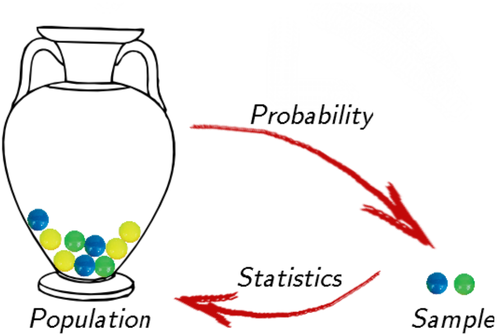
Probability Review

Rob Hall

September 9, 2010

What is Probability?



- Probability reasons about a sample, knowing the population.
- The goal of statistics is to estimate the population based on a sample.
- Both provide invaluable tools to modern machine learning.

Plan

- Facts about sets (to get our brains in gear).
- Definitions and facts about probabilities.
- Random variables and joint distributions.
- Characteristics of distributions (mean, variance, entropy).
- Some asymptotic results (a "high level" perspective).

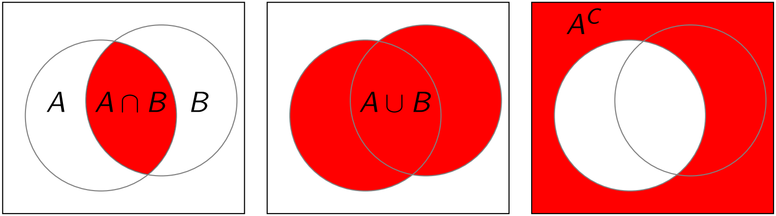
Goals: get some intuition about probability, learn how to formulate a simple proof, lay out some useful identities for use as a reference.

Non-goal: supplant an entire semester long course in probability.

Set Basics

A *set* is just a collection of *elements* denoted e.g.,
$S = \{s_1, s_2, s_3\}$, $R = \{r : \text{some condition holds on } r\}$.

- **Intersection**: the elements that are in both sets:
  $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- **Union**: the elements that are in either set, or both:
  $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- **Complementation**: all the elements that aren't in the set:
  $A^C = \{x : x \notin A\}$.



Properties of Set Operations

- **Commutativity**: $A \cup B = B \cup A$
- **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$.
- Likewise for intersection.
- Proof?

Properties of Set Operations

- **Commutativity**: $A \cup B = B \cup A$
- **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$.
- Likewise for intersection.
- Proof? Follows easily from commutative and associative properties of "and" and "or" in the definitions.

Properties of Set Operations

- **Commutativity**: $A \cup B = B \cup A$
- **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$.
- Likewise for intersection.
- Proof? Follows easily from commutative and associative properties of "and" and "or" in the definitions.
- **Distributive properties**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- Proof?

Properties of Set Operations

- **Commutativity**: $A \cup B = B \cup A$
- **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$.
- Likewise for intersection.
- Proof? Follows easily from commutative and associative properties of "and" and "or" in the definitions.
- **Distributive properties**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- Proof? Show each side of the equality contains the other.
- **DeMorgan's Law** ...see book.

Disjointness and Partitions

- A sequence of sets $A_1, A_2 \ldots$ is called **pairwise disjoint** or **mutually exclusive** if for all $i \neq j, A_i \cap A_j = \{\}$.
- If the sequence is pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then the sequence forms a **partition** of S.

Partitions are useful in probability theory and in life:

$$\begin{aligned} B \cap S &= B \cap \left(\bigcup_{i=1}^{\infty} A_i\right) \quad \text{(def of partition)} \\ &= \bigcup_{i=1}^{\infty} (B \cap A_i) \quad \text{(distributive property)} \end{aligned}$$

Note that the sets $B \cap A_i$ are also pairwise disjoint (proof?).

Disjointness and Partitions

- A sequence of sets $A_1, A_2 \ldots$ is called **pairwise disjoint** or **mutually exclusive** if for all $i \neq j, A_i \cap A_j = \{\}$.
- If the sequence is pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then the sequence forms a **partition** of S.

Partitions are useful in probability theory and in life:

$$\begin{aligned} B \cap S &= B \cap \left(\bigcup_{i=1}^{\infty} A_i\right) \quad \text{(def of partition)} \\ &= \bigcup_{i=1}^{\infty} (B \cap A_i) \quad \text{(distributive property)} \end{aligned}$$

Note that the sets $B \cap A_i$ are also pairwise disjoint (proof?).
If S is the whole space, what have we constructed?.

Probability Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	$\mathcal{F}$, E	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	$(\Omega, \mathcal{F}, P)$	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later).

Probability Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	$\mathcal{F}, E$	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	$(\Omega, \mathcal{F}, P)$	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later). e.g., rolling a fair die:

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Probability Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	$\mathcal{F}$, E	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	$(\Omega, \mathcal{F}, P)$	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later). e.g., rolling a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$\mathcal{F} = 2^{\Omega} = \{\{1\}, \{2\} \ldots \{1, 2\} \ldots \{1, 2, 3\} \ldots \{1, 2, 3, 4, 5, 6\}, \{\}\}$$

Probability Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	$\mathcal{F}, E$	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	$(\Omega, \mathcal{F}, P)$	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later). e.g., rolling a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$\mathcal{F} = 2^{\Omega} = \{\{1\}, \{2\} \ldots \{1, 2\} \ldots \{1, 2, 3\} \ldots \{1, 2, 3, 4, 5, 6\}, \{\}\}$$

$P(\{1\}) = P(\{2\}) = \ldots = \frac{1}{6}$ (i.e., a fair die)
$P(\{1, 3, 5\}) = \frac{1}{2}$ (i.e., half chance of odd result)
$P(\{1, 2, 3, 4, 5, 6\}) = 1$ (i.e., result is "almost surely" one of the faces).

Axioms for Probability

A set of conditions imposed on probability measures (due to Kolmogorov)

- $P(A) \geq 0, \forall A \in \mathcal{F}$
- $P(\Omega) = 1$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ where $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ are pairwise disjoint.

Axioms for Probability

A set of conditions imposed on probability measures (due to Kolmogorov)

- $P(A) \geq 0, \forall A \in \mathcal{F}$
- $P(\Omega) = 1$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ where $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ are pairwise disjoint.

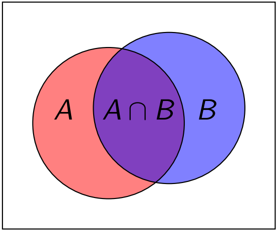
These quickly lead to:

- $P(A^C) = 1 - P(A)$ (since $P(A) + P(A^C) = P(A \cup A^C) = P(\Omega) = 1$).
- $P(A) \leq 1$ (since $P(A^C) \geq 0$).
- $P(\{\}) = 0$ (since $P(\Omega) = 1$).

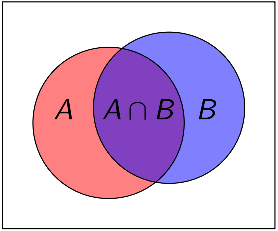
$P(A \cup B)$ – General Unions

Recall that A, A^C form a partition of Ω:

$$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$$



$P(A \cup B)$ – General Unions



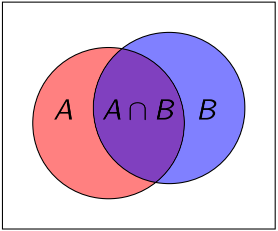
Recall that A, A^C form a partition of Ω:

$$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$$

And so: $P(B) = P(B \cap A) + P(B \cap A^C)$

For a general partition this is called the "law of total probability."

$P(A \cup B)$ – General Unions



Recall that A, A^C form a partition of Ω:

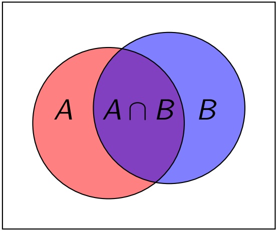
$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$

And so: $P(B) = P(B \cap A) + P(B \cap A^C)$

For a general partition this is called the "law of total probability."

$$P(A \cup B) \quad = \quad P(A \cup (B \cap A^C))$$

$P(A \cup B)$ – General Unions



Recall that A, A^C form a partition of Ω:

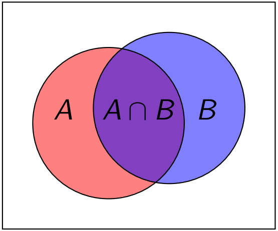
$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$

And so: $P(B) = P(B \cap A) + P(B \cap A^C)$

For a general partition this is called the "law of total probability."

$$
\begin{aligned}
P(A \cup B) &= P(A \cup (B \cap A^C)) \\
&= P(A) + P(B \cap A^C)
\end{aligned}
$$

$P(A \cup B)$ – General Unions



Recall that A, A^C form a partition of Ω:

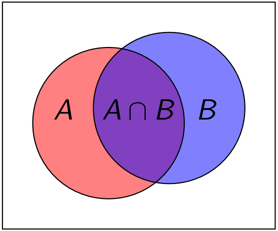
$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$

And so: $P(B) = P(B \cap A) + P(B \cap A^C)$

For a general partition this is called the "law of total probability."

$$
\begin{aligned}
P(A \cup B) &= P(A \cup (B \cap A^C)) \\
&= P(A) + P(B \cap A^C) \\
&= P(A) + P(B) - P(B \cap A)
\end{aligned}
$$

$P(A \cup B)$ – General Unions



Recall that A, A^C form a partition of Ω:

$B = B \cap \Omega = B \cap (A \cup A^C) = (B \cap A) \cup (B \cap A^C)$

And so: $P(B) = P(B \cap A) + P(B \cap A^C)$

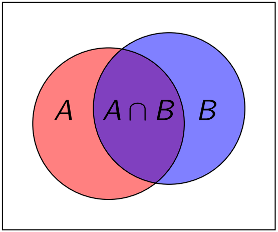
For a general partition this is called the "law of total probability."

$$\begin{aligned} P(A \cup B) &= P(A \cup (B \cap A^C)) \\ &= P(A) + P(B \cap A^C) \\ &= P(A) + P(B) - P(B \cap A) \\ &\leq P(A) + P(B) \end{aligned}$$

Very important difference between disjoint and non-disjoint unions.

Same idea yields the so-called "union bound" aka Boole's inequality

Conditional Probabilities

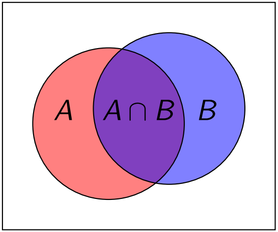


For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Interpretation: the outcome is definitely in B, so treat B as the entire sample space and find the probability that the outcome is also in A.

Conditional Probabilities



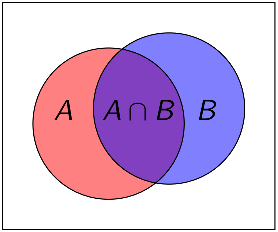
For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Interpretation: the outcome is definitely in B, so treat B as the entire sample space and find the probability that the outcome is also in A.

This rapidly leads to: $P(A|B)P(B) = P(A \cap B)$ aka the "chain rule for probabilities." (why?)

Conditional Probabilities



For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

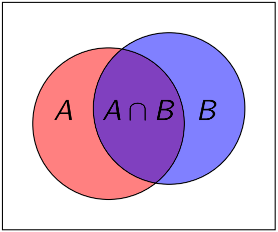
Interpretation: the outcome is definitely in B, so treat B as the entire sample space and find the probability that the outcome is also in A.

This rapidly leads to: $P(A|B)P(B) = P(A \cap B)$ aka the "chain rule for probabilities." (why?)

When $A_1, A_2 \ldots$ are a partition of Ω:

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

Conditional Probabilities



For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Interpretation: the outcome is definitely in B, so treat B as the entire sample space and find the probability that the outcome is also in A.

This rapidly leads to: $P(A|B)P(B) = P(A \cap B)$ aka the "chain rule for probabilities." (why?)

When $A_1, A_2 \ldots$ are a partition of Ω:

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

This is also referred to as the "law of total probability."

Conditional Probability Example

Suppose we throw a fair die:
$\Omega = \{1, 2, 3, 4, 5, 6\}, \; \mathcal{F} = 2^{\Omega}, \; P(\{i\}) = \frac{1}{6}, \; i = 1 \ldots 6$
$A = \{1, 2, 3, 4\}$ i.e., "result is less than 5,"
$B = \{1, 3, 5\}$ i.e., "result is odd."

$$P(A) \quad =$$

Conditional Probability Example

Suppose we throw a fair die:
$\Omega = \{1, 2, 3, 4, 5, 6\}, \; \mathcal{F} = 2^{\Omega}, \; P(\{i\}) = \frac{1}{6}, \; i = 1 \ldots 6$
$A = \{1, 2, 3, 4\}$ i.e., "result is less than 5,"
$B = \{1, 3, 5\}$ i.e., "result is odd."

$$
\begin{aligned}
P(A) &= \frac{2}{3} \\
P(B) &=
\end{aligned}
$$

Conditional Probability Example

Suppose we throw a fair die:
$\Omega = \{1,2,3,4,5,6\},\ \mathcal{F} = 2^{\Omega},\ P(\{i\}) = \frac{1}{6},\ i = 1 \ldots 6$
$A = \{1,2,3,4\}$ i.e., "result is less than 5,"
$B = \{1,3,5\}$ i.e., "result is odd."

$$\begin{aligned} P(A) &= \frac{2}{3} \\ P(B) &= \frac{1}{2} \\ P(A|B) &= \end{aligned}$$

Conditional Probability Example

Suppose we throw a fair die:
$\Omega = \{1, 2, 3, 4, 5, 6\},\ \mathcal{F} = 2^{\Omega},\ P(\{i\}) = \frac{1}{6},\ i = 1 \ldots 6$
$A = \{1, 2, 3, 4\}$ i.e., "result is less than 5,"
$B = \{1, 3, 5\}$ i.e., "result is odd."

$$
\begin{aligned}
P(A) &= \frac{2}{3} \\
P(B) &= \frac{1}{2} \\
P(A|B) &= \frac{P(A \cap B)}{P(B)}
\end{aligned}
$$

Conditional Probability Example

Suppose we throw a fair die:

$\Omega = \{1,2,3,4,5,6\},\ \mathcal{F} = 2^{\Omega},\ P(\{i\}) = \frac{1}{6},\ i = 1 \ldots 6$

$A = \{1,2,3,4\}$ i.e., "result is less than 5,"

$B = \{1,3,5\}$ i.e., "result is odd."

$$
\begin{aligned}
P(A) &= \frac{2}{3} \\
P(B) &= \frac{1}{2} \\
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(\{1,3\})}{P(B)}
\end{aligned}
$$

Conditional Probability Example

Suppose we throw a fair die:

$\Omega = \{1,2,3,4,5,6\},\ \mathcal{F} = 2^{\Omega},\ P(\{i\}) = \frac{1}{6},\ i = 1 \ldots 6$

$A = \{1,2,3,4\}$ i.e., "result is less than 5,"

$B = \{1,3,5\}$ i.e., "result is odd."

$$
\begin{aligned}
P(A) &= \frac{2}{3} \\
P(B) &= \frac{1}{2} \\
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(\{1,3\})}{P(B)} \\
&= \frac{2}{3}
\end{aligned}
$$

Conditional Probability Example

Suppose we throw a fair die:

$\Omega = \{1,2,3,4,5,6\}$, $\mathcal{F} = 2^{\Omega}$, $P(\{i\}) = \frac{1}{6}$, $i = 1 \ldots 6$

$A = \{1,2,3,4\}$ i.e., "result is less than 5,"

$B = \{1,3,5\}$ i.e., "result is odd."

$$\begin{aligned} P(A) &= \frac{2}{3} \\ P(B) &= \frac{1}{2} \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{1,3\})}{P(B)} \\ &= \frac{2}{3} \end{aligned}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Conditional Probability Example

Suppose we throw a fair die:
$\Omega = \{1,2,3,4,5,6\}$, $\mathcal{F} = 2^{\Omega}$, $P(\{i\}) = \frac{1}{6}$, $i = 1 \ldots 6$
$A = \{1,2,3,4\}$ i.e., "result is less than 5,"
$B = \{1,3,5\}$ i.e., "result is odd."

$$\begin{aligned} P(A) &= \frac{2}{3} \\ P(B) &= \frac{1}{2} \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{1,3\})}{P(B)} \\ &= \frac{2}{3} \end{aligned}$$

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{1}{2} \end{aligned}$$

Note that in general, $P(A|B) \neq P(B|A)$ however we may quantify their relationship.

Bayes' Rule

Using the chain rule we may see:

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

Rearranging this yields **Bayes' rule**:

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

Often this is written as:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

Where B_i are a partition of Ω (note the bottom is just the law of total probability).

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Two events A, B are called **conditionally independent given** C when $P(A \cap B|C) = P(A|C)P(B|C)$.

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Two events A, B are called **conditionally independent given** C when $P(A \cap B|C) = P(A|C)P(B|C)$.

When $P(A) > 0$ we may write $P(B|A, C) = P(B|C)$

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Two events A, B are called **conditionally independent given** C when $P(A \cap B|C) = P(A|C)P(B|C)$.

When $P(A) > 0$ we may write $P(B|A, C) = P(B|C)$
e.g., "the weather tomorrow is independent of the weather yesterday, knowing the weather today."

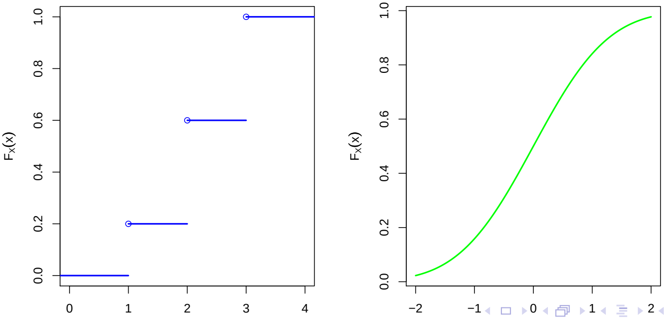
Random Variables – caution: hand waving

A **random variable** is a function $X : \Omega \rightarrow \mathbb{R}^d$
e.g.,

- Roll some dice, $X =$ sum of the numbers.
- Indicators of events: $X(\omega) = 1_A(\omega)$. e.g., toss a coin, $X = 1$ if it came up heads, 0 otherwise. Note relationship between the set theoretic constructions, and binary RVs.
- Give a few monkeys a typewriter, $X =$ fraction of overlap with complete works of Shakespeare.
- Throw a dart at a board, $X \in \mathbb{R}^2$ are the coordinates which are hit.

Distributions

- By considering random variables, we may think of probability measures as functions on the real numbers.
- Then, the probability measure associated with the RV is completely characterized by its **cumulative distribution function (CDF)**: $F_X(x) = P(X \leq x)$.
- If two RVs have the same CDF we call then **identically distributed**.
- We say $X \sim F_X$ or $X \sim f_X$ (f_X coming soon) to indicate that X has the distribution specified by F_X (resp, f_X).



Discrete Distributions

- If X takes on only a countable number of values, then we may characterize it by a **probability mass function (PMF)** which describes the probability of each value: $f_X(x) = P(X = x)$.

Discrete Distributions

- If X takes on only a countable number of values, then we may characterize it by a **probability mass function (PMF)** which describes the probability of each value: $f_X(x) = P(X = x)$.
- We have: $\sum_x f_X(x) = 1$ (why?)

Discrete Distributions

- If X takes on only a countable number of values, then we may characterize it by a **probability mass function (PMF)** which describes the probability of each value: $f_X(x) = P(X = x)$.
- We have: $\sum_x f_X(x) = 1$ (why?) – since each ω maps to one x, and $P(\Omega) = 1$.
- e.g., general discrete PMF: $f_X(x_i) = \theta_i$, $\sum_i \theta_i = 1, \theta_i \geq 0$.

Discrete Distributions

- If X takes on only a countable number of values, then we may characterize it by a **probability mass function (PMF)** which describes the probability of each value: $f_X(x) = P(X = x)$.
- We have: $\sum_x f_X(x) = 1$ (why?) – since each ω maps to one x, and $P(\Omega) = 1$.
- e.g., general discrete PMF: $f_X(x_i) = \theta_i, \ \sum_i \theta_i = 1, \theta_i \geq 0$.
- e.g., bernoulli distribution: $X \in \{0, 1\}, \ f_X(x) = \theta^x (1 - \theta)^{1-x}$
- A general model of binary outcomes (coin flips etc.).

Discrete Distributions

- Rather than specifying each probability for each event, we may consider a more restrictive parametric form, which will be easier to specify and manipulate (but sometimes less general).

Discrete Distributions

- Rather than specifying each probability for each event, we may consider a more restrictive parametric form, which will be easier to specify and manipulate (but sometimes less general).
- e.g., multinomial distribution:
  $X \in \mathbb{N}^d,\ \sum_{i=1}^d x_i = n,\ f_X(x) = \frac{n!}{x_1!x_2!\cdots x_d!} \prod_{i=1}^d \theta_i^{x_i}$.
- Sometimes used in text processing (dimensions correspond to words, n is the length of a document).
- What have we lost in going from a general form to a multinomial?

Continuous Distributions

- When the CDF is continuous we may consider its derivative $f_x(x) = \frac{d}{dx} F_X(x)$.
- This is called the **probability density function (PDF)**.

Continuous Distributions

- When the CDF is continuous we may consider its derivative $f_x(x) = \frac{d}{dx}F_X(x)$.
- This is called the **probability density function (PDF)**.
- The probability of an interval (a, b) is given by $P(a < X < b) = \int_a^b f_X(x)\ dx$.
- The probability of any specific point c is zero: $P(X = c) = 0$ (why?).

Continuous Distributions

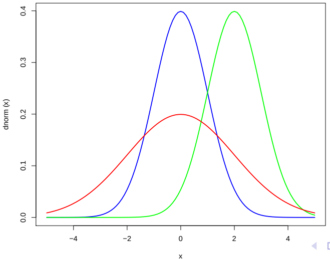
- When the CDF is continuous we may consider its derivative $f_x(x) = \frac{d}{dx} F_X(x)$.
- This is called the **probability density function (PDF)**.
- The probability of an interval (a, b) is given by $P(a < X < b) = \int_a^b f_X(x)\ dx$.
- The probability of any specific point c is zero: $P(X = c) = 0$ (why?).
- e.g., Uniform distribution: $f_X(x) = \frac{1}{b-a} \cdot 1_{(a,b)}(x)$

Continuous Distributions

- When the CDF is continuous we may consider its derivative $f_x(x) = \frac{d}{dx} F_X(x)$.
- This is called the **probability density function (PDF)**.
- The probability of an interval (a, b) is given by $P(a < X < b) = \int_a^b f_X(x)\ dx$.
- The probability of any specific point c is zero: $P(X = c) = 0$ (why?).
- e.g., Uniform distribution: $f_X(x) = \frac{1}{b-a} \cdot 1_{(a,b)}(x)$
- e.g., Gaussian aka "normal:" $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{(x-\mu)^2}{2\sigma^2}\}$

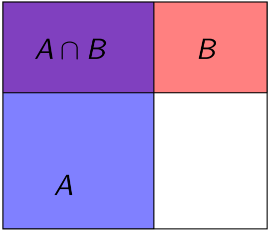
Continuous Distributions

- When the CDF is continuous we may consider its derivative $f_x(x) = \frac{d}{dx} F_X(x)$.
- This is called the **probability density function (PDF)**.
- The probability of an interval (a, b) is given by $P(a < X < b) = \int_a^b f_X(x)\ dx$.
- The probability of any specific point c is zero: $P(X = c) = 0$ (why?).
- e.g., Uniform distribution: $f_X(x) = \frac{1}{b-a} \cdot 1_{(a,b)}(x)$
- e.g., Gaussian aka "normal:" $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{(x-\mu)^2}{2\sigma^2}\}$
- Note that both families give probabilities for every interval on the real line, yet are specified by only two numbers.



Multiple Random Variables

We may consider multiple functions of the same sample space, e.g., $X(\omega) = 1_A(\omega)$, $Y(\omega) = 1_B(\omega)$:



May represent the **joint distribution** as a table:

	X=0	X=1
Y=0	0.25	0.15
Y=1	0.35	0.25

We write the joint PMF or PDF as $f_{X,Y}(x, y)$

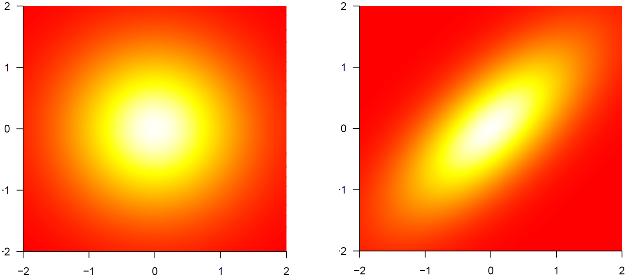
Multiple Random Variables

Two random variables are called **independent** when the joint PDF factorizes:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

When RVs are independent and identically distributed this is usually abbreviated to "i.i.d."

Relationship to independent events: X, Y ind. iff $\{\omega : X(\omega) \leq x\}, \{\omega : Y(\omega) \leq y\}$ are independent events for all x, y.



Working with a Joint Distribution

We have similar constructions as we did in abstract prob. spaces:

- **Marginalizing**: $f_X(x) = \int_{\mathcal{Y}} f_{X,Y}(x,y)\ dy$.
  Similar idea to the law of total probability (identical for a discrete distribution).
- **Conditioning**: $f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{\mathcal{X}} f_{X,Y}(x,y)\ dx}$.
  Similar to previous definition.

Old?	Blood pressure?	Heart Attack?	P
0	0	0	0.22
0	0	1	0.01
0	1	0	0.15
0	1	1	0.01
1	0	0	0.18
...	...	...	...

How to compute $P(\text{heart attack}|\text{old})$?

Characteristics of Distributions

We may consider the **expectation** (or "mean") of a distribution:

$$E(X) = \begin{cases} \sum_x x f_X(x) & \text{X is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\ dx & \text{X is continuous} \end{cases}$$

Characteristics of Distributions

We may consider the **expectation** (or "mean") of a distribution:

$$E(X) = \begin{cases} \sum_x x f_X(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & X \text{ is continuous} \end{cases}$$

Expectation is linear:

$$\begin{aligned}
E(aX + bY + c) &= \sum_{x,y} (ax + by + c) f_{X,Y}(x,y) \\
&= \sum_{x,y} ax f_{X,Y}(x,y) + \sum_{x,y} by f_{X,Y}(x,y) + \sum_{x,y} c f_{X,Y}(x,y) \\
&= a \sum_{x,y} x f_{X,Y}(x,y) + b \sum_{x,y} y f_{X,Y}(x,y) + c \sum_{x,y} f_{X,Y}(x,y) \\
&= a \sum_{x} x \sum_{y} f_{X,Y}(x,y) + b \sum_{y} y \sum_{x} f_{X,Y}(x,y) + c \\
&= a \sum_{x} x f_X(x) + b \sum_{y} y f_Y(y) + c \\
&= aE(X) + bE(Y) + c
\end{aligned}$$

Characteristics of Distributions

Questions:

1. $E[EX] =$

Characteristics of Distributions

Questions:

1. $E[EX] = \sum_x (EX) f_X(x) =$

Characteristics of Distributions

Questions:

1. $E[EX] = \sum_x (EX) f_X(x) = (EX) \sum_x f_X(x) = EX$

Characteristics of Distributions

Questions:

1. $E[EX] = \sum_x (EX) f_X(x) = (EX) \sum_x f_X(x) = EX$
2. $E(X \cdot Y) = E(X)E(Y)$?

Characteristics of Distributions

Questions:

1. $E[EX] = \sum_x (EX) f_X(x) = (EX) \sum_x f_X(x) = EX$
2. $E(X \cdot Y) = E(X)E(Y)$?

   Not in general, although when $f_{X,Y} = f_X f_Y$:

$$E(X \cdot Y) = \sum_{x,y} xy f_X(x) f_Y(y) = \sum_x x f_X(x) \sum_y y f_Y(y) = EX \cdot EY$$

Characteristics of Distributions

We may consider the **variance** of a distribution:

$$\mathrm{Var}(X) = E(X - EX)^2$$

This may give an idea of how “spread out” a distribution is.

Characteristics of Distributions

We may consider the **variance** of a distribution:

$$\mathrm{Var}(X) = E(X - EX)^2$$

This may give an idea of how "spread out" a distribution is.

A useful alternate form is:

$$\begin{aligned} E(X - EX)^2 &= E[X^2 - 2XE(X) + (EX)^2] \\ &= E(X^2) - 2E(X)E(X) + (EX)^2 \\ &= E(X^2) - (EX)^2 \end{aligned}$$

Characteristics of Distributions

We may consider the **variance** of a distribution:

$$\mathrm{Var}(X) = E(X - EX)^2$$

This may give an idea of how "spread out" a distribution is.

A useful alternate form is:

$$\begin{aligned} E(X - EX)^2 &= E[X^2 - 2XE(X) + (EX)^2] \\ &= E(X^2) - 2E(X)E(X) + (EX)^2 \\ &= E(X^2) - (EX)^2 \end{aligned}$$

Variance of a coin toss?

Characteristics of Distributions

Variance is non-linear but the following holds:

$$\mathrm{Var}(aX) = E(aX - E(aX))^2 = E(aX - aEX)^2 = a^2 E(X - EX)^2 = a^2 \mathrm{Var}(X)$$

Characteristics of Distributions

Variance is non-linear but the following holds:

$$\text{Var}(aX) = E(aX - E(aX))^2 = E(aX - aEX)^2 = a^2 E(X - EX)^2 = a^2 \text{Var}(X)$$

$$\text{Var}(X + c) = E(X + c - E(X + c))^2 = E(X - EX + c - c)^2 = E(X - EX)^2 = \text{Var}(X)$$

Characteristics of Distributions

Variance is non-linear but the following holds:

$$\mathrm{Var}(aX) = E(aX - E(aX))^2 = E(aX - aEX)^2 = a^2 E(X - EX)^2 = a^2 \mathrm{Var}(X)$$

$$\mathrm{Var}(X+c) = E(X+c-E(X+c))^2 = E(X-EX+c-c)^2 = E(X-EX)^2 = \mathrm{Var}(X)$$

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E(X - EX + Y - EY)^2 \\
&= \underbrace{E(X - EX)^2}_{\mathrm{Var}(X)} + \underbrace{E(Y - EY)^2}_{\mathrm{Var}(Y)} + 2\underbrace{E(X - EX)(Y - EY)}_{\mathrm{Cov}(X,Y)}
\end{aligned}$$

Characteristics of Distributions

Variance is non-linear but the following holds:

$$\mathrm{Var}(aX) = E(aX - E(aX))^2 = E(aX - aEX)^2 = a^2 E(X - EX)^2 = a^2 \mathrm{Var}(X)$$

$$\mathrm{Var}(X+c) = E(X+c-E(X+c))^2 = E(X-EX+c-c)^2 = E(X-EX)^2 = \mathrm{Var}(X)$$

$$\begin{aligned} \mathrm{Var}(X+Y) &= E(X - EX + Y - EY)^2 \\ &= \underbrace{E(X-EX)^2}_{\mathrm{Var}(X)} + \underbrace{E(Y-EY)^2}_{\mathrm{Var}(Y)} + 2\underbrace{E(X-EX)(Y-EY)}_{\mathrm{Cov}(X,Y)} \end{aligned}$$

So when X, Y are independent we have:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

(why?)

Putting it all together

Say we have $X_1 \dots X_n$ i.i.d., where $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$$E(\bar{X}_n) =$$

Putting it all together

Say we have $X_1 \dots X_n$ i.i.d., where $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n) = E[\frac{1}{n}\sum_{i=1}^{n} X_i] =$$

Putting it all together

Say we have $X_1 \dots X_n$ i.i.d., where $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$$E(\bar{X}_n) = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n E(X_i) =$$

Putting it all together

Say we have $X_1 \ldots X_n$ i.i.d., where $EX_i = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n) = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}n\mu = \mu$$

Putting it all together

Say we have $X_1 \dots X_n$ i.i.d., where $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$$E(\bar{X}_n) = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\text{Var}(\bar{X}_n) = \text{Var}(\frac{1}{n}\sum_{i=1}^n X_i) =$$

Putting it all together

Say we have $X_1 \dots X_n$ i.i.d., where $EX_i = \mu$ and $\text{Var}(X_i) = \sigma^2$.

We want to know the expectation and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$$E(\bar{X}_n) = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

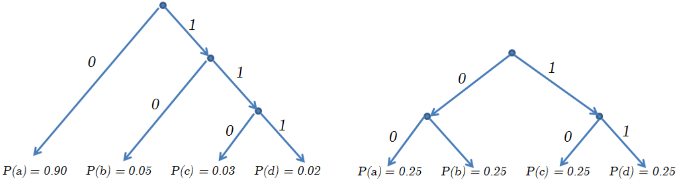
$$\text{Var}(\bar{X}_n) = \text{Var}(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Entropy of a Distribution

Entropy is a measure of uniformity in a distribution.

$$H(X) = -\sum_{x} f_X(x) \log_2 f_X(x)$$

Imagine you had to transmit a sample from f_X, so you construct the optimal encoding scheme:



Entropy gives the mean depth in the tree (= mean number of bits).

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

We had: $E\bar{X}_n = \mu, \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Distribution appears to be "contracting:" as n increases, variance is going to 0.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

We had: $E\bar{X}_n = \mu, \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Distribution appears to be "contracting:" as n increases, variance is going to 0.

Using Chebyshev's inequality:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$$

For any fixed ϵ, as $n \to \infty$.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

The **weak law of large numbers**:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

In English: choose ϵ and a probability that $|\bar{X}_n - \mu| < \epsilon$, I can find you an n so your probability is achieved.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

The **weak law of large numbers**:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

In English: choose ϵ and a probability that $|\bar{X}_n - \mu| < \epsilon$, I can find you an n so your probability is achieved.

The **strong law of large numbers**:

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$$

In English: the mean converges to the expectation "almost surely" as n increases.

Law of Large Numbers (LLN)

Recall our variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We may wonder about its behavior as $n \to \infty$.

The **weak law of large numbers**:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

In English: choose ϵ and a probability that $|\bar{X}_n - \mu| < \epsilon$, I can find you an n so your probability is achieved.

The **strong law of large numbers**:

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$$

In English: the mean converges to the expectation "almost surely" as n increases.

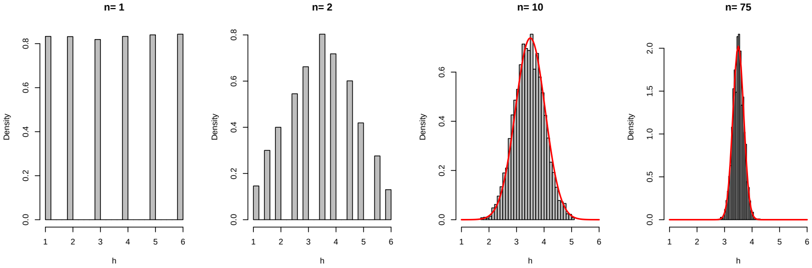
Two different versions, each holds under different conditions, but i.i.d. and finite variance is enough for either.

Central Limit Theorem (CLT)

The distribution of $\bar{X}_n$ also converges weakly to a Gaussian,

$$\lim_{n \to \infty} F_{\bar{X}_n}(x) = \Phi\left(\frac{x-\mu}{\sqrt{n}\sigma}\right)$$

Simulated n dice rolls and took average, 5000 times:

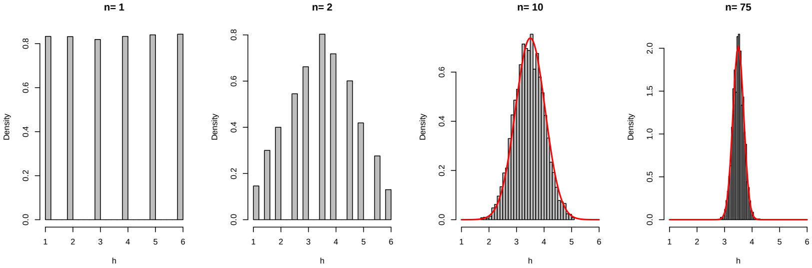


Central Limit Theorem (CLT)

The distribution of $\bar{X}_n$ also converges weakly to a Gaussian,

$$\lim_{n\to\infty} F_{\bar{X}_n}(x) = \Phi\left(\frac{x-\mu}{\sqrt{n}\sigma}\right)$$

Simulated n dice rolls and took average, 5000 times:



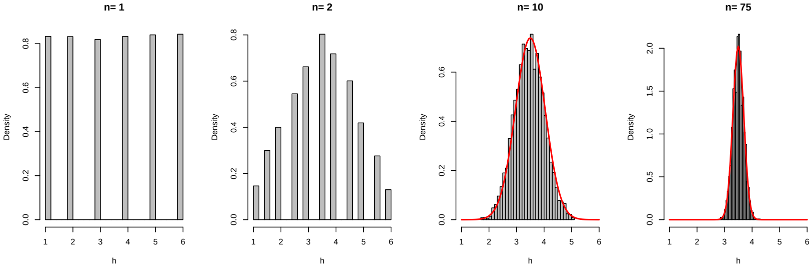
Two kinds of convergence went into this picture (why 5000?):

Central Limit Theorem (CLT)

The distribution of $\bar{X}_n$ also converges weakly to a Gaussian,

$$\lim_{n \to \infty} F_{\bar{X}_n}(x) = \Phi\left(\frac{x - \mu}{\sqrt{n}\sigma}\right)$$

Simulated n dice rolls and took average, 5000 times:



Two kinds of convergence went into this picture (why 5000?):

1. True distribution converges to a Gaussian (CLT).
2. Empirical distribution converges to true distribution (Glivenko-Cantelli).

Asymptotics Opinion

Ideas like these are crucial to machine learning:

- We want to minimize error on a whole population (e.g., classify text documents as well as possible)
- We minimize error on a training set of size n.
- What happens as $n \to \infty$?
- How does the complexity of the model, or the dimension of the problem affect convergence?