Course Overview

Statistics for Data Science
CSE357 - Fall 2021
Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts

(Stanford Encyclopedia of Philosophy, 2014)
Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts

(Stanford Encyclopedia of Philosophy, 2014)

Data Science - a field focused on using statistical, scientific, and computational techniques to gain insights from data.
Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts

(Stanford Encyclopedia of Philosophy, 2014)

Data Science - a field focused on using statistical, scientific, and computational techniques to gain insights from data.
Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts

(Stanford Encyclopedia of Philosophy, 2014)

Data Science - a field focused on using statistical, scientific, and computational techniques to gain insights from data.
Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts

(Stanford Encyclopedia of Philosophy, 2014)

Data Science - a field focused on using statistical, scientific, and computational techniques to gain insights from data.

Approximately equal:

Data Science ≈ Data Mining ≈ Analytics ≈ Quantitative Science

Highly Related

Data Science, Big Data, Machine Learning, Artificial Intelligence
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)

A pathway to knowledge about...

... what was, (past)

... what is, (present)

... what is likely (future)
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)

Why?!?

A pathway to knowledge about...

... what was,  (past)
... what is,   (present)
... what is likely (future, the full population)
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)

\textit{Why?!?}

A pathway to knowledge about...

... what was, (past)

... what is, (present)

... what is likely (future)

Jobs
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)

**Why?!?**

A pathway to knowledge about...

... what was, (past)

... what is, (present)

... what is likely (future)

**Jobs**

**Decisions**
Statistics for Data Science

Statistical methods for gaining knowledge and insights from data.

-- designed for those already proficient in programming (i.e. computing)

Why?!?

A pathway to knowledge about...

... what was,  (past)  
... what is,    (present)  
... what is likely (future)

Jobs

Decisions

Truth / Meaning in Life

The answer to the "ultimate question of life, the universe, and everything" (Adams)
In other words, so you can go on Twitter and say

"The data say …"

"I did my research."

… and change no one's mind but at least understand it better yourself.
Course Website

https://www3.cs.stonybrook.edu/~has/CSE357/index.html
What is Probability?
What is Probability?

Examples

(1) outcome of flipping a coin

(2) amount of snowfall

(3) mentioning "happy"

(4) mentioning "happy" *a lot*
What is Probability?

The chance that something will happen.

Given infinite observations of an event, the proportion of observations where a given outcome happens.

Strength of belief that something is true.
What is Probability?

The chance that something will happen.

Given infinite observations of an event, the proportion of observations where a given outcome happens.

Strength of belief that something is true.

“Mathematical language for quantifying uncertainty” - Wasserman
Probability (review)

$\Omega$ : Sample Space, set of all outcomes of a random experiment

$A$ : Event ($A \subseteq \Omega$), collection of possible outcomes of an experiment

$P(A)$: Probability of event $A$, $P$ is a function: events $\rightarrow \mathbb{R}$
Probability (review)

$\Omega$: Sample Space, set of all outcomes of a random experiment

$A$: Event ($A \subseteq \Omega$), collection of possible outcomes of an experiment

$P(A)$: Probability of event $A$, $P$ is a function: events $\rightarrow \mathbb{R}$

(1) $P(\Omega) = 1$

(2) $P(A) \geq 0$, for all $A$

(3) If $A_1, A_2, \ldots$ are disjoint events then:

\[
P(\bigcup_{i} A_i) = \sum_{i} P(A_i)
\]
Probability (review)

Ω: Sample Space, set of all outcomes of a random experiment

A: Event (A ⊆ Ω), collection of possible outcomes of an experiment

P(A): Probability of event A, P is a function: events → ℝ

P is a probability measure, if and only if

1. \( P(Ω) = 1 \)
2. \( P(A) \geq 0 \), for all A
3. If \( A_1, A_2, \ldots \) are disjoint events then:
   \[
P(\bigcup_{i} A_i) = \sum_{i} P(A_i)
   \]
Probability (review)

\( \Omega \): Sample Space, set of all outcomes of a random experiment

\( A \): Event \((A \subseteq \Omega)\), collection of possible outcomes of an experiment

\( P(A) \): Probability of event \( A \), \( P \) is a function: \( \text{events} \rightarrow \mathbb{R} \)

\( P \) is a *probability measure*, if and only if

1. \( P(\Omega) = 1 \)
2. \( P(A) \geq 0 \), for all \( A \)
3. If \( A_1, A_2, \ldots \) are disjoint events then:

\[
P(\bigcup_{i} A_i) = \sum_{i} P(A_i)
\]
Probability (review)

\( \Omega \): Sample Space, set of all outcomes of a random experiment

\( A \): Event (\( A \subseteq \Omega \)), collection of possible outcomes of an experiment

\( P(A) \): Probability of event \( A \), \( P \) is a function: \( \text{events} \rightarrow \mathbb{R} \)

\( P \) is a **probability measure**, if and only if

1. \( P(\Omega) = 1 \)
2. \( P(A) \geq 0 \), for all \( A \)
3. If \( A_1, A_2, \ldots \) are disjoint events then:
   \[
P(\bigcup_{i} A_i) = \sum_{i} P(A_i)
   \]

**Examples**

1. outcome of flipping a coin
2. amount of snowfall
3. mentioning "happy"
4. mentioning "happy" a lot
Probability (review)

Some Properties:

If \( B \subseteq A \) then \( P(A) \geq P(B) \)

\[ P(A \cup B) \leq P(A) + P(B) \]

\[ P(A \cap B) \leq \min(P(A), P(B)) \]

\[ P(\neg A) = P(\Omega / A) = 1 - P(A) \]

/ is set difference

\( P(A \cap B) \) will be notated as \( P(A, B) \)

Examples

1. outcome of flipping a coin
2. amount of snowfall
3. mentioning "happy"
4. mentioning "happy" a lot
Independence

Two Events: $A$ and $B$

Does knowing something about $A$ tell us whether $B$ happens (and vice versa)?
Independence

Two Events: $A$ and $B$

Does knowing something about $A$ tell us whether $B$ happens (and vice versa)?

(1) A: first flip of a fair coin; B: second flip of the same fair coin
(2) A: mention or not of the first word is “happy”
    B: mention or not of the second word is “birthday”
Independence

Two Events: $A$ and $B$

Does knowing something about $A$ tell us whether $B$ happens (and vice versa)?

(1) $A$: first flip of a fair coin; $B$: second flip of the same fair coin
(2) $A$: mention or not of the first word is “happy”
    $B$: mention or not of the second word is “birthday”

Two events, $A$ and $B$, are independent iff $P(A, B) = P(A)P(B)$
Independence

Two Events: A and B

Does knowing something about A tell us whether B happens (and vice versa)?

(1) A: first flip of a fair coin; B: second flip of the same fair coin

(2) A: mention or not of the first word is “happy”
    B: mention or not of the second word is “birthday”

Two events, A and B, are independent iff \( P(A, B) = P(A)P(B) \)
Disjoint Sets vs. Independent Events

**Independence:** Two events, A and B are independent iff $P(A,B) = P(A)P(B)$

**Disjoint Sets:** If two events, A and B, come from disjoint sets, then $P(A,B) = 0$
Disjoint Sets vs. Independent Events

**Independence:** … iff \( P(A,B) = P(A)P(B) \)

**Disjoint Sets:** If two events, \( A \) and \( B \), come from disjoint sets, then \( P(A,B) = 0 \)

Does *independence* imply *disjoint*?
Disjoint Sets vs. Independent Events

**Independence:** … iff \( P(A,B) = P(A)P(B) \)

**Disjoint Sets:** If two events, \( A \) and \( B \), come from disjoint sets, then
\[
P(A,B) = 0
\]

Does independence imply disjoint? No
Proof: A counterexample: ?
Disjoint Sets vs. Independent Events

**Independence:** … iff \( P(A,B) = P(A)P(B) \)

**Disjoint Sets:** If two events, \( A \) and \( B \), come from disjoint sets, then
\[
P(A,B) = 0
\]

Does independence imply disjoint? No

Proof: A counterexample: \( A \): flip of fair coin \( A \) is heads,
\( B \): flip of fair coin \( B \) is heads;

independence tell us \( P(A)P(B) = P(A,B) = 0.25 \),
but disjoint tells us \( P(A, B) = 0 \)
Probability (Review)

Conditional Probability

\[ P(A, B) \]
\[ P(A | B) = \frac{---------}{P(B)} \]
Probability (Review)

Conditional Probability

\[ P(A, B) \]

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

H: mention “happy” in message, \( m \)
B: mention “birthday” in message, \( m \)

P(H) = .01 \quad P(B) = .001 \quad P(H, B) = .0005

P(H|B) = ??
Probability (Review)

Conditional Probability

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

H: mention “happy” in message, m
B: mention “birthday” in message, m

\[P(H) = .01 \quad P(B) = .001 \quad P(H, B) = .0005\]
\[P(H \mid B) = .50\]

H1: first flip of a fair coin is heads
H2: second flip of the same coin is heads

\[P(H2) = 0.5 \quad P(H1) = 0.5 \quad P(H2, H1) = 0.25\]
\[P(H2 \mid H1) = 0.5\]
Probability (Review)

Conditional Probability

\[ P(A, B) \]

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

Two events, A and B, are independent iff

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

H1: first flip of a fair coin is heads
H2: second flip of the same coin is heads

P(H2) = 0.5 \quad P(H1) = 0.5 \quad P(H2, H1) = 0.25

P(H2|H1) = 0.5

P(A, B) = P(A)P(B) \quad \text{iff} \quad P(A|B) = P(A)
# Probability (Review)

## Conditional Probability

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

Two events, A and B, are independent iff

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

\[
P(A, B) = P(A)P(B) \text{ iff } P(A \mid B) = P(A)
\]

**Interpretation of Independence:**

Observing B has no effect on probability of A.

H1: first flip of a fair coin is heads

H2: second flip of the same coin is heads

\[
P(H2) = 0.5 \quad P(H1) = 0.5 \quad P(H2, H1) = 0.25
\]

\[
P(H2 \mid H1) = 0.5
\]

\[
P(H2|H1) = 0.5
\]
Why Probability?
Why Probability?

A formality to make sense of the world.

1) To quantify uncertainty
   *Should we believe something or not? Is it a meaningful difference?*

2) To be able to generalize from one situation or point in time to another.
   *Can we rely on some information? What is the chance Y happens?*

3) To organize data into meaningful groups or “dimensions”
   *Where does X belong? What words are similar to X?*
Probabilities over >2 events...

Independence:

\( A_1', A_2', \ldots, A_n \) are independent iff

\[
P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i)
\]
Probabilities over >2 events...

Independence:

\[ A_1', A_2', \ldots, A_n \] are independent iff 

\[
P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i)
\]

Conditional Probability:

\[
P(A_1, A_2, \ldots, A_{n-1} | A_n) = \frac{P(A_1, A_2, \ldots, A_{n-1}, A_n)}{P(A_n)}
\]
Probabilities over $>2$ events...

Independence:

$A_1, A_2, \ldots, A_n$ are independent iff

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i)$$

Conditional Probability:

$$P(A_1, A_2, \ldots, A_{n-1}|A_n) = \frac{P(A_1, A_2, \ldots, A_{n-1}, A_n)}{P(A_n)}$$

$$P(A_1, A_2, \ldots, A_{m-1}|A_m, A_{m+1}, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_{m-1}, A_m, A_{m+1}, \ldots, A_n)}{P(A_n)}$$

just think of multiple events happening as a single event:

$$Z = A_1, A_2, \ldots, A_{m-1} = A_1 \cap A_2 \cap \ldots \cap A_{m-1} \quad \text{then} \quad P(Z|A_n)$$
Conditional Probabilities are **Fundamental** to Data Science

for example

**Machine Learning:** Most modern deep learning techniques try to estimate

\[ P(\text{outcome} \mid \text{data}) \]

**Causal inference:** Does treatment cause outcome?

\[ P(\text{outcome} \mid \text{treatment}) \neq P(\text{outcome}) \]

*also requires random sampling of treatment conditions*
Conditional Independence

$A$ and $B$ are conditionally independent, given $C$, IFF

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

Equivalently, $P(A \mid B, C) = P(A \mid C)$

Interpretation: *Once we know $C$, then $B$ doesn’t tell us anything useful about $A$.*
Bayes Theorem - Lite

GOAL: Relate \((1) \ P(A|B)\) to \((2) \ P(B|A)\)
Bayes Theorem - Lite

GOAL: Relate (1) \( P(A|B) \) to (2) \( P(B|A) \)

Let’s try:

(3) \( P(A|B) = \frac{P(A,B)}{P(B)}, \) def. of conditional probability on (1)
Bayes Theorem - Lite

GOAL: Relate (1) $P(A|B)$ to (2) $P(B|A)$

Let’s try:

(3) $P(A|B) = P(A,B) / P(B)$, def. of conditional probability on (1)

(4) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of cond prob on (2); sym of set intrsct
Bayes Theorem - Lite

GOAL: Relate (1) $P(A|B)$ to (2) $P(B|A)$

Let’s try:

(3) $P(A|B) = P(A,B)/P(B)$, def. of conditional probability on (1)

(4) $P(B|A) = P(B,A)/P(A) = P(A,B)/P(A)$, def. of cond prob on (2); sym of set intrsct

(5) $P(B|A)P(A) = P(A,B)$, algebra on (4) ← known as “Multiplication Rule”
Bayes Theorem - Lite

GOAL: Relate \( P(A|B) \) to \( P(B|A) \)

Let’s try:

(3) \( P(A|B) = \frac{P(A,B)}{P(B)} \), def. of conditional probability on (1)

(4) \( P(B|A) = \frac{P(B,A)}{P(A)} = \frac{P(A,B)}{P(A)} \), def. of cond prob on (2); sym of set intersec

(5) \( P(B|A)P(A) = P(A,B) \), algebra on (4) ← known as “Multiplication Rule”

(6) \( P(A|B) = \frac{(P(B|A)P(A))}{P(B)} \), Substitute \( P(A,B) \) from (5) into (3)
Bayes Theorem - Lite

GOAL: Relate (1) \( P(A|B) \) to (2) \( P(B|A) \)

Let’s try:

(3) \( P(A|B) = \frac{P(A,B)}{P(B)} \), def. of conditional probability on (1)

(4) \( P(B|A) = \frac{P(B,A)}{P(A)} = \frac{P(A,B)}{P(A)} \), def. of cond prob on (2); sym of set intrsct

(5) \( P(B|A)P(A) = P(A,B) \), algebra on (4) ← known as “Multiplication Rule”

(6) \( P(A|B) = \frac{(P(B|A)P(A))}{P(B)} \), Substitute \( P(A,B) \) from (5) into (3)
Bayes Theorem - Lite

Why?

We often want to know $P(A|B)$ but we are only given $P(B|A)$ and $P(A)$.

Example: You want to know if an email is likely spam given a word appearing in it: $P(\text{spam} \mid \text{word})$. However, you only have a dataset of words and spam: $P(\text{word} \mid \text{spam})$ and you can look up the frequency of spam emails in general to get $P(\text{spam})$ as well as the frequency of "word" in general for $P(\text{word})$. 

(6) $P(A \mid B) = \frac{(P(B \mid A)P(A))}{P(B)}$
Bayes Theorem - Heavy (with multiple events partitioning Ω)

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$,
for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $Ω$
First: Law of Total Probability

GOAL: Relate \( P(A_i | B) \) to \( P(B | A_i) \),
for all \( i = 1 \ldots k \), where \( A_1 \ldots A_k \) partition \( \Omega \)

partition: \( P(A_1 \cup A_2 \ldots \cup A_k) = \Omega \)
\( P(A_i, A_j) = 0 \), for all \( i \neq j \)
First: Law of Total Probability

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$, for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$

**partition:**

$P(A_1 \cup A_2 \ldots \cup A_k) = \Omega$

$P(A_i, A_j) = 0$, for all $i \neq j$

When both of these conditions are true, we say "$A_1, \ldots, A_k$ partition $\Omega"
First: Law of Total Probability

GOAL: Relate \( P(A_i | B) \) to \( P(B | A_i) \),
for all \( i = 1 \ldots k \), where \( A_1 \ldots A_k \) partition \( \Omega \)

\textbf{partition:} \( P(A_1 \cup A_2 \ldots \cup A_k) = \Omega \)
\( P(A_i, A_j) = 0 \), for all \( i \neq j \)

\textbf{law of total probability:} If \( A_1 \ldots A_k \) partition \( \Omega \),
then for any event, \( B \):

\[
P(B) = \sum_{i=1}^{k} P(B | A_i)P(A_i)
\]
Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$,
for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$

Let’s try:

$$P(B) = \sum_{i=1}^{k} P(B | A_i) P(A_i)$$

Law of Total Probability
Law of Total Probability and Bayes Theorem

GOAL: Relate \( P(A_i | B) \) to \( P(B | A_i) \),
for all \( i = 1 \ldots k \), where \( A_1 \ldots A_k \) partition \( \Omega \)

Let’s try:

(1) \( P(A_i | B) = \frac{P(A_i, B)}{P(B)} \)

(2) \( \frac{P(A_i, B)}{P(B)} = \frac{P(B | A_i)}{P(A_i)} \) \( P(A_i) / P(B) \), by multiplication rule

\[
P(B) = \sum_{i=1}^{k} P(B | A_i) P(A_i)
\]

P(A,B) = P(B | A)P(A)
Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$, for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$

Let’s try:

1. $P(A_i | B) = \frac{P(A_i, B)}{P(B)}$  

2. $\frac{P(A_i, B)}{P(B)} = \frac{P(B | A_i) P(A_i)}{P(B)}$, by multiplication rule

*but in practice, we might not know $P(B)$*
Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$, for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$

Let’s try:

(1) $P(A_i | B) = P(A_i, B) / P(B)$

(2) $P(A_i, B) / P(B) = P(B | A_i) P(A_i) / P(B)$, by multiplication rule
   
   but in practice, we might not know $P(B)$

(3) $P(B | A_i) P(A_i) / P(B) = P(B | A_i) P(A_i) / (\sum_{i=1}^{k} P(B | A_i) P(A_i))$, by law of total probability
Law of Total Probability and Bayes Theorem

GOAL: Relate \( P(A_i | B) \) to \( P(B | A_i) \),
for all \( i = 1 \ldots k \), where \( A_1 \ldots A_k \) partition \( \Omega \)

Let’s try:

1. \( P(A_i | B) = \frac{P(A_i, B)}{P(B)} \)

2. \( \frac{P(A_i, B)}{P(B)} = \frac{P(B | A_i) P(A_i)}{P(B)} \), by multiplication rule
   
   *but in practice, we might not know \( P(B) \)*

3. \( \frac{P(B | A_i) P(A_i)}{P(B)} = \frac{P(B | A_i) P(A_i)}{ \left( \sum_{i=1}^{k} P(B | A_i) P(A_i) \right) } \), by law of total probability

Thus,

\[
P(A_i | B) = \frac{P(B | A_i) P(A_i)}{\sum_{i=1}^{k} P(B | A_i) P(A_i)}
\]
Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i | B)$ to $P(B | A_i)$, for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$

Let’s try:

(1) $P(A_i | B) = \frac{P(A_i, B)}{P(B)}$

(2) $P(A_i, B) / P(B) = P(B | A_i) \frac{P(A_i)}{P(B)}$, by multiplication rule; but in practice, we might not know $P(B)$

(3) $P(B | A_i) \frac{P(A_i)}{P(B)} = \frac{P(B | A_i) P(A_i)}{\sum_{i=1}^{k} P(B | A_i) P(A_i)}$, by law of total probability

Thus,

$$P(A_i | B) = \frac{P(B | A_i) P(A_i)}{\sum_{i=1}^{k} P(B | A_i) P(A_i)}$$
Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$, for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ partition $\Omega$.

Let’s try:

1. $P(A_i|B) = \frac{P(A_i,B)}{P(B)}$
2. $\frac{P(A_i,B)}{P(B)} = P(B|A_i) \frac{P(A_i)}{P(B)}$, by multiplication rule, but in practice, we might not know $P(B)$
3. $P(B|A_i) \frac{P(A_i)}{P(B)} = \frac{P(B|A_i) P(A_i)}{\sum_{i=1}^{k} P(B|A_i) P(A_i)}$, by law of total probability

Thus,

$$P(A_i|B) = \frac{P(B|A_i) P(A_i)}{\sum_{i=1}^{k} P(B|A_i) P(A_i)}$$
Probability Review:

- What constitutes a probability measure?
- Independence
- Conditional probability
- Conditional independence
- How to derive Bayes Theorem
- Multiplication Rule
- Partition of Sample Space
- Law of Total Probability
- Bayes Theorem in Practice