Probability Theory

Probability and Statistics for Data Science
CSE594 - Spring 2016

What is Probability?

What is Probability?

Examples

- outcome of flipping a coin (seminal example)
- amount of snowfall
- mentioning a word
- mentioning a word “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.

“Mathematical language for quantifying uncertainty” - Wasserman

Probability (review)

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

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

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

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(Ω)** = 1
- **P(*A*)** ≥ 0 **,** for all ***A***
- If A_1, A_2, … are disjoint events then: $$\mathrm{P}\left(\bigcup_i^{\infty} A_i\right) = \sum_i^{\infty} \mathrm{P}(A_i)$$

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

- **P(Ω)** = 1
- **P(*A*)** ≥ 0 **,** for all ***A***
- If A_1, A_2, … are disjoint events then: $$\mathrm{P}\left(\bigcup_i^\infty A_i\right) = \sum_i^\infty \mathrm{P}(A_i)$$

Probability

Examples

- outcome of flipping a coin (seminal example)
- amount of snowfall
- mentioning a word
- mentioning a word “a lot”

Probability (review)

Some Properties:

If $\boldsymbol{B} \subseteq \boldsymbol{A}$ then $\mathbf{P}(\boldsymbol{A}) \geq \mathbf{P}(\boldsymbol{B})$

$\mathbf{P}(\boldsymbol{A} \cup \boldsymbol{B}) \leq \mathbf{P}(\boldsymbol{A}) + \mathbf{P}(\boldsymbol{B})$

$\mathbf{P}(\boldsymbol{A} \cap \boldsymbol{B}) \leq \mathbf{min}(\mathbf{P}(\boldsymbol{A}), \mathbf{P}(\boldsymbol{B}))$

$\mathbf{P}(\neg \boldsymbol{A}) = \mathbf{P}(\Omega \,/\, \boldsymbol{A}) = 1 - \mathbf{P}(\boldsymbol{A})$

/ is set difference
$P(A \cap B)$ will be notated as $P(A, B)$

Probability (Review)

Independence

Two Events: A and B

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

Probability (Review)

Independence

Two Events: *A* and *B*

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

- A: first flip of a fair coin; B: second flip of the same fair coin
- A: mention or not of the word “happy”
  B: mention or not of the word “birthday”

Probability (Review)

Independence

Two Events: *A* and *B*

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

- A: first flip of a fair coin; B: second flip of the same fair coin
- A: mention or not of the word “happy”
  B: mention or not of the word “birthday”

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

Probability (Review)

Conditional Probability

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

Probability (Review)

Conditional Probability

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

H: mention “happy” in message, m

B: mention “birthday” in message, m

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

P(H|B) = ??

Probability (Review)

Conditional Probability

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

H: mention "happy" in message, m

B: mention "birthday" in message, m

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

P(H|B) = .50

H1: first flip of a fair coin is heads

H2: second flip of the same coin is heads

P(H2) = **0.5** P(H1) = 0.5 P(H2, H1) = 0.25

P(H2|H1) = **0.5**

Probability (Review)

Conditional Probability

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

H1: first flip of a fair coin is heads

H2: second flip of the same coin is heads

P(H2) = **0.5** P(H1) = 0.5 P(H2, H1) = 0.25

P(H2|H1) = **0.5**

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

P(*A*, *B*) = P(*A*)P(*B*) iff **P(*A*|*B*) = P(A)**

Probability (Review)

Conditional Probability

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

H1: first flip of a fair coin is heads

H2: second flip of the same coin is heads

P(H2) = **0.5** P(H1) = 0.5 P(H2, H1) = 0.25

P(H2|H1) = **0.5**

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

P(*A*, *B*) = P(*A*)P(*B*) iff **P(*A*|*B*) = P(A)**

Interpretation of Independence:

Observing B has no effect on probability of A.

Why Probability?

Why Probability?

A formality to make sense of the world.

- To quantify uncertainty
  Should we believe something or not? Is it a meaningful difference?

- 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?

- To organize data into meaningful groups or “dimensions”
  Where does X belong? What words are similar to X?

Random Variables

X: A mapping from $\boldsymbol{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Random Variables

X: A mapping from $\mathbf{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\mathbf{\Omega}$ = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X only has 6 possible values: 0, 1, 2, 3, 4, 5

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with k = 4 tails?

$\mathbf{P}(X(\omega) = k)$ where $\omega \in \mathbf{\Omega}$

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ ) \qquad \text{where } \omega \in \mathbf{\Omega}$$

Random Variables

X: A mapping from $\mathbf{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\mathbf{\Omega}$ = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

$X(<HHHHH>) = 0$

$X(<HHHTH>) = 1$

$X(<TTTHT>) = 4$

$X(<HTTTT>) = 4$

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ )$ where $\omega \in \mathbf{\Omega}$

$X(\omega) = 4$ for 5 out of 32 sets in $\mathbf{\Omega}$. Thus, assuming a fair coin, $\mathbf{P}(X = 4) = 5/32$

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with k = 4 tails?

$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ )$ where $\omega \in \mathbf{\Omega}$

X(ω) = 4 for 5 out of 32 sets in **Ω**. Thus, assuming a fair coin, $\mathbf{P}(X = 4) = 5/32$

(Not a variable, but a function that we end up notating a lot like a variable)

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {**<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>**…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X is a *discrete random variable* if it takes only a countable number of values.

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$\mathbf{P}(X = k) := \mathbf{P}(\{\omega : X(\omega) = k\})$ where $\omega \in \mathbf{\Omega}$

$X(\omega) = 4$ for 5 out of 32 sets in **Ω**. Thus, assuming a fair coin, $\mathbf{P}(X = 4) = 5/32$

(Not a variable, but a function that we end up notating a lot like a variable)

Random Variables

$\mathbf{X}$: A mapping from $\boldsymbol{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

$\mathbf{X}$ is a *discrete random variable* if it takes only a countable number of values.

Random Variables

$\mathbf{X}$: A mapping from $\boldsymbol{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\boldsymbol{\Omega}$ = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

Random Variables

$\mathbf{X}$: A mapping from $\boldsymbol{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\boldsymbol{\Omega}$ = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

Random Variables

$\mathbf{X}$: A mapping from $\mathbf{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\mathbf{\Omega}$ = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

What is the probability we receive (at least) a *inches?*

$\mathbf{P}(X \geq a) := \mathbf{P}(\ \{\omega : X(\omega) \geq a\}\ )$

What is the probability we receive between a *and* b *inches?*

$\mathbf{P}(a \leq X \leq b) := \mathbf{P}(\ \{\omega : a \leq X(\omega) \leq b\}\ )$

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

$\mathbf{P}(X = i) := 0$, for all $i \in \mathbf{\Omega}$

(probability of receiving exactly i inches of snowfall is zero)

What is the probability we receive (at least) a *inches?*

$\mathbf{P}(X \geq a) := \mathbf{P}(\ \{\omega : X(\omega) \geq a\}\ )$

What is the probability we receive between a *and* b *inches?*

$\mathbf{P}(a \leq X \leq b) := \mathbf{P}(\ \{\omega : a \leq X(\omega) \leq b\}\ )$

Probability Review: 1-26

- what constitutes a probability measure?
- independence
- conditional probability
- random variables
  - discrete
  - continuous

Language Models Review: 1-28

- Why are language models (LMs) useful?
- Maximum Likelihood Estimation for Binomials
- Idea of Chain Rule, Markov assumptions
- Why is word sparsity an issue?
- Further interest: Leplace Smoothing, Good-Turing Smoothing, LMs in topic modeling.

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

$$P(A,B) = 0$$

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

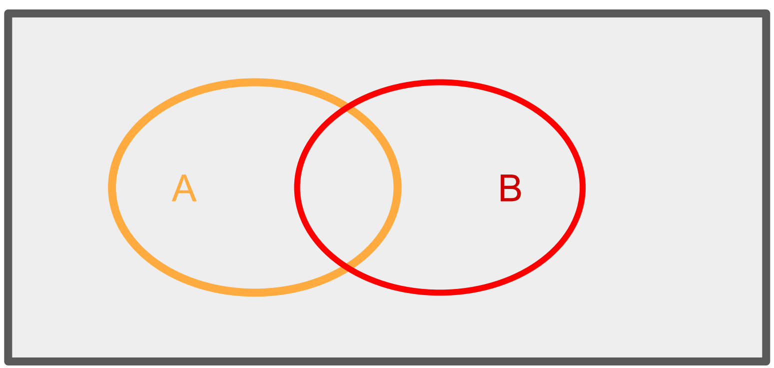
Does independence imply disjoint?

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: A: first coin flip is heads, B: second coin flip is heads;
P(A)P(B) = P(A,B), but .25 = P(A, B) =/= 0



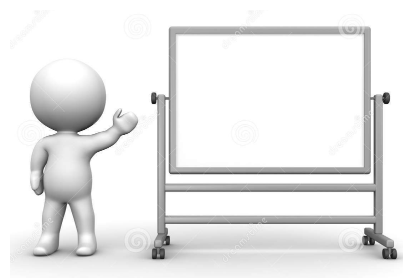
Disjoint Sets vs. Independent Events

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? No
Proof: A counterexample: A: first coin flip is heads, B: second coin flip is heads;
P(A)P(B) = P(A,B), but .25 = P(A, B) =/= 0

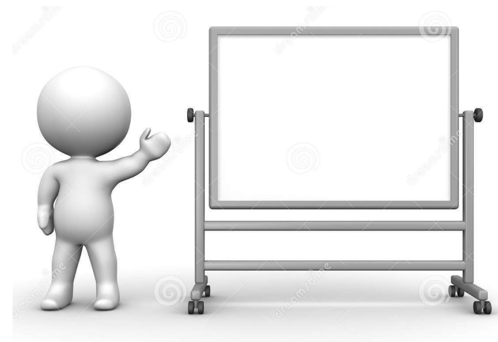
Does disjoint imply independence?



Tools for Decomposing Probabilities

Whiteboard Time!

- Table
- Tree



Examples:

- urn with 3 balls (with and without replacement)
- conversation lengths
- championship bracket

Probabilities over >2 events...

Independence:

$A_1, A_2, \ldots, A_n$ are independent iff $P(A_1, A_2, \ldots, A_n) = \prod P(A_i)$

Probabilities over >2 events...

Independence:

$A_1, A_2, \ldots, A_n$ are independent iff $P(A_1, A_2, \ldots, A_n) = \prod P(A_i)$

Conditional Probability:

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

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

(just think of multiple events happening as a single event)

Conditional Independence

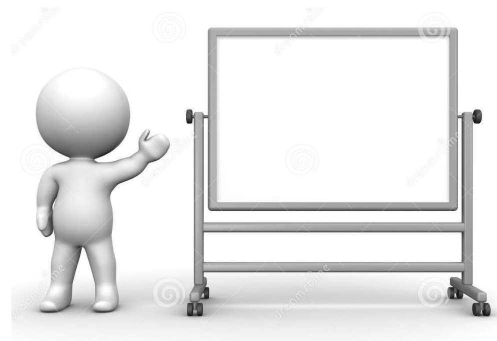
A and *B* are conditionally independent, given C, IFF

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

Equivalently, $P(A|B,C) = P(A|C)$

Interpretation: *Once we know C, B doesn't tell us anything useful about A.*

Example: Championship bracket



Bayes Theorem - Lite

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

Let's try:

Bayes Theorem - Lite

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

Let's try:

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

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

Bayes Theorem - Lite

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

Let's try:

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

(2) P(B|A) = P(B,A) / P(A) = P(A,B) / P(A), def. of conf. prob; sym of set union

(3) P(A,B) = P(B|A)P(A), algebra on (2) ← known as "Multiplication Rule"

Bayes Theorem - Lite

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

Let's try:

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

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

(3) $P(A,B) = P(B|A)P(\mathrm{A})$, algebra on (2) ← known as "Multiplication Rule"

(4) $P(A|B) = P(B|A)P(A) / P(B)$, Substitute $P(\mathrm{A,B})$ from (3) into (1)

Bayes Theorem - Lite

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

Let's try:

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

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

(3) $P(A,B) = P(B|A)P(A)$, algebra on (2) ← known as "Multiplication Rule"

(4) $\boxed{P(A|B) = P(B|A)P(A) / P(B)}$, Substitute P(A,B) from (3) into (1)

Law of Total Probability and Bayes Theorem

GOAL: Relate P($A_i|B$) to P($B|A_i$),

for all i = 1 ... k, where A_1 ... A_k **partition** Ω

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_\mathrm{i})$,
for all i = 1 ... k, where $A_1 \dots A_k$ **partition** Ω

partition: $P(A_1 \cup A_2 \dots \cup A_k) = \Omega$
$P(A_i, A_j) = 0$, for all i ≠ j

Law of Total Probability and Bayes Theorem

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

partition: $P(A_1 \cup A_2 \ldots \cup A_k) = \Omega$
$P(A_i, A_j) = 0$, for all i ≠ j

law of total probability: If $A_1 \ldots A_k$ **partition** Ω,
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 $P(B|A_i)$,

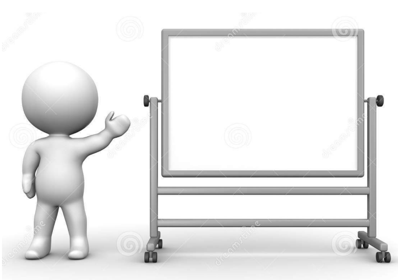
for all i = 1 ... k, where $A_1 \dots A_k$ **partition** Ω

partition: $P(A_1 \cup A_2 \dots \cup A_k) = \Omega$

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

law of total probability: If $A_1 \dots A_k$ **partition** Ω,

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 ... k, where $A_1 \ldots A_k$ **partition** Ω

Let's try:

Law of Total Probability and Bayes Theorem

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

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

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where $A_1 \dots A_k$ **partition** Ω

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)

Law of Total Probability and Bayes Theorem

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

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) $P(B|A_i)\, P(A_i) / P(B) = P(B|A_i)\, P(A_i) / \left(\sum_{i=1}^{k} P(B|A_i)P(A_i) \right)$, 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 ... k, where $A_1 \ldots A_k$ **partition** Ω

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

Thus, $\boxed{P(A_i|B) = P(B|A_i)\, P(A_i) / \left(\sum_{i=1}^{k} P(B|A_i)P(A_i) \right)}$

Probability Theory Review: 2-2

- Conditional Independence
- How to derive Bayes Theorem
- Law of Total Probability
- Bayes Theorem in Practice

Working with data in python

= refer to python notebook

Random Variables, Revisited

$\mathbf{X}$: A mapping from Ω to $\mathbb{R}$ that describes the question we care about in practice.

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

$\mathbf{X}$ is a *discrete random variable* if it takes only a countable number of values.

Random Variables, Revisited

$\mathbf{X}$: A mapping from $\boldsymbol{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\boldsymbol{\Omega}$ = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

$\mathbf{P}(X = i) := 0$, for all $i \in \boldsymbol{\Omega}$

(probability of receiving exactly i inches of snowfall is zero)

What is the probability we receive (at least) a *inches?*

$\mathbf{P}(X \geq a) := \mathbf{P}(\ \{\omega : X(\omega) \geq a\}\ )$

What is the probability we receive between a *and* b *inches?*

$\mathbf{P}(a \leq X \leq b) := \mathbf{P}(\ \{\omega : a \leq X(\omega) \leq b\}\ )$

Random Variables, Revisited

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

$\mathbf{P}(X = i) := 0$, for all $i \in \mathbf{\Omega}$

(probability of receiving <u>exactly</u> i inches of snowfall is zero)

How to model?

s?

inches?

Continuous Random Variables

Discretize them!

(group into discrete bins)

How to model?

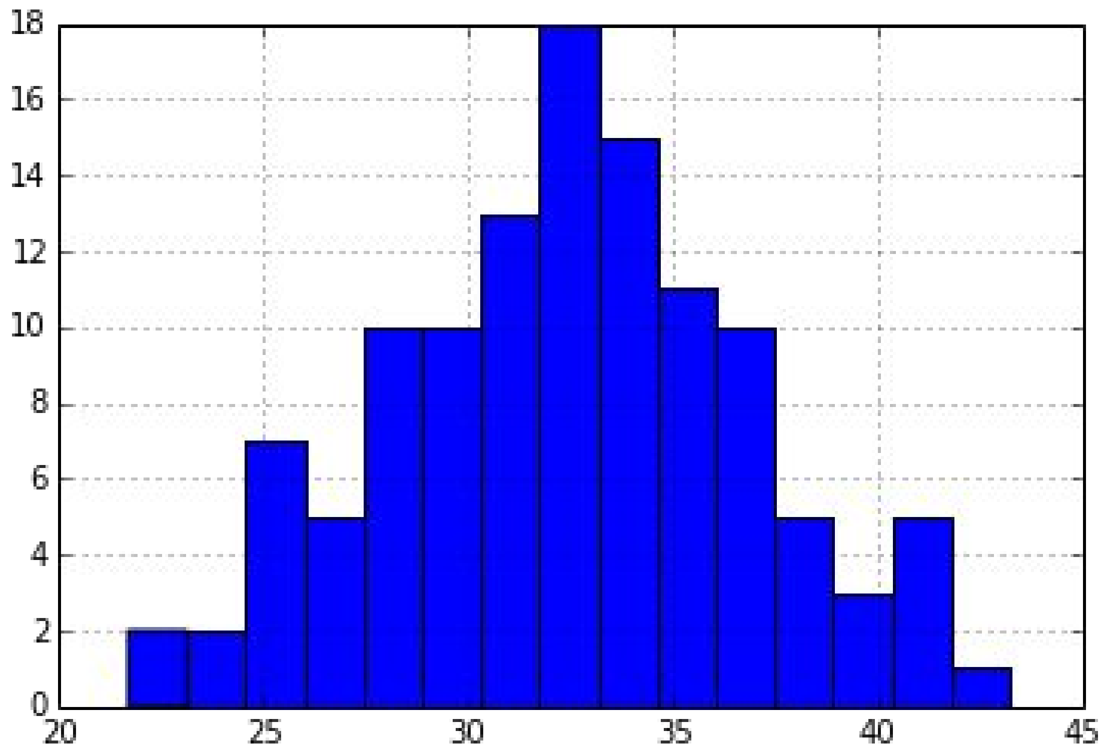
Continuous Random Variables

Discretize them!
(group into discrete bins)

How to model?

Histograms

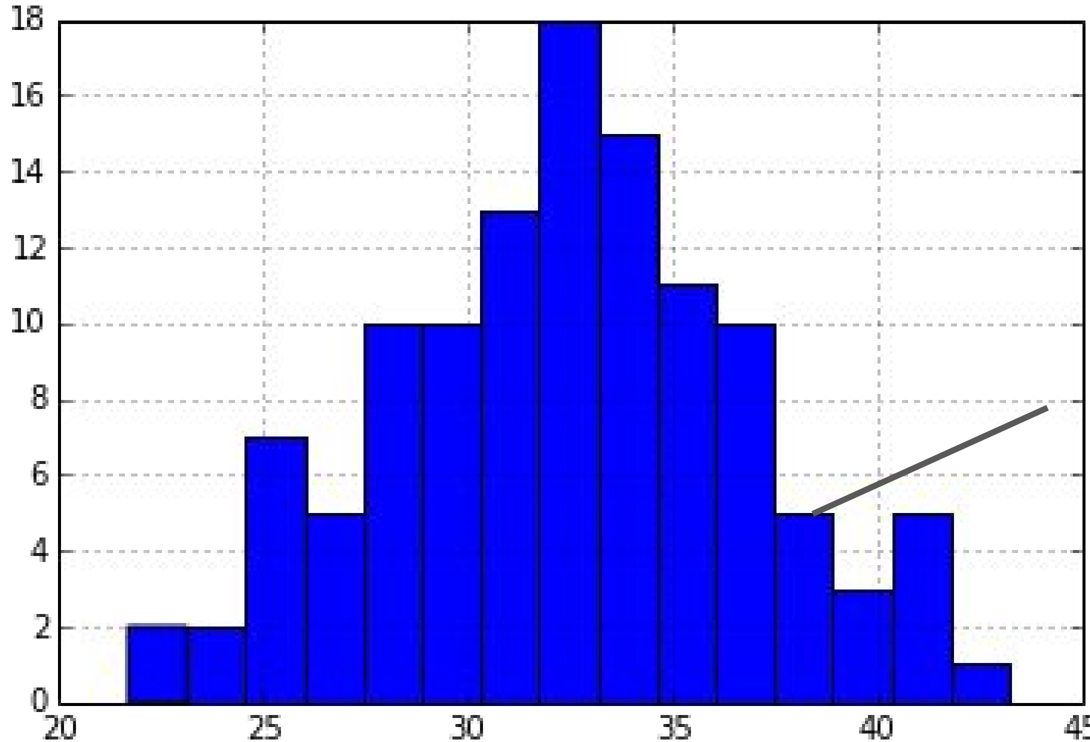
Continuous Random Variables



Continuous Random Variables

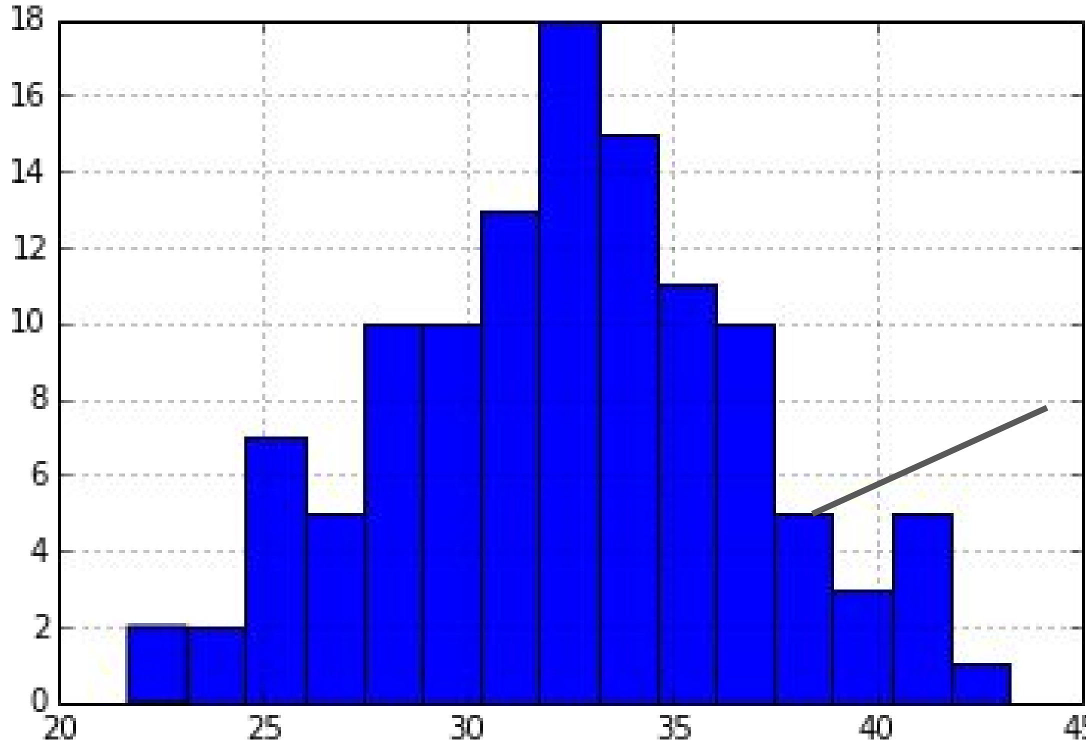
$P(bin=8) = .32$

$P(bin=12) = .08$



Continuous Random Variables

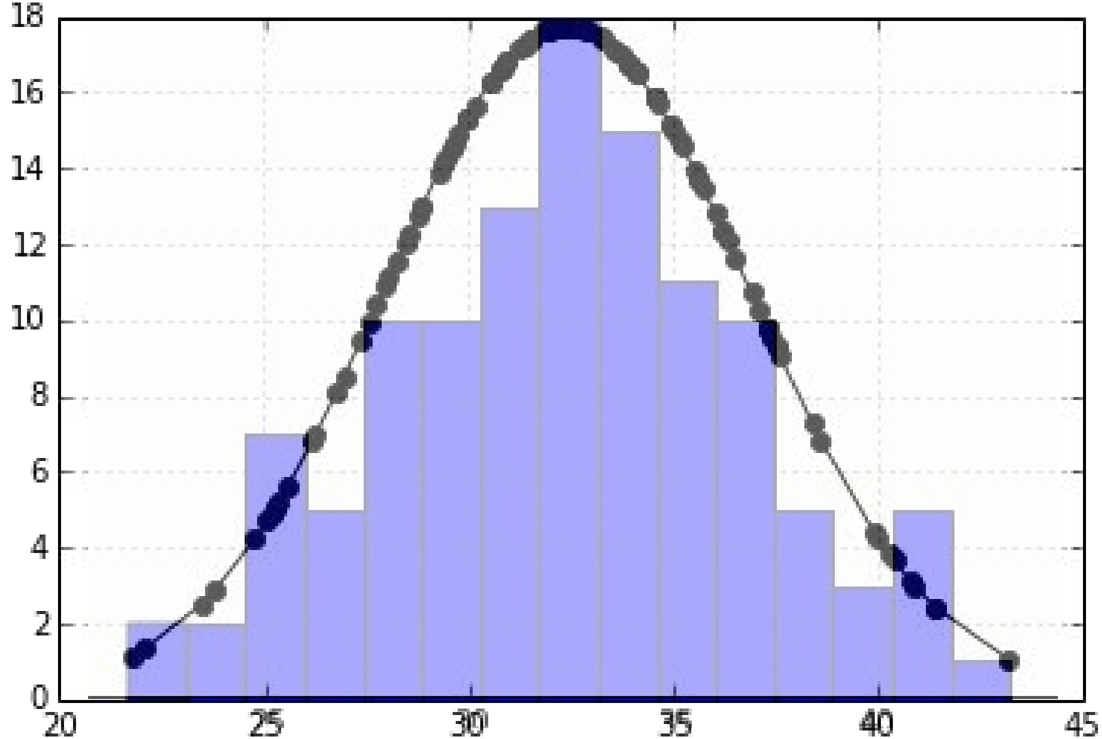
P(*bin*=8) = .32



P(*bin*=12) = .08

But aren't we throwing away information?

Continuous Random Variables



Continuous Random Variables

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *continuous random variable* if there exists a function *fx* such that:

$$f_X(x) \geq 0, \text{for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \quad \text{and}$$

$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

Continuous Random Variables

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *continuous random variable* if there exists a function *fx* such that:

$$f_X(x) \geq 0, \text{for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \quad \text{and}$$

$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

fx **: "probability density function" (pdf)**

Continuous Random Variables

PDFs

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *continuous random variable* if there exists a function *fx* such that:

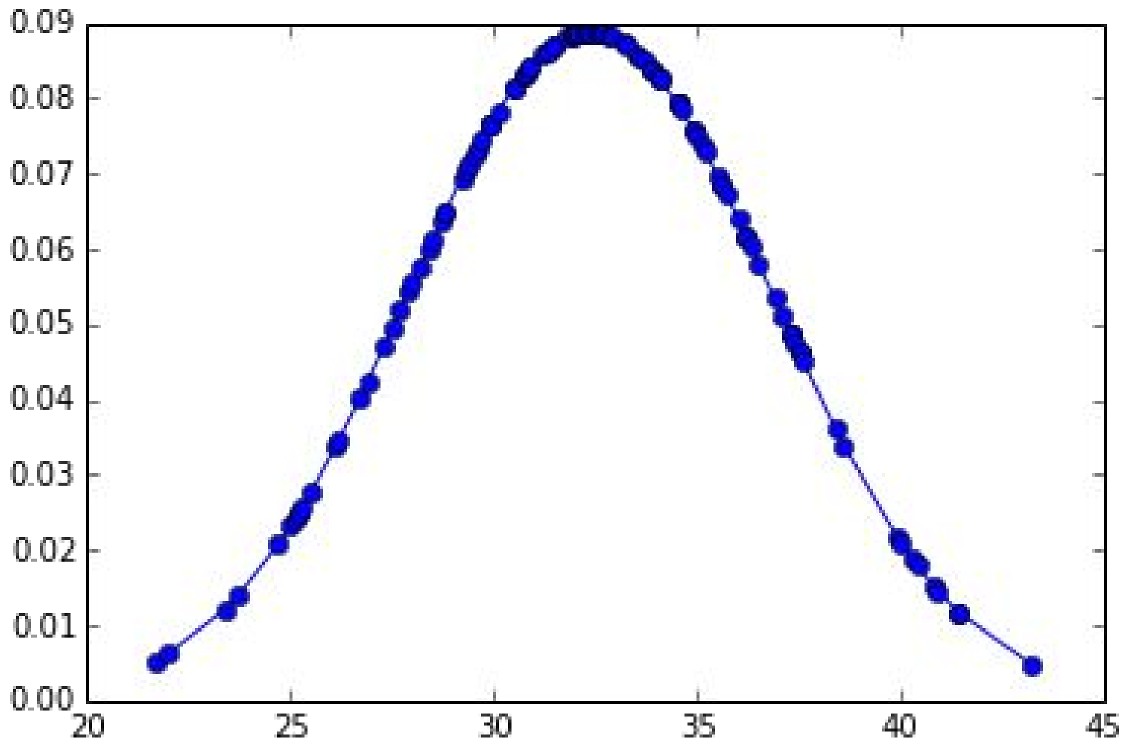
$$f_X(x) \geq 0, \text{for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \quad \text{and}$$

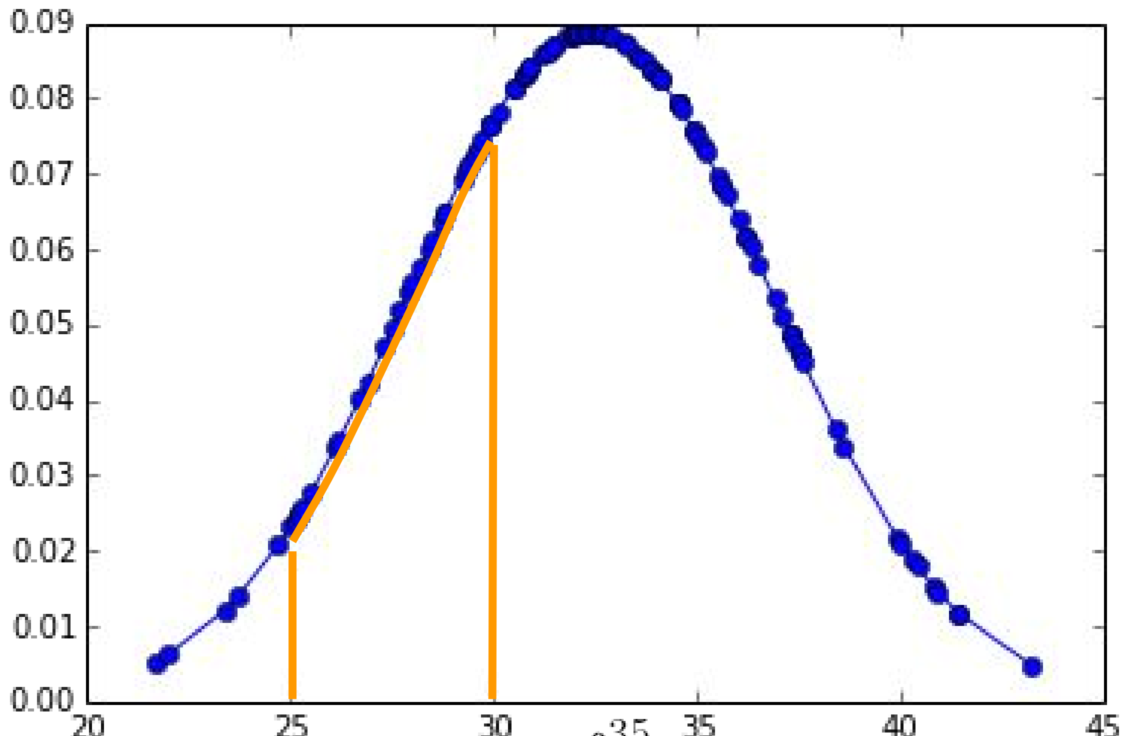
$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

***fx* : "probability density function" (pdf)**

Continuous Random Variables



Continuous Random Variables



$$P(25 < X < 35) = \int_{25}^{35} fx(x)dx$$

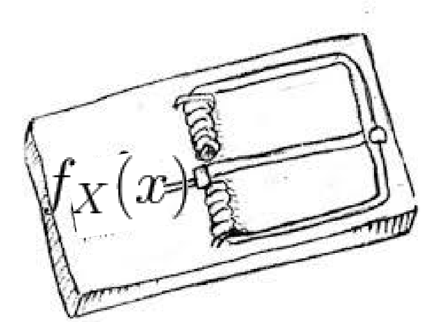
CRV Review: 2-4

- Concept of PDF
- Formal definition of a pdf
- How to create a continuous random variable in python
- Plot Histograms
- Plot PDFs

Continuous Random Variables

Common Trap

- $f_X(x)$ does not yield a probability
  - $\int_a^b f_X(x)dx$ does
  - x may be anything ($\mathbb{R}$)
    - thus, $f_X(x)$ may be > 1



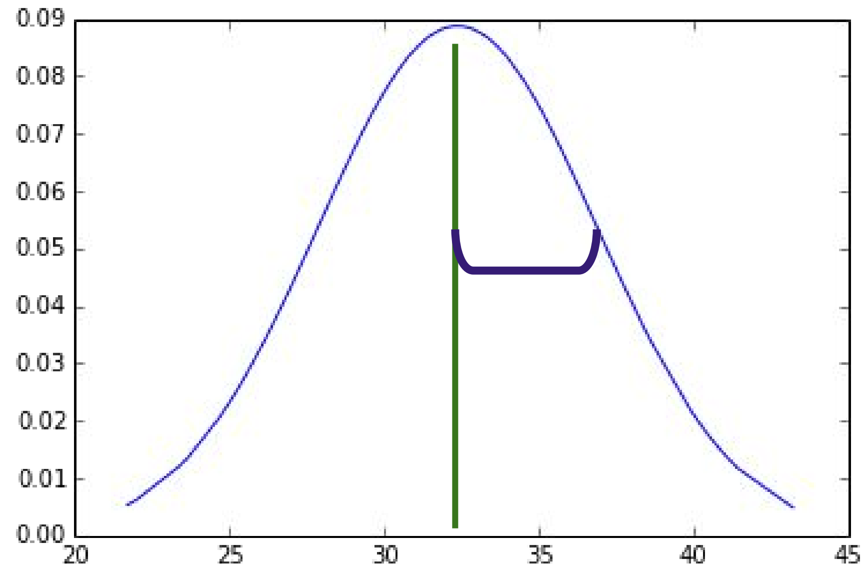
Continuous Random Variables

Some Common Probability Density Functions

Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

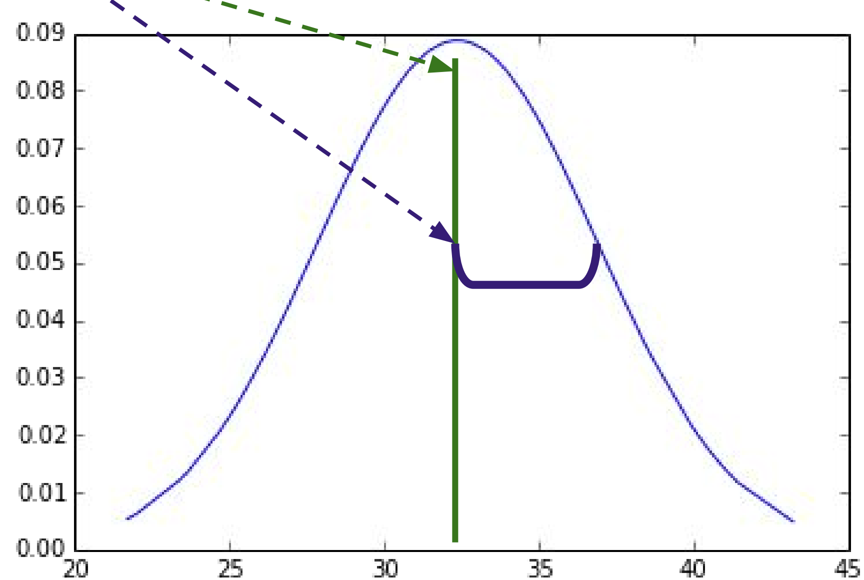
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

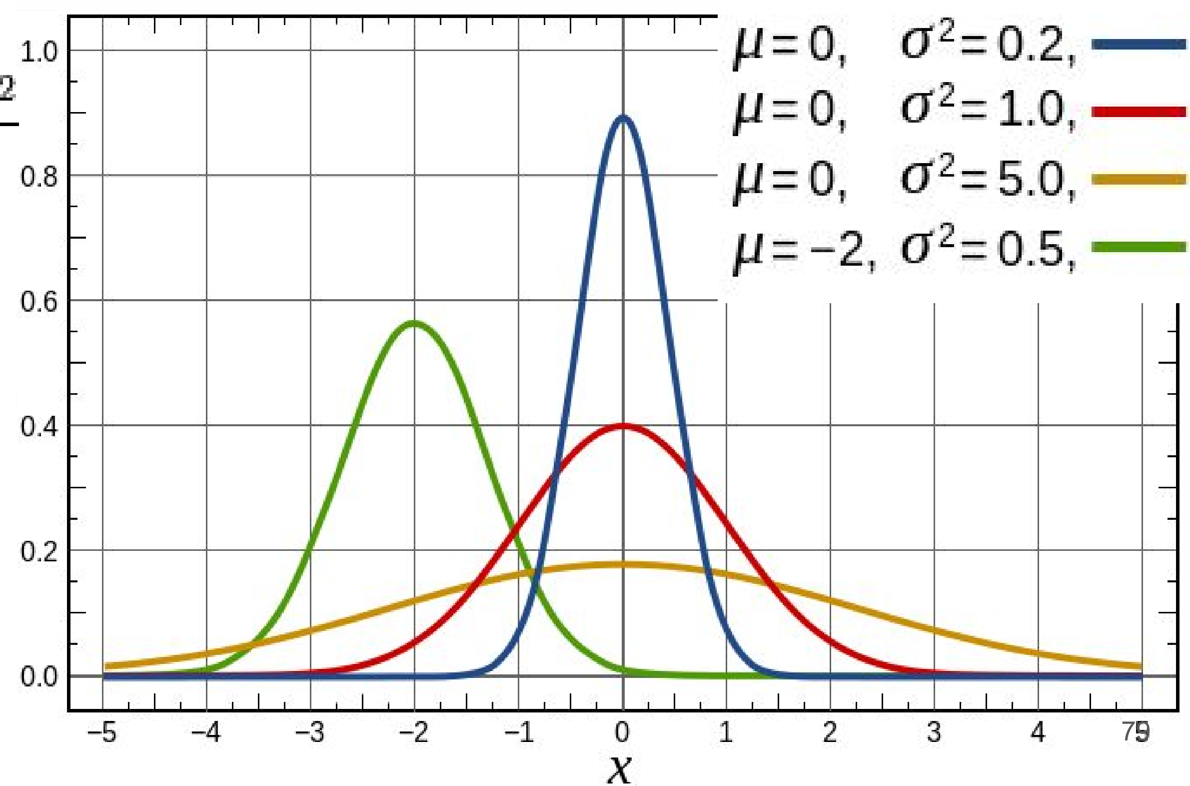
σ: standard deviation

Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

Credit: Wikipedia

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

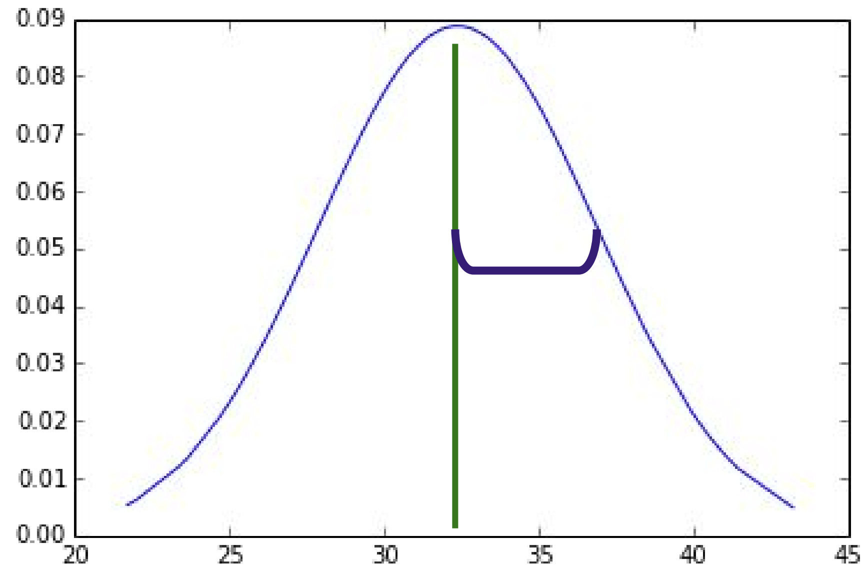
σ: standard deviation

Continuous Random Variables

Common *pdfs*: $\mathrm{Normal}(\mu, \sigma^2)$

$X \sim \mathrm{Normal}(\mu, \sigma^2)$, examples:

- height
- intelligence/ability
- **measurement error**
- averages (or sum) of
  lots of random variables



Continuous Random Variables

Common *pdf*s: Normal(0, 1) ("standard normal")

How to "standardize" any normal distribution:

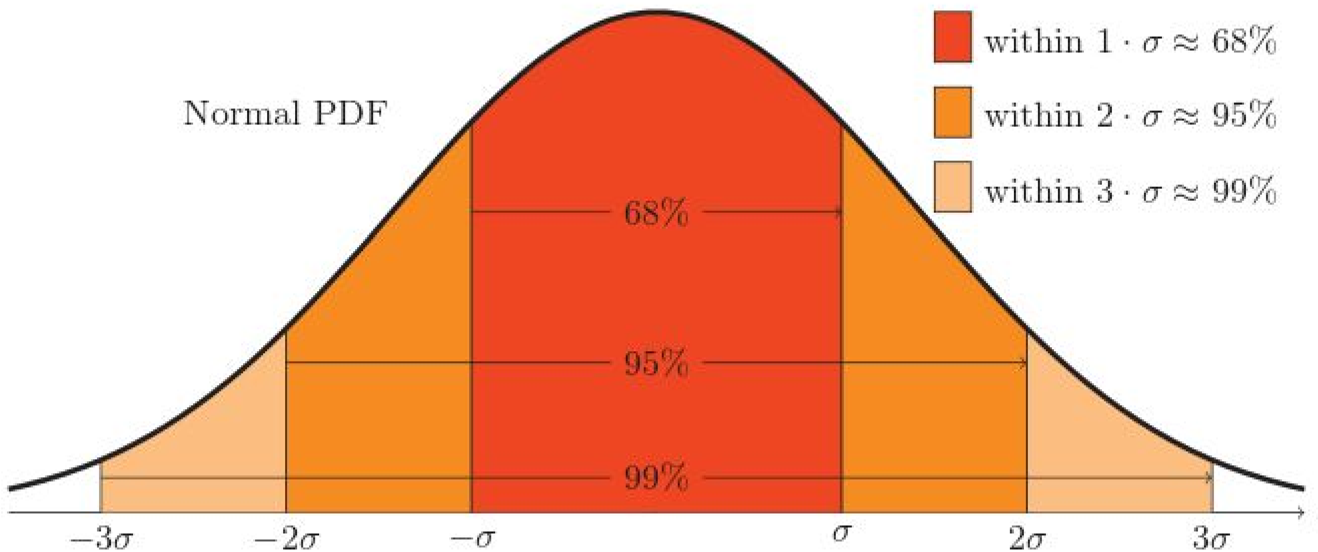
- subtract the mean, μ (aka "mean centering")
- divide by the standard deviation, σ

$z = (x - \mu) / \sigma$, (aka "z score")

Continuous Random Variables

Common *pdfs*: $\text{Normal}(0, 1)$

$$P(-1 \le Z \le 1) \approx .68, \quad P(-2 \le Z \le 2) \approx .95, \quad P(-3 \le Z \le 3) \approx .99$$

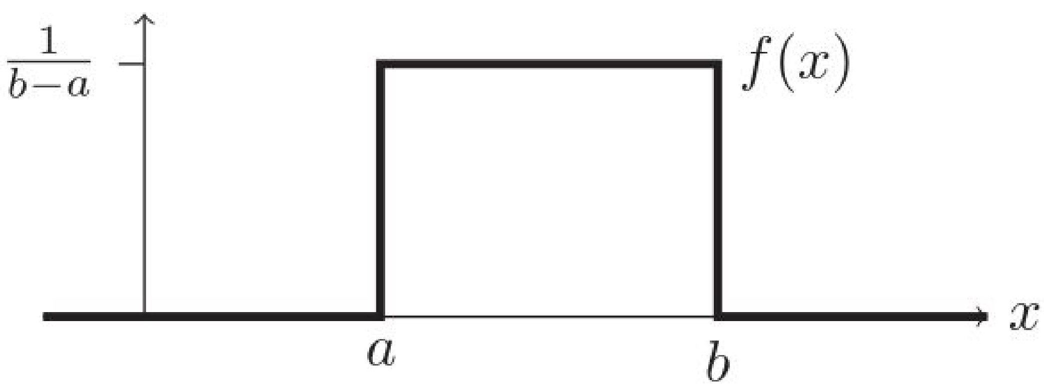


Credit: MIT Open Courseware: Probability and Statistics

Continuous Random Variables

Common *pdfs*: Uniform(a, b)

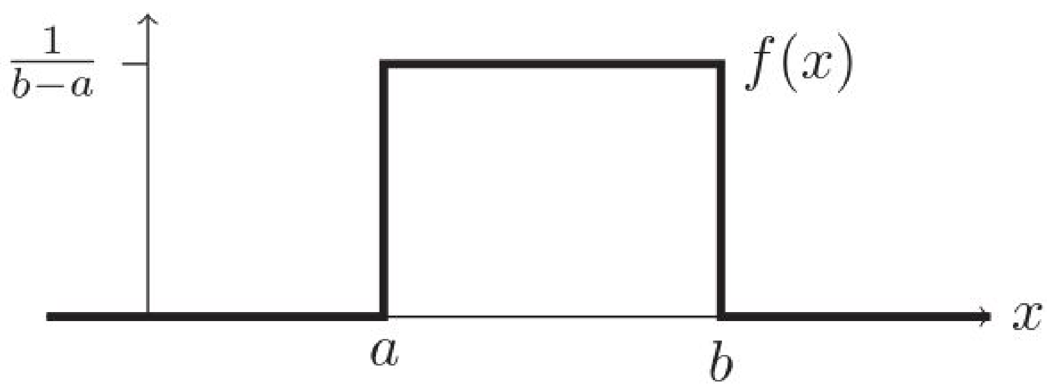
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



Continuous Random Variables

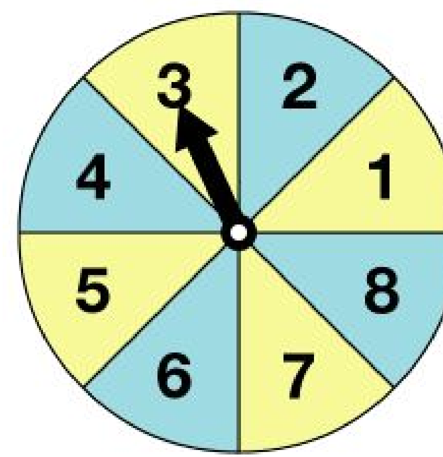
Common *pdfs*: Uniform(a, b)

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



$X \sim \text{Uniform}(a, b)$, examples:

- spinner in a game
- random number generator
- analog to digital rounding error



Continuous Random Variables

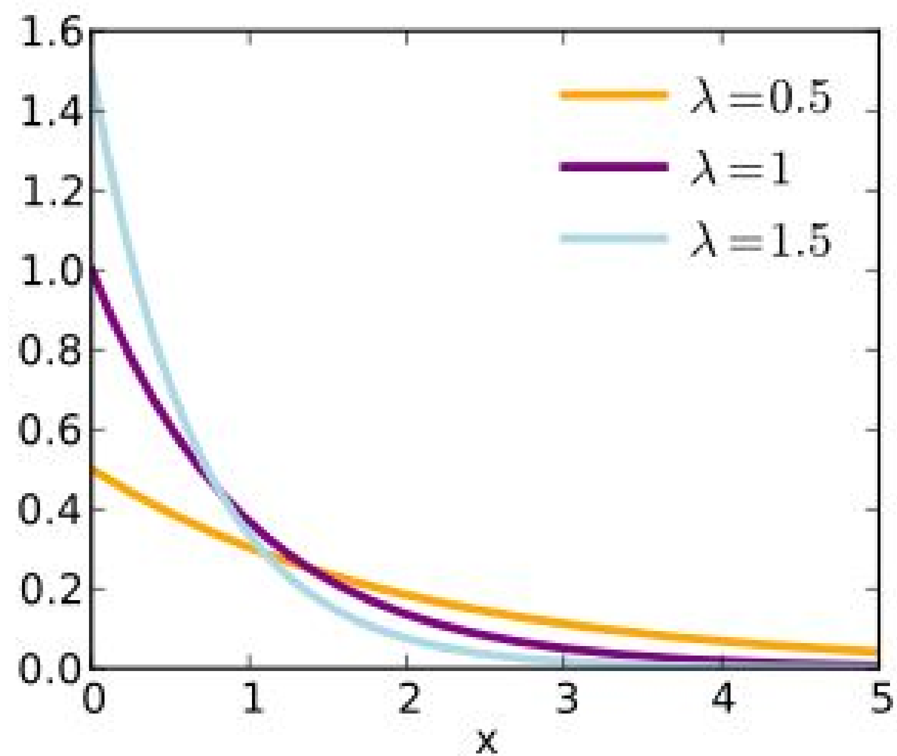
Common *pdfs*: Exponential(λ)

$$f_X(x) = \lambda e^{-\lambda x}, x > 0$$

λ: rate or inverse scale

β: scale $\left(\lambda = \frac{1}{\beta} \right)$

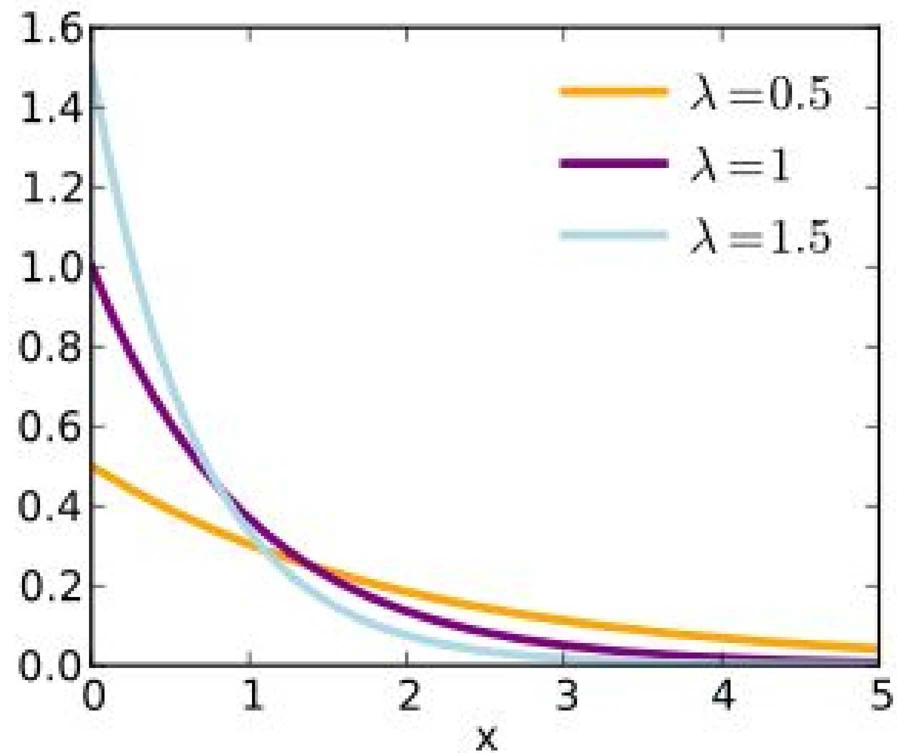
Credit: Wikipedia



Continuous Random Variables

Common *pdf*s: Exponential(λ)

Credit: Wikipedia



$X \sim \text{Exp}(\lambda)$, examples:

- lifetime of electronics
- waiting times between rare events (e.g. waiting for a taxi)
- recurrence of words across documents

Continuous Random Variables

How to decide which pdf is best for my data?

Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- Kernel Density Estimator

Continuous Random Variables

How to decide which pdf is best for my data?

Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- **Kernel Density Estimator**

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

K: kernel function, *h:* bandwidth

(for every data point, draw *K* and add to density)

Continuous Random Variables

How to decide which pdf is best for my data?

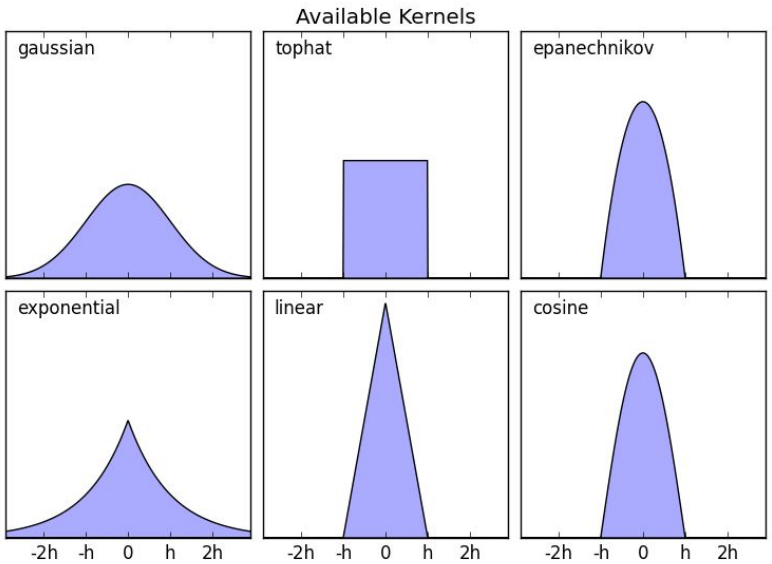
Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- **Kernel Density Estimator**

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

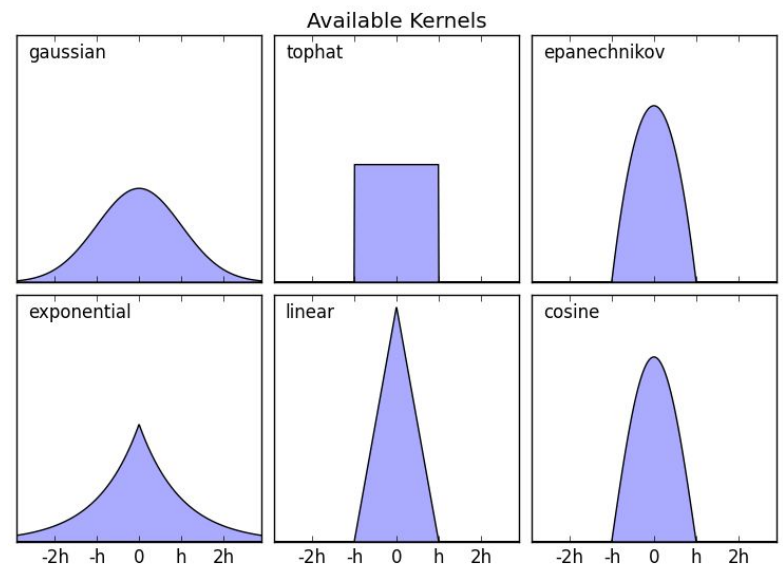
K: kernel function, *h:* bandwidth

(for every data point, draw *K* and add to density)



Continuous Random Variables

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

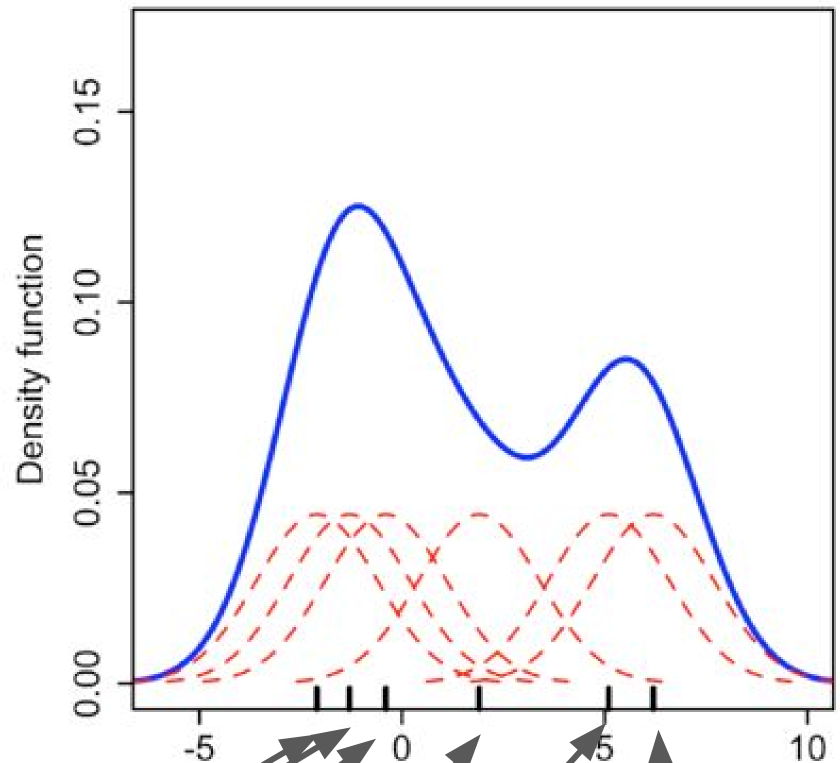


Continuous Random Variables

just like a pdf, this function takes in an x and returns the appropriate y on an estimated distribution curve

to figure out y for a given x, take the sum of where each where each kernel (a density plot for each data point in the original X) puts that x.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

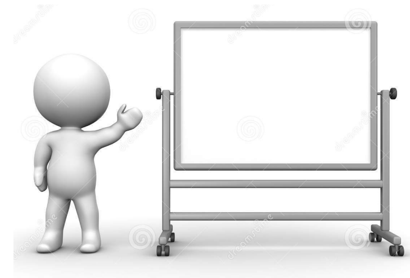


Continuous Random Variables

Analogies

- Funky dartboard

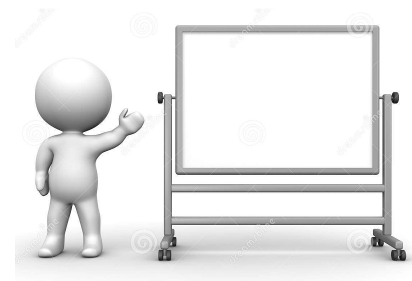
Credit: MIT Open Courseware: Probability and Statistics



Continuous Random Variables

Analogies

- Funky dartboard



- Random number generator

Cumulative Distribution Function

- Random number generator

Cumulative Distribution Function

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

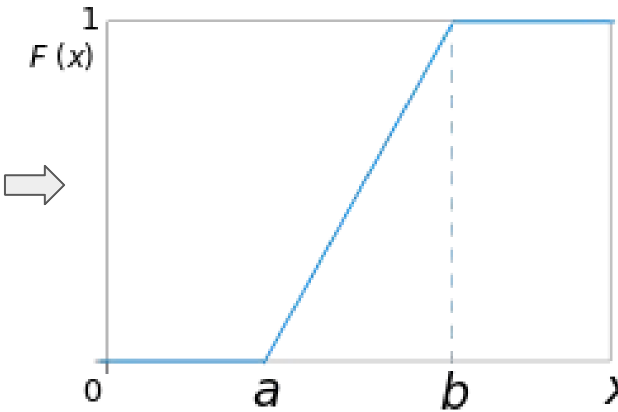
$$F_X(x) = \mathrm{P}(X \leq x)$$

Cumulative Distribution Function

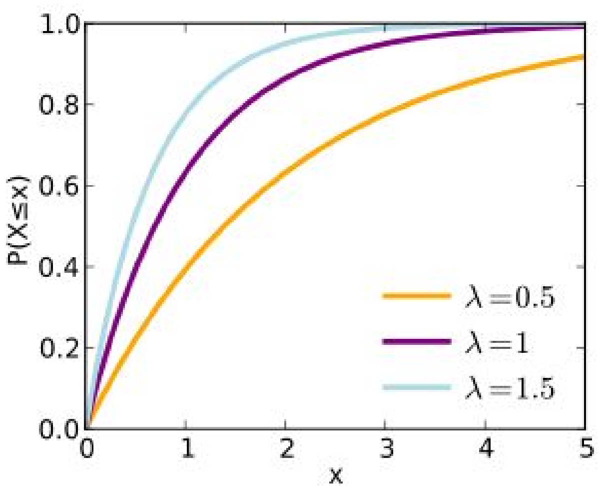
For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

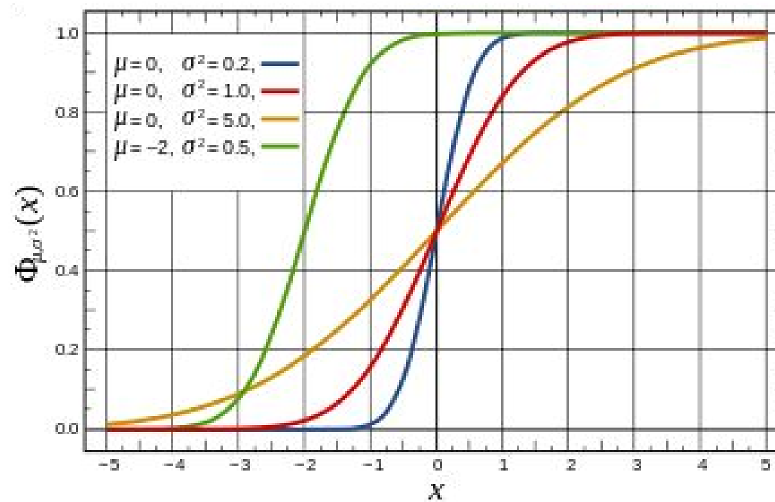
Uniform



Exponential



Normal

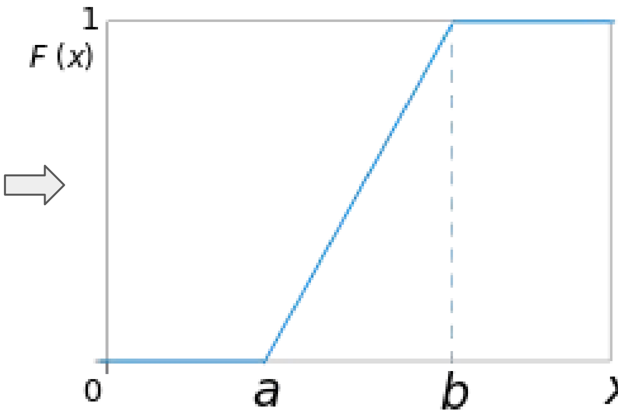


Cumulative Distribution Function

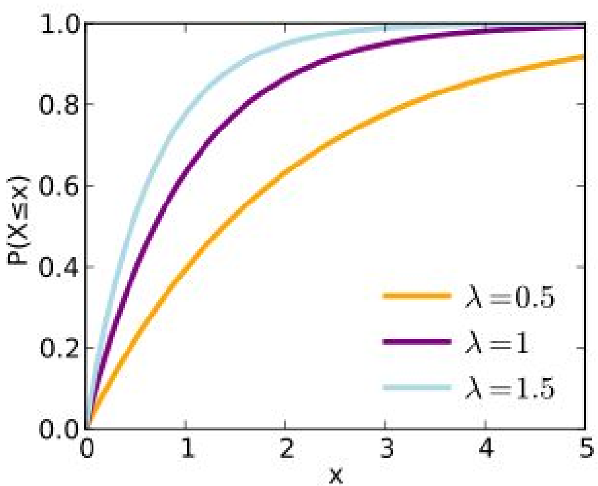
For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

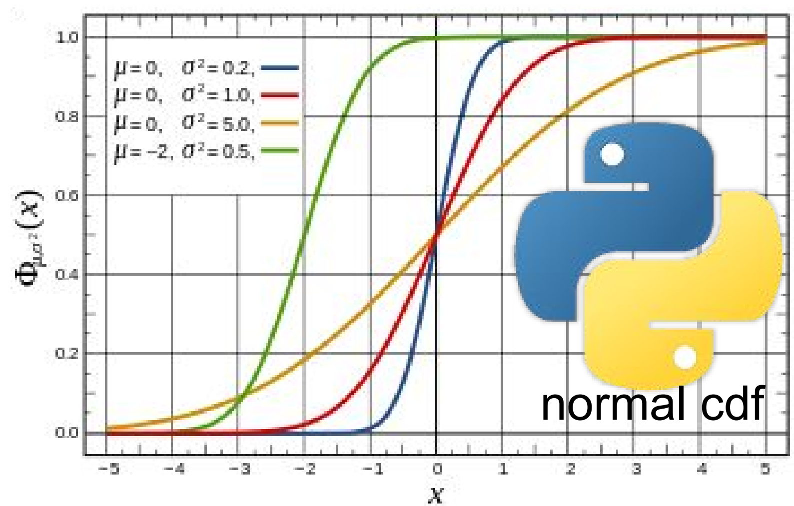
Uniform



Exponential



Normal

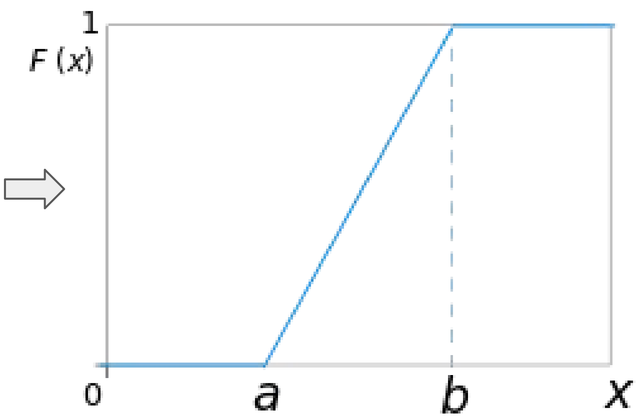


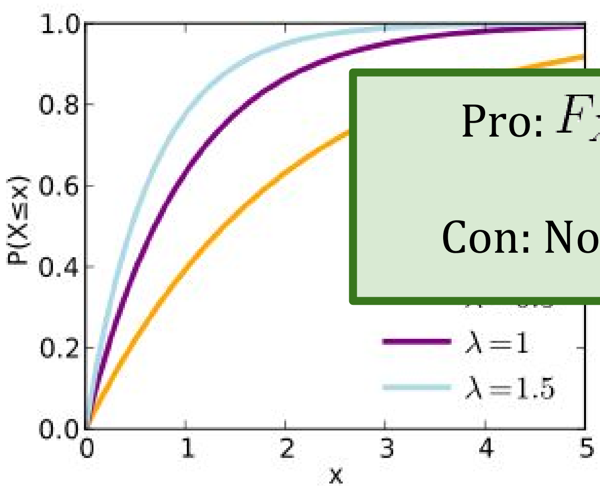
Cumulative Distribution Function

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \to [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \le x)$$

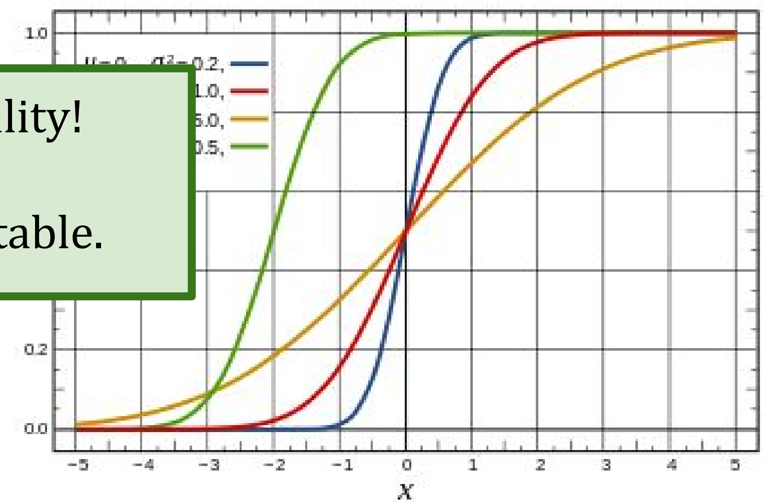
Uniform ⇨





Pro: $F_X(x)$ yields a probability!

Con: Not intuitively interpretable.



Random Variables, Revisited

$\mathbf{X}$: A mapping from Ω to $\mathbb{R}$ that describes the question we care about in practice.

$\mathbf{X}$ is a *continuous random variable* if it can take on an infinite number of values between any two given values.

$\mathbf{X}$ is a *discrete random variable* if it takes only a countable number of values.

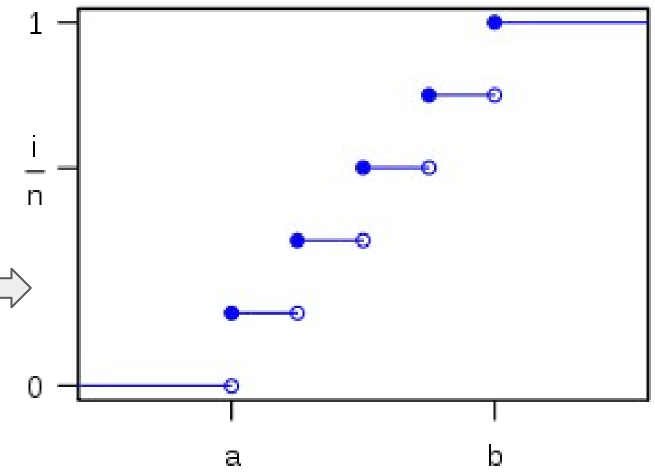
Discrete Random Variables

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \to [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

X is a *discrete random variable* if it takes only a countable number of values.

Discrete Random Variables

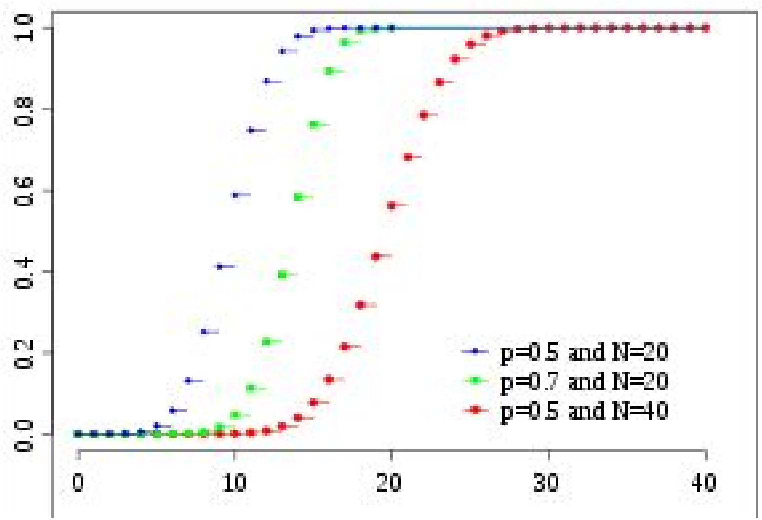


Discrete Uniform

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \to [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

X is a *discrete random variable* if it takes only a countable number of values.



Binomial (n, p)

(like normal)

Discrete Random Variables

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

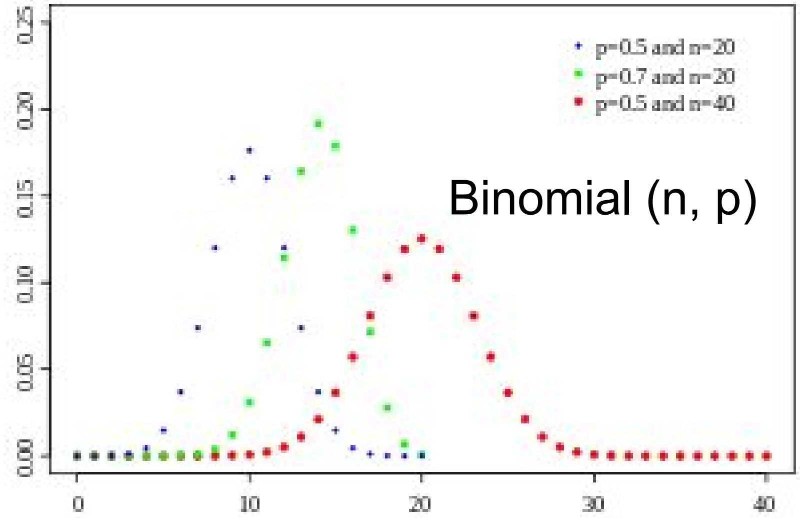
$$F_X(x) = \mathrm{P}(X \leq x)$$

For a given discrete random variable X, *probability mass function* (*pmf*), *fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$f_X(x) = \mathrm{P}(X = x)$$

X is a *discrete random variable* if it takes only a countable number of values.

Discrete Random Variables



For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

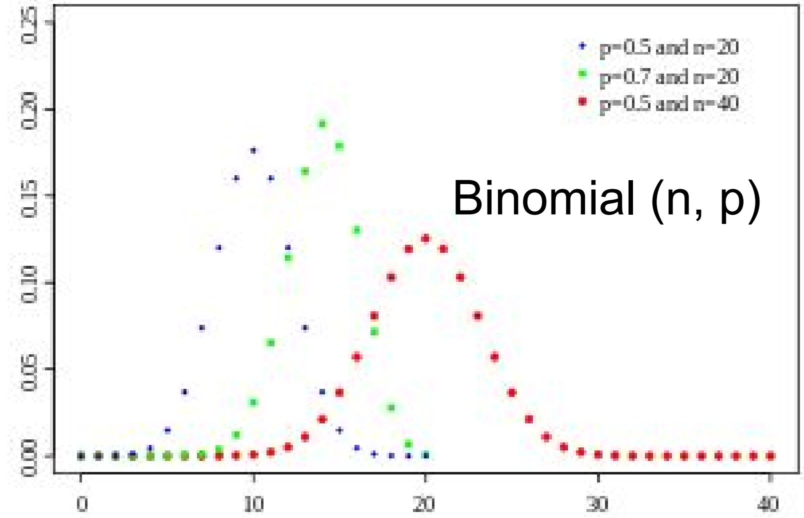
$$F_X(x) = \mathrm{P}(X \leq x)$$

For a given discrete random variable X, *probability mass function* (*pmf*), *fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$f_X(x) = \mathrm{P}(X = x)$$

X is a *discrete random variable* if it takes only a countable number of values.

Discrete Random Variables



For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

For a given discrete random variable X, *probability mass function* (*pmf*), *fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

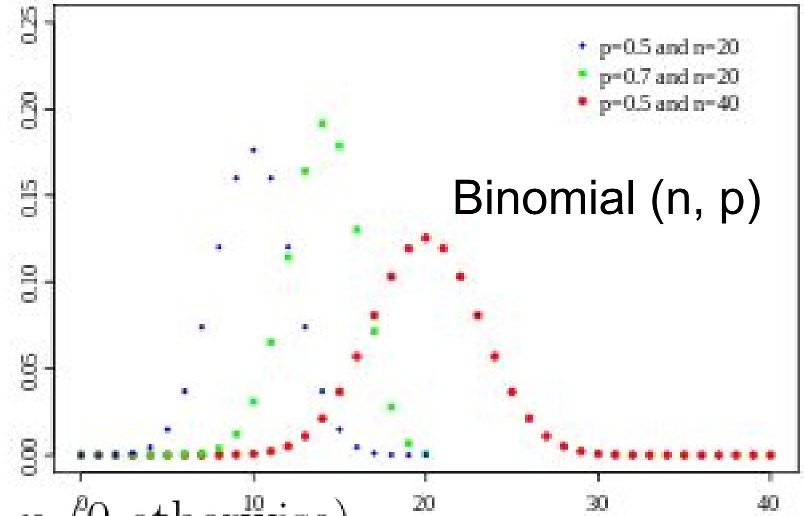
$$f_X(x) = \mathrm{P}(X = x)$$

X is a *discrete random variable* if it takes only a countable number of values.

$$\sum_i f_X(x) = 1$$

$$F_X(f) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} f_X(x)$$

Discrete Random Variables



Common Discrete Random Variables

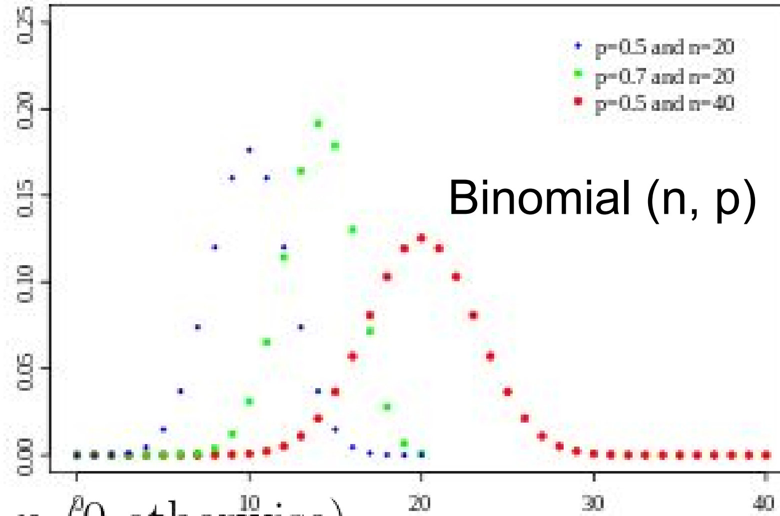
- Binomial(n, p)

  $$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } 0 \leq x \leq n \text{ (0 otherwise)}$$

  example: number of heads after n coin flips (p, probability of heads)
- Bernoulli(p) = Binomial(1, p)

  example: one trial of success or failure

Discrete Random Variables



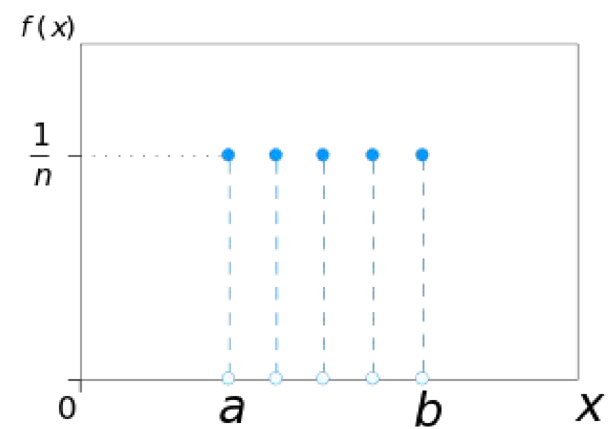
Common Discrete Random Variables

- Binomial(n, p)

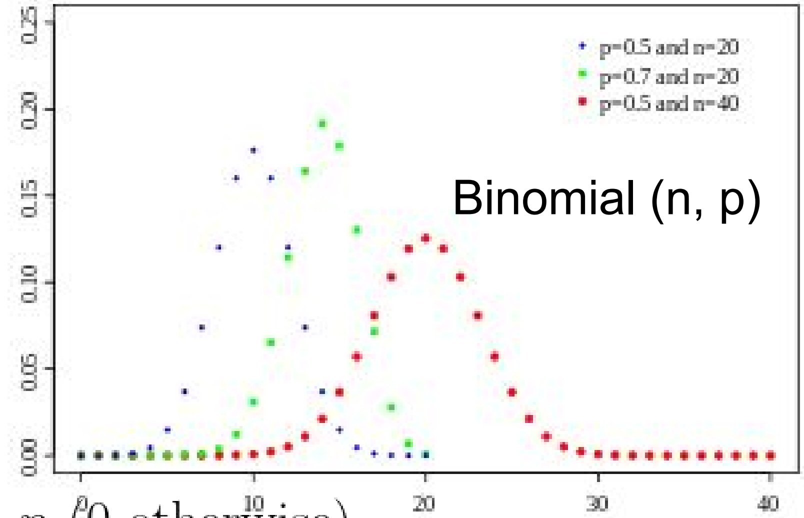
  $$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } 0 \le x \le n \text{ (0 otherwise)}$$

  example: number of heads after n coin flips (p, probability of heads)
- Bernoulli(p) = Binomial(1, p)

  example: one trial of success or failure
- Discrete Uniform(a, b)



Discrete Random Variables



Common Discrete Random Variables

- Binomial(n, p)

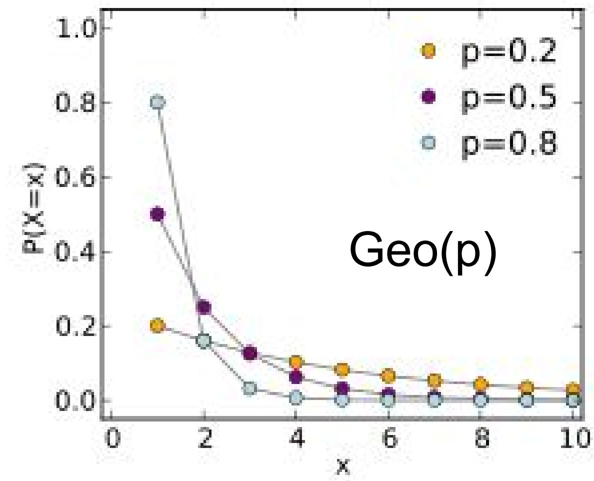
  $$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } 0 \le x \le n \text{ (0 otherwise)}$$

  example: number of heads after n coin flips (p, probability of heads)
- Bernoulli(p) = Binomial(1, p)

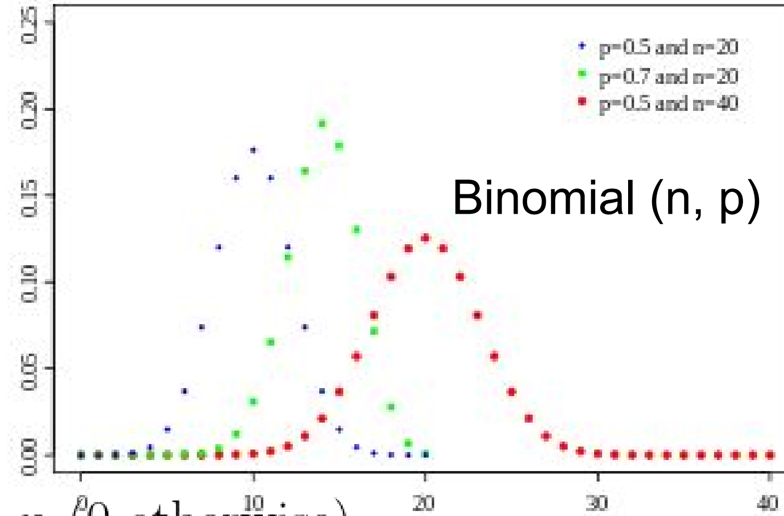
  example: one trial of success or failure
- Discrete Uniform(a, b)
- Geometric(p)

  $P(X = k) = p(1 - p)^{k-1}, \; k \geq 1$

  example: coin flips until first head



Discrete Random Variables



Common Discrete Random Variables

- Binomial(n, p)

  $$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } 0 \le x \le n \text{ (0 otherwise)}$$

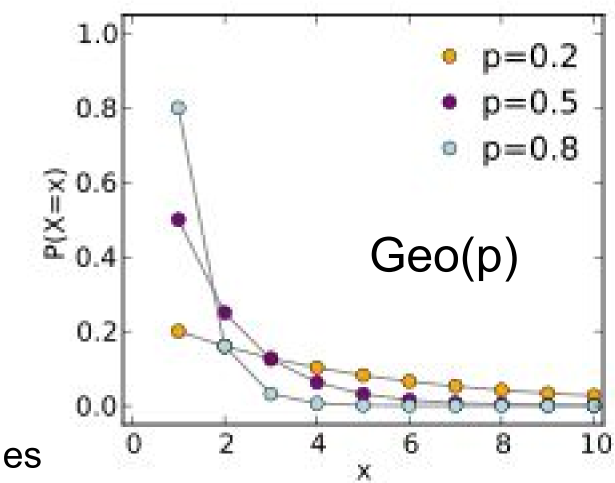
  example: number of heads after n coin flips (p, probability of heads)
- Bernoulli(p) = Binomial(1, p)

  example: one trial of success or failure
- Discrete Uniform(a, b)
- Geometric(p)

  $P(X = k) = p(1 - p)^{k-1},\ k \geq 1$

  example: coin flips until first head

discrete random variables



Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X_1, X_2, \dots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X_1, X_2, \ldots, X_n$ ~ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

$$l_n(p) = S\, log\, p + (n - s)\, log(1 - p)$$

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{where } S = \sum_i X_i$$

$$l_n(p) = S\, log\, p + (n - s)\, log(1 - p)$$

take the derivative and set to 0 to find: $\hat{p} = \frac{S}{n}$

Probability Theory Review: 2-11

- common pdfs: Normal, Uniform, Exponential
- how does kernel density estimation work?
- common pmfs: Binomial (Bernoulli), Discrete Uniform, Geometric
- cdfs (and how to transform out from a random number generator (i.e. uniform distribution) into another distribution)
- how to plot: pdfs, cdfs, and pmfs in python.
- MLE revisited: how to derive the parameter estimate from the likehood function

Maximum Likelihood Estimation (parameter estimation)

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X_1, X_2, \dots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

$$l_n(p) = S\, log\, p + (n - s)\, log(1 - p)$$

take the derivative and set to 0 to find: $\hat{p} = \frac{S}{n}$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X \sim$ Normal(μ, σ), then

GOAL: take the derivative and set to 0 to find:

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim \text{Normal}(\mu, \sigma)$, then

$$f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$\boxed{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Normal pdf

GOAL: take the derivative and set to 0 to find:

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $$f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}}$$

GOAL: take the derivative and set to 0 to find:

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}}$

$$log(f(x_1, \ldots, x_n; \mu, \sigma)) = -n \log \sqrt{2\pi} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

GOAL: take the derivative and set to 0 to find:

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$

$$log(f(x_1, \ldots, x_n; \mu, \sigma)) = -n \log \sqrt{2\pi} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

first, we find μ using partial derivatives:

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{2\sigma^2} = 0$$

GOAL: take the derivative and set to 0 to find:

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$

$$log(f(x_1, \ldots, x_n; \mu, \sigma)) = -n \log \sqrt{2\pi} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

first, we find μ using partial derivatives:

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{2\sigma^2} = 0 \ , \quad \sum_{i=1}^{n} x_i - n\mu = 0 \ , \quad \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} = \bar{x}$$

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $$f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$log(f(x_1, \ldots, x_n; \mu, \sigma)) = -n \log \sqrt{2\pi} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

first, we find μ using partial derivatives*:*

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{2\sigma^2} = 0 \quad , \quad \sum_{i=1}^{n} x_i - n\mu = 0 \quad , \quad \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} = \bar{x}$$

now σ:

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $f(x_1, x_2, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$

$$log(f(x_1, \ldots, x_n; \mu, \sigma)) = -n \log\sqrt{2\pi} - n\log\sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

first, we find μ using partial derivatives:

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{2\sigma^2} = 0 \quad , \quad \sum_{i=1}^{n} x_i - n\mu = 0 \quad , \quad \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} = \bar{x}$$

now σ:

$$\frac{\partial f(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \quad , \quad \frac{n}{\sigma^2} + \sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad , \quad \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \hat{\sigma}^2$$

Maximum Likelihood Estimation

Example: $X \sim$ Normal(μ, σ), then $f(x_1, x_2, \dots, x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$

$$log(f(x_1, \dots, x_n; \mu, \sigma)) = -n\log\sqrt{2\pi} - n\log\sigma - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

first, we find μ using partial derivatives:

sample mean

$$\frac{\partial f(x_1, \dots, x_n; \mu, \sigma)}{\partial\mu} = \sum_{i=1}^{n}\frac{(x_i-\mu)}{2\sigma^2} = 0 \quad , \quad \sum_{i=1}^{n} x_i - n\mu = 0 \quad , \quad \boxed{\frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} = \bar{x}}$$

now σ:

$$\frac{\partial f(x_1, \dots, x_n; \mu, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} = 0 \quad , \quad \frac{n}{\sigma^2} + \sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad , \quad \boxed{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \hat{\sigma}^2}$$

sample variance

Maximum Likelihood Estimation

Try yourself:

Example: $X \sim$ Exponential(λ),

hint: should arrive at something almost familiar; then recall $\lambda = \dfrac{1}{\beta}$

Expectation, revisited

Conceptually: Just given the distribution and no other information: what value should I expect?

Expectation, revisited

Conceptually: Just given the distribution and no other information: what value should I expect?

Formally: The ***expected value*** of X is:

$$\mathbf{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

denoted: $\mathbf{E}(X) = \mathbf{E}X = (x) = \mu = \mu x$

Expectation, revisited

Conceptually: Just given the distribution and no other information: what value should I expect?

Formally: The ***expected value*** of X is:

$$\mathbf{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

denoted: $\mathbf{E}(X) = \mathbf{E}X = (x) = \mu = \mu x$

"expectation" "mean" "first moment"

Expectation, revisited

Conceptually: Just given the distribution and no other information: what value should I expect?

Formally: The ***expected value*** of X is:

$$\mathbf{E}(X) = \int x\,dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

denoted: $\mathbf{E}(X) = \mathbf{E}X = (x) = \mu = \mu x$

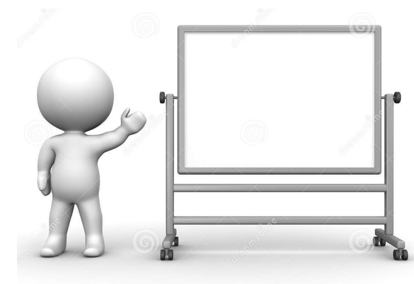
"expectation" "mean" "first moment"

Alternative Conceptualization: If I had to summarize a distribution with only one number, what would do that best?

(the average of a large number of randomly generated numbers from the distribution)

Expectation, revisited

Examples:



$X \sim$ Bernoulli(p):

$X \sim$ Uniform(-3,1):

The ***expected value*** of X is:

$$\mathbf{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

denoted: $\mathbf{E}(X) = \mathbf{E}X = (x) = \mu = \mu x$

Probability Theory Review: 2-16

- MLE over a continuous random variable
- mean and variance
- The concept of expectation
- Calculating expectation for
  - discrete variables
  - continuous variables