 0.450 \]
These genes don’t impact T

Only Causal Structure That Fits

\[ p(T \mid G_1 \text{ and } G_2) = 0.9 \]

\[ p(T \mid \sim G_1 \text{ or } \sim G_2) = 0.2 \]

\[ p(G_1) = 0.5 \]

\[ p(G_2 \mid G_5) = 0.9 \]

\[ p(G_2 \mid \sim G_5) = 0.2 \]

\[ p(G_5) = 0.6 \]

Piech, CS109, Stanford University