Motivation

Supervised Learning: Learning from a randomly sampled batch of labeled data

Dataset (tuple of tuples) $\xrightarrow{\text{Learner (function)}}$ Predictor = Model = Hypothesis (function)

$$D = \left((\vec{X}_1, Y_1), (\vec{X}_2, Y_2), \dots, (\vec{X}_n, Y_n)\right) \in (\mathcal{X} \times \mathcal{Y})^n$$

$(\vec{X}_i, Y_i) \sim \mathbb{P}_{\vec{X},Y}$ independent for all $i \in \{1, \dots, n\}$

D — n feature-label pairs

$\mathcal{X}$ — set of d-dimensional features

$\mathcal{Y}$ — set of labels/targets

$$f : \mathcal{X} \to \mathcal{Y}$$

f a function from features to labels

Ex $f(x) = 200x + 100$, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$$

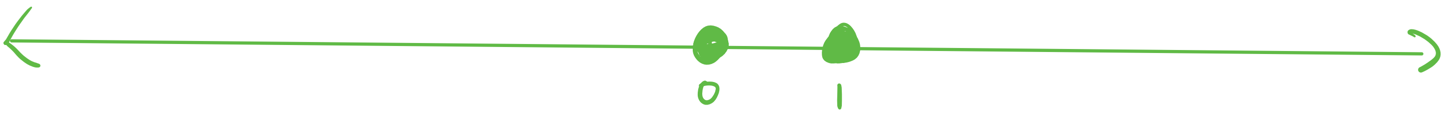
$\mathcal{A}$ a function from datasets to predictors

Ex: $\mathcal{A}(D) = \hat{f}$ where $f : \mathcal{X} \to \mathcal{Y}$

$$\hat{f}(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i \in \{1, \dots, n\} \\ & \text{Pick } i \text{ to be lowest index} \\ 0 & \text{otherwise} \end{cases}$$

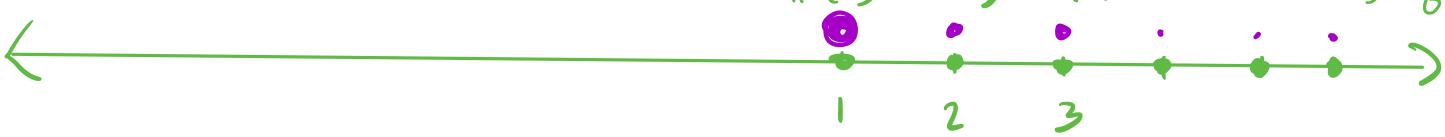
$\{0,1\}$ $\qquad \mathbb{P}(0)+\mathbb{P}(1)=1$

$\mathbb{P}(0)=\frac{1}{2} \qquad \mathbb{P}(1)=\frac{1}{2}$



$X \in \overbrace{\{1,2,3,4,5,6\}}^{\text{outcomes}} \qquad \sum_{i=1}^{6} \mathbb{P}(\{i\})=1$

$\overset{3/6 \, =}{\mathbb{P}(\{1\})} = \overset{1/6 \, =}{\mathbb{P}(\{2\})} = \overset{1/6 \, =}{\mathbb{P}(\{3\})} = \overset{1/18 \, = \quad 1/18 \, =}{\cdots} = \overset{1/18 \, =}{\mathbb{P}(\{6\})} = \frac{1}{6}$



$\mathbb{P}(\underbrace{\{1,2,3\}}_{\text{event}}) = \mathbb{P}(\{1\}\cup\{2\}\cup\{3\}) = \mathbb{P}(\{1\})+\mathbb{P}(\{2\})+\mathbb{P}(\{3\})$

$\mathbb{P}(X=3) \overset{\text{def}}{=} \mathbb{P}(\{3\})$

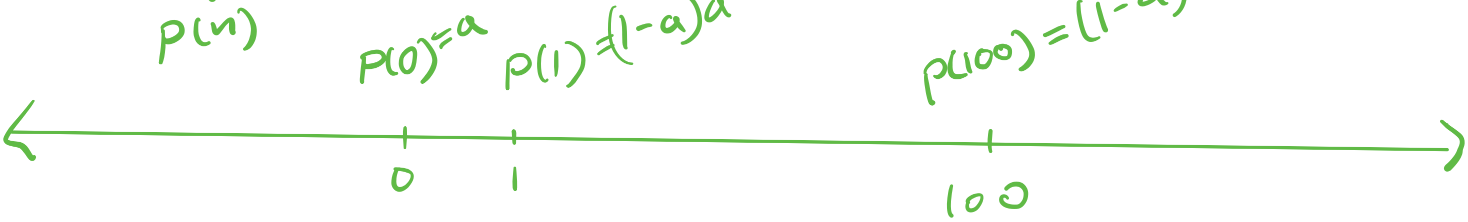
$\mathbb{P}(X\le 3) = \mathbb{P}(\{1,2,3\}) = \mathbb{P}(X=1)+\mathbb{P}(X=2)+\mathbb{P}(X=3)$

$= p(1)+p(2)+p(3)$

$\mathbb{P}(X=x)=p(x)$ probability mass function

$X \in \{0, 1, 2, \ldots\} = \mathbb{N}$ $\quad \sum_{n \in \mathbb{N}} p(n) = \sum_{n=0}^{\infty} p(n) = p(0) + p(1) + p(2) + \ldots = 1$

$\underbrace{\sum_{n=0}^{\infty} (1-a)\, a^{(n+1)}}_{p(n)} = 1 \qquad a < 1$

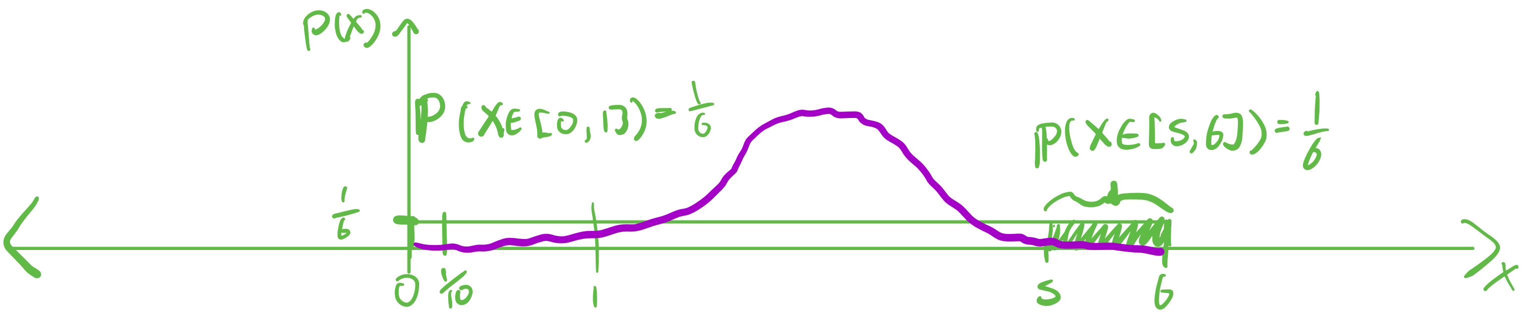


$\mathbb{P}(X \le 100) = \sum_{n=0}^{100} p(n)$

if outcome space of r.v. X is countable (finite / countably infinite)

then X is called a discrete r.v.

$X \in [0, 6]$ $\quad \sum_{n \in \mathbb{R}} p(n) = 1 \qquad p(n) > 0$ for all $n \in \mathbb{R}$



$\mathbb{P}(X \in [0, \tfrac{1}{10}]) = \frac{1}{60} \qquad p(x \in [a,b]) = \frac{1}{c}$

$\mathbb{P}(X = x) = 0 \qquad \mathbb{P}(X \in [0,6]) = 1 = \int_0^6 p(x)\,dx$

$\mathbb{P}(X \in [a,b]) = \int_a^b p(x)\,dx \qquad = \int_0^6 \frac{1}{6}\,dx = \frac{1}{6} x \Big|_0^6 = \frac{1}{6} \cdot (6-0) = 1$

$\mathbb{P}(X = 4) \longrightarrow + \int_4^4 p(x)\,dx = 0$

$\mathbb{P}(X \in [0,1] \cup [5,6] \cup \{4\}) = \int_{x \in [0,1] \cup [5,6]} p(x)\,dx = \int_0^1 p(x)\,dx + \int_5^6 p(x)\,dx = \frac{1}{6} + \frac{1}{6}$

Probability

Note: Humans have a bad intuition when it comes to randomness
- Thinking Fast and Slow by: Daniel Kahneman

- Outcomes, Events, Distributions
- Random Variables
- Calculating probabilities using pmf and pdf
- Multivariate random variables
  - Conditional and marginal probabilities
- Representing random features, labels, and datasets
- Functions of random variables
- Expectation and variance

Warning: If some things seem informal, it is likely because we would need tools from Measure Theory, which we will not cover in this course.

Experiment: A process that generates an uncertain outcome

Ex: flipping a coin, rolling a dice

Outcome Space/Set: The set of all outcomes from the experiment

Ex: $\mathcal{Y} = \{0, 1\}$ Heads, Tails — flipping a coin

$\mathcal{X} = \{1,2,3,4,5,6\}$ — rolling a dice

$[0, 900]$ — amount of a chemical in a wine

$\mathbb{R}$ — measurement error

Event: A subset of the outcome space (imprecise)

Ex: Outcome space: $\mathcal{Y} = \{0, 1\}$

Events: $\{0\}, \{1\}, \{0,1\} = \mathcal{Y}, \emptyset$

Ex: Outcome space: $\mathcal{X} = \{1,2,3,4,5,6\}$

Events: $\emptyset, \mathcal{X}, \{1\}, \{1,3,5\}, \{2,4,6\}, \{1,2,3\}, \ldots$

Ex: Outcome space: $[0, 900]$

Events: $\emptyset, [0,900], [0,4], [1,2] \cup [7,30], \ldots$

Probability Distribution: A function $\mathbb{P}$ defining the likelihood of each event (and satisfying certain properties)

$\mathbb{P}$: event space/set $\to [0,1]$

Think of this as a set containing all the events

A complicated set (σ-algebra) that we will not define

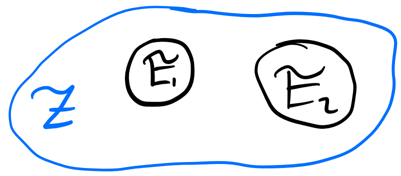
Properties: (imprecise)

Outcome space: $\mathcal{Z}$

1. $\mathbb{P}(\mathcal{Z}) = 1$ $\qquad \subset = \subseteq$

2. If $E_1 \subseteq \mathcal{Z}, E_2 \subseteq \mathcal{Z}$ and $E_1 \cap E_2 = \emptyset$, then

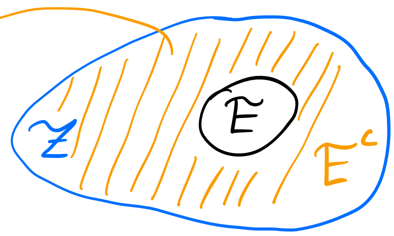
$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$$



Ex: (of property 2.)

Events: E, E^c

$E \cap E^c = \emptyset$, $E \cup E^c = \mathcal{Z}$



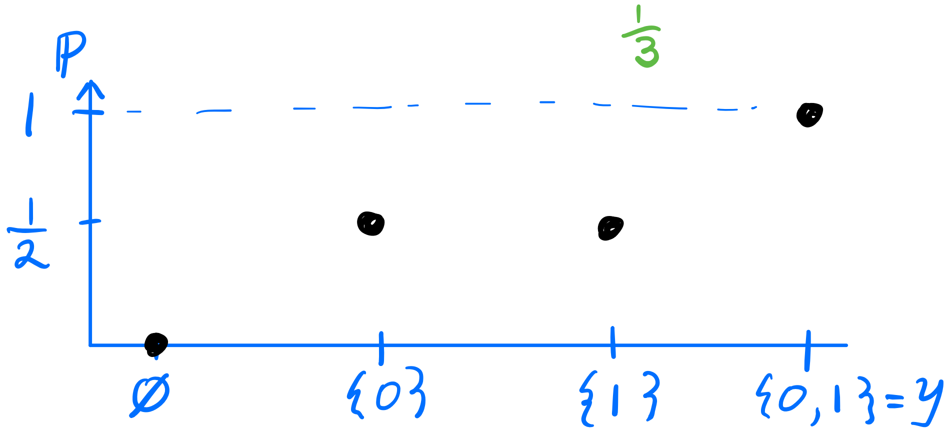
$E = \{1\}$ $\quad E^c = \{2,3,4,5,6\}$

$$\underbrace{\mathbb{P}(E \cup E^c)}_{= \mathbb{P}(\mathcal{Z}) = 1} = \mathbb{P}(E) + \mathbb{P}(E^c)$$

rearranging: $\boxed{\mathbb{P}(E) = 1 - \mathbb{P}(E^c)}$

Ex: Outcome space: $\mathcal{Y} = \{0,1\} = \{0\} \cup \{1\}$

$\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\mathcal{Y}) = 1$, $\mathbb{P}(\{0\}) = \frac{1}{2}$ ($\frac{1}{3}$), $\mathbb{P}(\{1\}) = \frac{1}{2}$ ($\frac{2}{3}$)



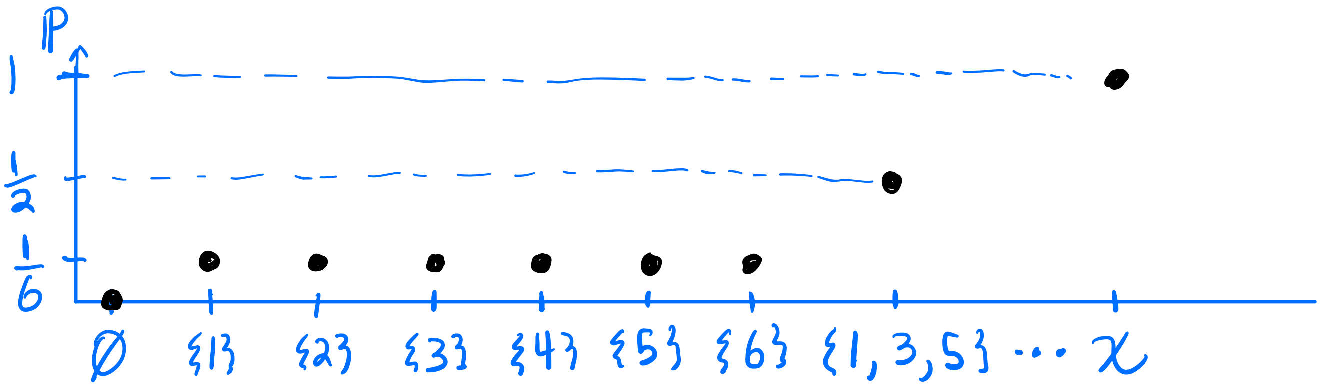
Ex: Outcome space: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$

$\mathbb{R}$

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \cdots = \mathbb{P}(\{6\}) = \frac{1}{6}$$

$$\underbrace{\{1\} \cup \{3\} \cup \{5\}}\ \mathbb{P}(\{1, 3, 5\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{5\}) = \frac{1}{2}$$

$$\mathbb{P}(\mathcal{X}) = 1$$



Random Variables

Random Variable (r.v.): (imprecise) A variable that takes a value based on the outcome of an experiment, and is associated with a probability distribution.

Ex: $X \in \mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ with $\mathbb{P}$ from prev. example

$Y \in \mathcal{Y} \in \{0, 1\}$ with $\mathbb{P}$ from prev. example

> A random variable is actually a function (satisfying certain properties) from one outcome space to another outcome space. Ex $X(T) = 0, X(H) = 1$.
> *It will not be necessary to know this for this course*

Probability Distributions with r.v.

Ex: Outcome space: $\mathcal{X}$ r.v.: $X \in \mathcal{X}$

$$\mathbb{P}(\{1,3,5\}) \stackrel{def}{=} \mathbb{P}(X \in \{1,3,5\})$$

$$\mathbb{P}(\{4,5,6\}) = \mathbb{P}(X \in \{4,5,6\}) = \mathbb{P}(X \geq 4)$$

$$\mathbb{P}(\{4\}) = \mathbb{P}(X \in \{4\}) = \mathbb{P}(X = 4)$$

Notation: $Z \sim \mathbb{P}$ "Z is sampled according to distribution $\mathbb{P}$"

Discrete r.v.: A r.v. that takes values from:

- A countable outcome space, or
- an uncountable outcome space, but there is a countable event that has probability 1

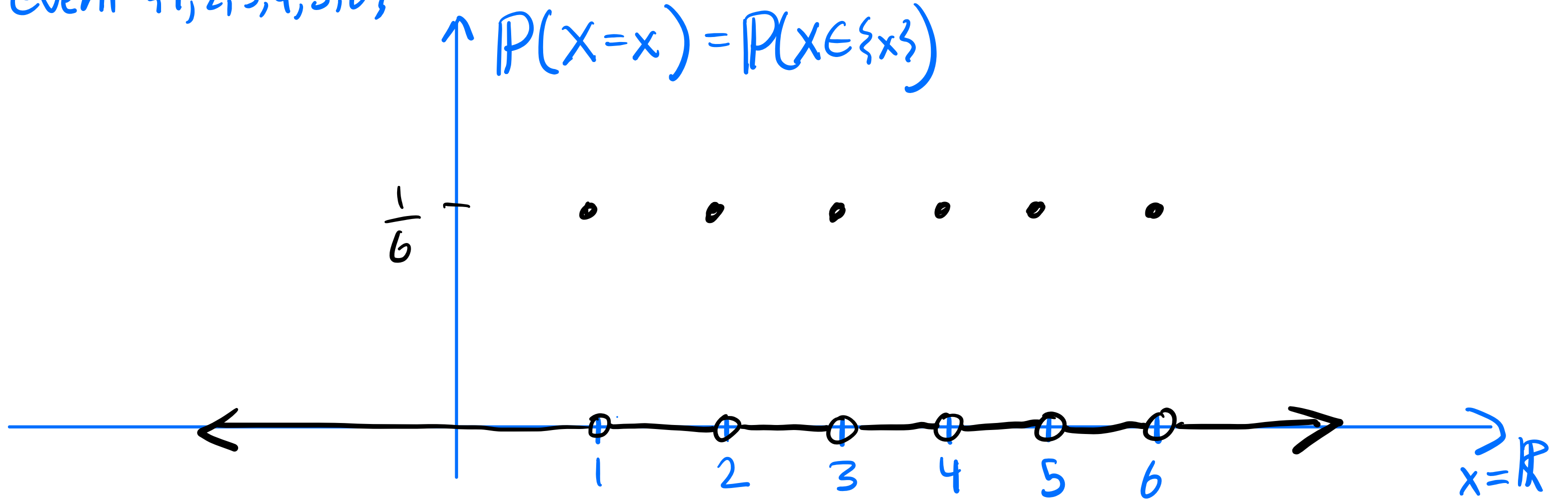
Ex: $Y \in \mathcal{Y} = \{0,1\}$, $X \in \mathcal{X} = \{1,2,3,4,5,6\}$, $Z \in \mathbb{N}$

Ex: $X \in \mathbb{R}$ where $\mathbb{P}(X=1) = \dots = \mathbb{P}(X=6) = \frac{1}{6}$

Probability 1 for countable event $\{1,2,3,4,5,6\}$

so $\mathbb{P}(X \in \{1,2,3,4,5,6\}) = 1$

and $\mathbb{P}(\mathbb{R} \setminus \{1,2,3,4,5,6\}) = 0$



Note: You can always take a r.v. defined on a countable outcome space and define it on a larger uncountable outcome space by setting the probability of the event containing all the new outcomes to zero

Continuous r.v.: A r.v. that takes values from:

- an uncountable outcome space and the probability of any single outcome is zero

Ex: $Z \in [0,900]$ and $\mathbb{P}(Z=z)=\mathbb{P}(Z\in\{z\})=0$ for all $z\in[0,900]$

but $\mathbb{P}(Z\in[0,900])=1$

Ex: $Z \in \mathbb{R}$ and $\mathbb{P}(Z=z)=\mathbb{P}(Z=\{z\})=0$ for all $z\in\mathbb{R}$

but $\mathbb{P}(Z\in\mathbb{R})=1$

Calculating Probabilities

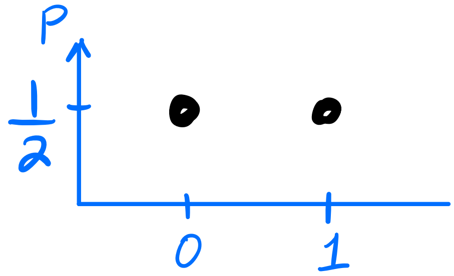
Motivation: It is hard to define the values of a probability distribution $\mathbb{P}$ for all the events

Probability Mass Function (pmf): A function $p:\mathcal{Z}\to[0,1]$ where $\mathcal{Z}$ is a contable outcome space and $\sum_{z\in\mathcal{Z}} p(z)=1$.

The probability of an event $\mathcal{E}\subset\mathcal{Z}$ is:

$$\boxed{\mathbb{P}(Z\in\mathcal{E}) \stackrel{\text{def}}{=} \sum_{z\in\mathcal{E}} p(z)} \qquad \text{where } Z\in\mathcal{Z}$$

Ex: Outcome space: $\mathcal{Y}$



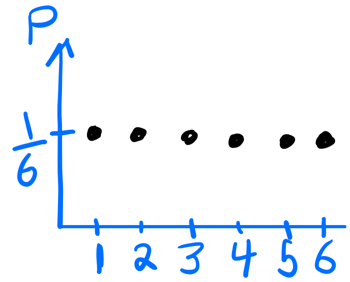
$p(0)=\frac{1}{2},\ p(1)=\frac{1}{2}$

$$\mathbb{P}(Y\in\{0,1\})=\sum_{y\in\{0,1\}}p(y)=p(0)+p(1)=1$$

$$\mathbb{P}(Y=0)=\mathbb{P}(Y=1)=\mathbb{P}(Y\in\{1\})=\sum_{y\in\{1\}}p(y)=p(1)=\frac{1}{2}$$

$$\mathbb{P}(Y\in\emptyset)=\sum_{y\in\emptyset}p(y)=0$$

Ex: Outcome space: $\mathcal{X}$



$p(1)=p(2)=\dots=p(6)=\frac{1}{6}$

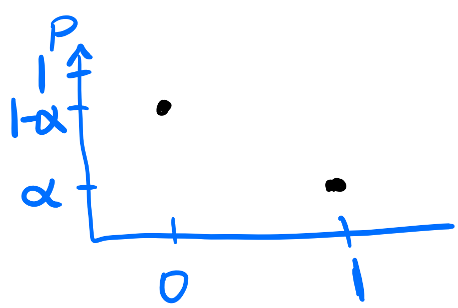
$$\mathbb{P}(X\in\{1,3,5\})=\sum_{x\in\{1,3,5\}}p(x)=p(1)+p(3)+p(5)=\frac{1}{2}$$

Discrete Probability Distributions with special names:

Bernoulli distribution (parameter: $\alpha\in[0,1]$):

Outcome space: $\{0,1\}$

pmf: $p(1)=\alpha,\ p(0)=1-\alpha$



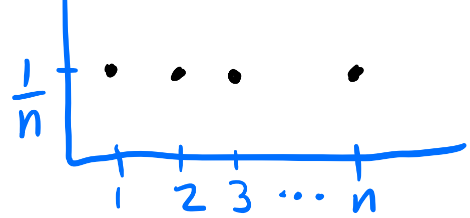
$\boxed{\text{Distribution } \mathbb{P}=\text{Bernoulli}(\alpha)}$

$\mathbb{P}(Z=1)=\mathbb{P}(Z\in\{1\})=p(1)=\alpha$ $\quad Z\in\{0,1\}$ is a "Bernoulli r.v."

Discrete Uniform Distribution (parameter: n):

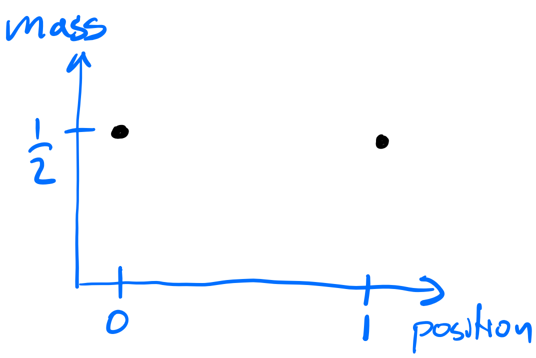
Outcome space: $\{1,2,\dots,n\}$

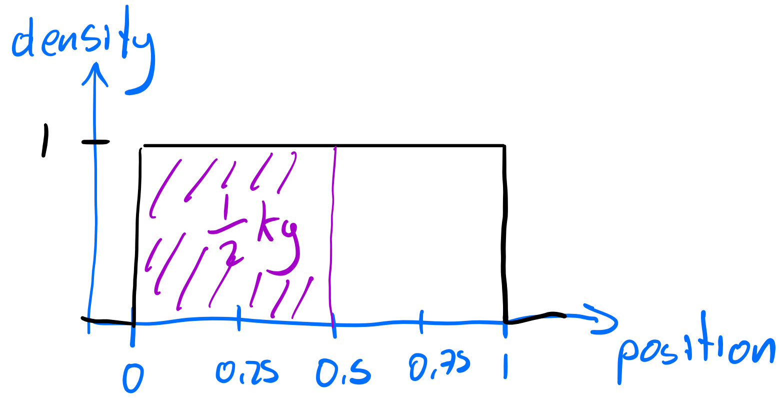
pmf: $p(1)=p(2)=\dots=p(n)=\frac{1}{n}$

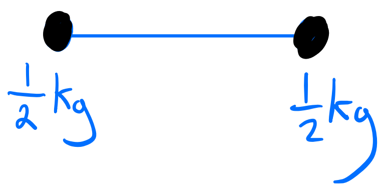


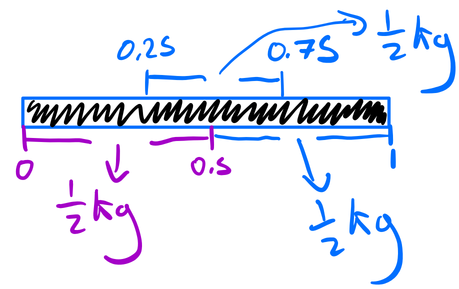
$$\boxed{\text{Distribution } \mathbb{P} = \text{Uniform}(n)}$$

Intuition with a rod in physics









Probability Density Function (pdf): a function $p: \mathcal{Z} \to [0,\infty)$ where $\mathcal{Z}$ is an uncountable outcome space and $\int_{\mathcal{Z}} p(z)\,dz = 1$

The probability of an event $E \subset \mathcal{Z}$ is:

$$\boxed{\mathbb{P}(Z\in E) \stackrel{def}{=} \int_E p(z)\,dz}$$

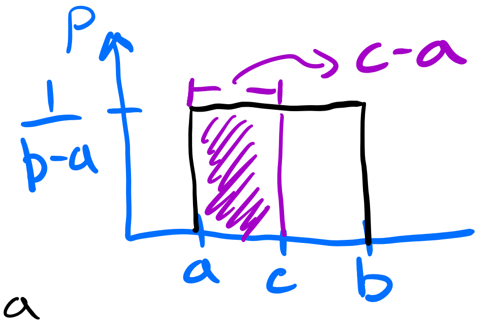
$$\boxed{\mathbb{P}(Z=z)=\mathbb{P}(Z\in\{z\})=0}$$ where $z\in\mathcal{Z}$

Continuous Probability Distributions with special names:

Continuous Uniform Distribution (parameters: $a \in \mathbb{R}, b \in \mathbb{R}$):

Outcome space: $[\underset{\substack{\downarrow \\ 0}}{a}, \underset{\substack{\downarrow \\ 6}}{b}]$

pdf: $p(z) = \frac{1}{b-a} = \frac{1}{6}$



$$\boxed{\text{Distribution } \mathbb{P} = \text{Uniform}(a, b)}$$

$$\mathbb{P}(a \leq Z \leq c) = \mathbb{P}(Z \in [a, c]) = \int_a^c p(z)\,dz = \frac{c-a}{b-a}$$

where $a \leq c \leq b$

Gaussian/Normal Distribution (parameters: $\mu \in \mathbb{R}, \sigma^2 > 0$):

Outcome space: $\mathbb{R}$

pdf: $p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right)$

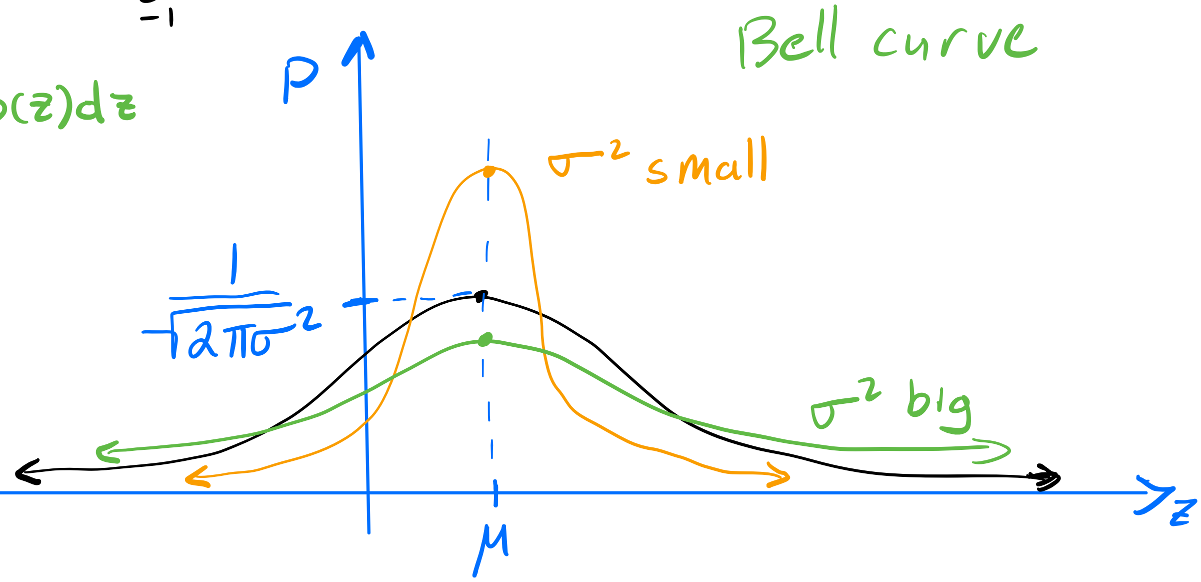
$$\boxed{\text{Distribution } \mathbb{P} = \mathcal{N}(\mu, \sigma^2) = \text{Gaussian}(\mu, \sigma^2)}$$

$\mathcal{N}(0,1) \leftarrow$ standard normal

Ex $\mathbb{P}(-1 \leq Z \leq 1) = \int_{-1}^{1} p(z)\,dz$

Bell curve

$\mathbb{P}(Z \in \mathbb{R}) = \int_{-\infty}^{\infty} p(z)\,dz$

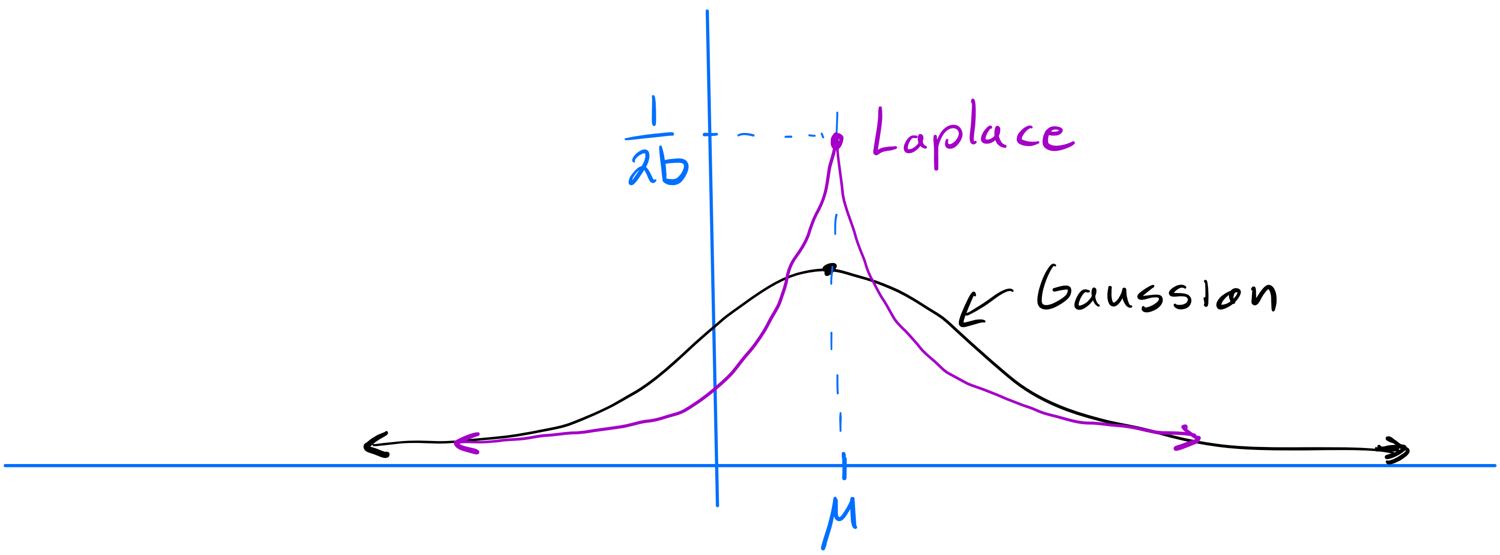


Laplace Distribution (parameters: $\mu \in \mathbb{R}, b>0$):

Outcome space: $\mathbb{R}$

pdf: $p(z) = \frac{1}{2b} \exp\left(-\frac{1}{b}|x-\mu|\right)$

$$\boxed{\text{Distribution } \mathbb{P} = \text{Laplace}(\mu, b)}$$



Multivariate Random Variables

Motivation: To be able to talk about the probability of different types of events at the same time

Ex: The probability of getting heads and rolling a 3

The probability of a wine containing 2.5mg of one chemical and 4mg of another chemical

The probability of a house having 4 rooms and 2 washrooms and being less than 10min from a university

The probability of being young and having arthritis

Multivariate Random Variable: A tuple of more than one random variable

Ex: (Flipping 2 coins)

Heads

Outcome space: $\mathcal{X} = \{0,1\} \times \{0,1\} = \{(0,0),(0,1),(1,0),(1,1)\}$

r.v.: $X = (X_1, X_2) \in \mathcal{X}$

(Collecting the info of one house (ex: # of rooms, age))

Outcome space: $\mathcal{X} = \mathbb{N} \times \underbrace{[0,\infty)}_{}$ age

\# of rooms

r.v.: $X = (X_1, X_2) \in X$

(Collecting the info of one house and its price)

Outcome space: $\mathcal{Z} = (\mathbb{N} \times \underbrace{[0,\infty)}_{}) \times \underbrace{[0,\infty)}_{}$ age

\# of rooms Price

r.v.: $Z = (X, Y)$ $\in$ $\in$

$= ((X_1, X_2), Y) = (X_1, X_2, Y)$

Calculating Joint Probabilities

Ex: (If you have arthritis and if you are young or old)

Old Arthritis

Outcome space: $\mathcal{Z} = \{0,1\} \times \{0,1\} = \{(0,0),(0,1),(1,0),(1,1)\}$

Young No arthritis

$\in \{0,1\}$

r.v.: $Z = (X, Y) \in \mathcal{Z}$

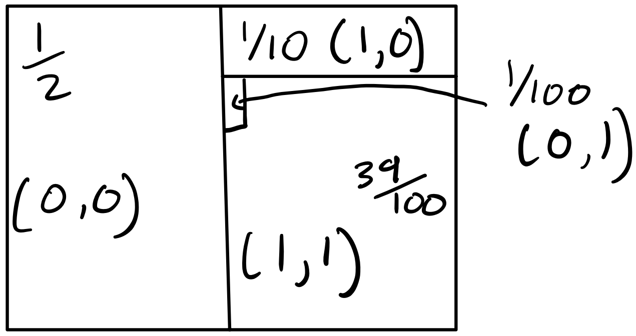
$\in \{0,1\}$

pmf: $p : \mathcal{Z} \to [0,1]$

Not based on real data
$$\begin{cases} p((0,0)) = p(0,0) = \frac{1}{2}, & p(0,1) = \frac{1}{100} \\ p(1,0) = \frac{1}{10}, & p(1,1) = \frac{39}{100} \end{cases}$$

		Y: 0	1
X	0	1/2	1/100
	1	1/10	39/100

$p(X,Y)$



$$\sum_{z \in \mathcal{Z}} p(z) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) = \frac{1}{2} + \frac{1}{100} + \frac{1}{10} + \frac{39}{100} = 1$$

What is the probability of being young (i.e. $X=0$)?

$$\mathcal{E} = \{(0,0),(0,1)\} = \{0\} \times \{0,1\} \subset \mathcal{Z}$$

$$P(X=0, Y\in\{0,1\}) = P(Z \in \mathcal{E}) = \sum_{z \in \mathcal{E}} p(z)$$

$$= \underbrace{\sum_{x\in\{0\}} \sum_{Y\in\{0,1\}}}_{= \sum_{z \in \{(0,0),(0,1)\}}} p(x,y)$$

$$= p(0,0) + p(0,1)$$

$$= \frac{1}{2} + \frac{1}{100} = \frac{51}{100}$$

Marginal Distribution: The distribution over a subset of random variables

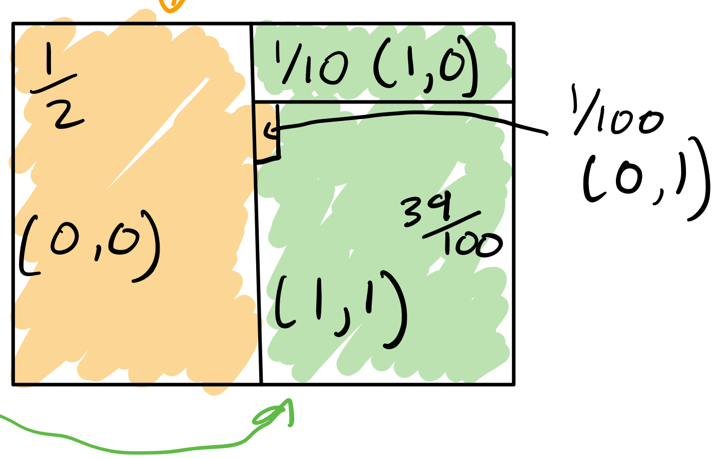
Ex: Continuing with the arthritis example

Marginal Distribution: $\mathbb{P}_X(X \in E_x)$ where $E_x \in \mathcal{X}$

Marginal pmf: $p_X : \mathcal{X} \to [0,1]$, $p_X(x) = \sum_{y \in \mathcal{Y}} p(x,y)$

$$\mathbb{P}_X(X=0) \overset{\mathbb{P}(X \in \{0\})}{=} p_X(0) = \sum_{x \in \{0\}} \sum_{y \in \mathcal{Y}} p(0,y) = \frac{51}{100}$$

$$\mathbb{P}_X(X=1) = \frac{1}{10} + \frac{39}{100} = \frac{49}{100}$$



Discrete r.v. $X_1, \dots, X_d$

$$X = (X_1, \dots, X_d) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d = \mathcal{X}, \quad p : \mathcal{X} \to [0,1]$$

$$p_{X_i} : \mathcal{X}_i \to [0,1], \quad i \in \{1, \dots, d\}$$

Marginal pmf:

$$\boxed{p_{X_i}(x_i) \overset{\text{def}}{=} \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, \dots, x_i, \dots, x_d)}$$

$$\boxed{\mathbb{P}_{X_i}(X_i \in E_i) = \sum_{x_i \in E_i} p_{X_i}(x_i)} \quad \text{where } E_i \subset \mathcal{X}_i$$

Continuous r.v. $X_1, \dots, X_d$

$$X = (X_1, \dots, X_d) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d = \mathcal{X}, \quad p : \mathcal{X} \to [0, \infty)$$

$p_{X_i}: \mathcal{X}_i \to [0,\infty)$, $i \in \{1,\dots,d\}$

Marginal pdf: $\boxed{p_{X_i}(x_i) = \int_{\mathcal{X}_1}\cdots\int_{\mathcal{X}_{i-1}}\int_{\mathcal{X}_{i+1}}\cdots\int_{\mathcal{X}_d} p(x_1,\dots,x_d)\,dx_1\cdots dx_{i-1}\,dx_{i+1}\,dx_d}$

Distribution: $\boxed{\mathbb{P}_{X_i}(X_i \in \mathcal{E}_i) = \int_{\mathcal{E}_i} p_{X_i}(x_i)\,dx_i}$ where $\mathcal{E}_i \subset \mathcal{X}_i$

Conditional Distribution: Probability of a r.v. given info about another r.v.

Ex: Probability that I have arthritis given I am young

Let r.v. $= Y \in \mathcal{Y}$, $X \in \mathcal{X}$

Discrete Y for any $x \in \mathcal{X}$ that $p_X(x) \neq 0$

$p_{Y|X=x}: \mathcal{Y} \to [0,1]$, $\quad p_{Y|X=x}(y) = p_{Y|X}(y|x)$

conditional pmf: $\boxed{p_{Y|X}(y|x) \stackrel{\text{def}}{=} \frac{p(y,x)}{p_X(x)}}$ implies $\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x) = 1$

Distribution: $\boxed{\mathbb{P}_{Y|X}(Y \in \mathcal{E}_Y | X = x) \stackrel{\text{def}}{=} \sum_{y \in \mathcal{E}_Y} p_{Y|X}(y|x)}$ where $\mathcal{E}_Y \subset \mathcal{Y}$

Continuous Y for any $x \in \mathcal{X}$ that $p_X(x) \neq 0$

$p_{Y|X=x}: \mathcal{Y} \to [0,\infty)$, $\quad p_{Y|X=x}(y) = p_{Y|X}(y|x)$

conditional pdf: $\boxed{p_{Y|X}(y|x) \overset{\text{def}}{=} \frac{p(y,x)}{p_X(x)}}$ implies $\int_{\mathcal{Y}} p_{Y|X}(y|x)\,dy = 1$

Distribution: $\boxed{\mathbb{P}_{Y|X}(Y \in E \mid X = x) \overset{\text{def}}{=} \int_{E} p_{Y|X}(y|x)\,dy}$ where $E_Y \subset \mathcal{Y}$

Chain Rule: $\boxed{p(x,y) = p(x|y)p(y) = p(y|x)p(x)}$

More generally:

$$p(x_1, x_2, \ldots, x_d) = p(x_d \mid x_1, \ldots, x_{d-1}) \cdots p(x_3 \mid x_1, x_2)\, p(x_2 \mid x_1)\, p(x_1)$$

Bayes' Rule: $\boxed{p(x|y) = \frac{p(y|x)p(x)}{p(y)}}$

Note: Sometimes the subscripts are not used for marginal and conditional distributions when it is clear from the context

$$p(x,y) = p_{X,Y}(x,y)$$

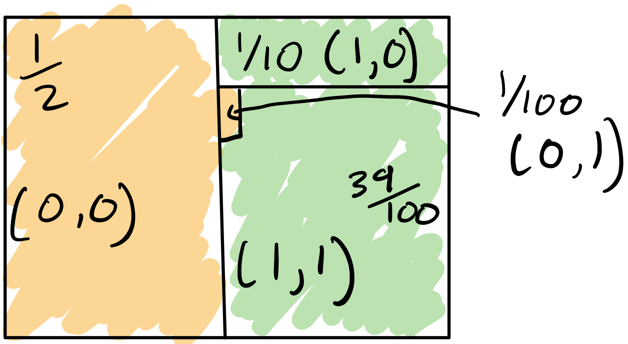
$$p(x) = p_X(x)$$

$$p(y|x) = p_{Y|X}(y|x)$$

Ex: Probability that I have arthritis given I am young

$$\mathbb{P}(Y=1 \mid X=0) = \sum_{y \in \{1\}} p_{Y|X}(y|0)$$

Arthritis young



$$= P_{Y|X}(1|0)$$

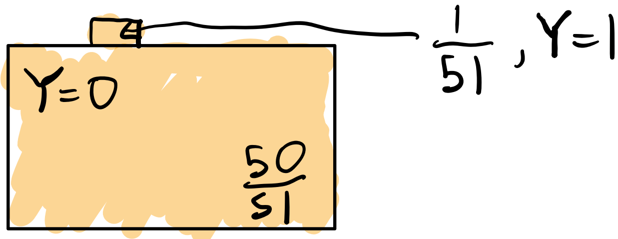
$$= \frac{p(0,1)}{P_X(0)}$$

$$= \frac{p(0,1)}{p(0,1)+p(0,0)}$$

$$= \frac{1/100}{1/100 + 1/2}$$

$$= \frac{1/100}{51/100} = \frac{1}{51}$$

condition on being young



<u>Ex</u>: Probability of being young given I have arthritis

$$\mathbb{P}(X=0 \mid Y=1) = P_{X|Y}(0|1)$$

Bayes' Rule

$$\overset{\text{Bayes' Rule}}{=} \frac{P_{Y|X}(1|0)\, p(0)}{P_Y(1)}$$

$$= \frac{\frac{1}{51} \cdot \frac{51}{100}}{40/100} = \frac{1}{40}$$

<u>Independence</u>: Changing the value of one r.v. doesn't affect the probability of another r.v.

$\boxed{\text{r.v. } X, Y \text{ are independent if: } p(x,y) = p(x)p(y)}$

Since $p(x,y) = p(x|y)p(y) = p(x)p(y|x) = p(x)p(y)$

$\boxed{\text{independence implies: } p(x|y) = p(x),\ p(y|x) = p(y)}$

More generally:

$X_1, X_2, \ldots, X_d$ are independent if: $p(x_1, \ldots, x_d) = p(x_1) \cdots p(x_d)$

Similarly for distributions:

$\boxed{\text{r.v. } X, Y \text{ are independent if: } \mathbb{P}(X \in E_X, Y \in E_Y) = \mathbb{P}(X \in E_X)\mathbb{P}(Y \in E_Y)}$

Ex: X, Y are not independent for Arthritis ex

$$p(0,1) = \frac{1}{100} \neq p_X(0)p_Y(1) = \frac{51}{100}\frac{40}{100} = 0.204$$

Ex: $X_1, X_2 \in \{0, 1\}$ are flips of two different fair coins

$$p(x_1, x_2) = \frac{1}{4} \quad \text{for all } x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$$

$$p_{X_1}(x_1)p_{X_2}(x_2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

		X_2: H	X_2: T
X_1	H	1/4	1/4
	T	1/4	1/4

What happens when $Z=(X,Y)$ with Y discrete and X continuous?

$p: \mathcal{X} \times \mathcal{Y} \to ?$ pmf or pdf? Ans: neither

Instead we will write $p(x,y)$ in terms of a marginal pdf for X and a conditional pmf for $Y|X$

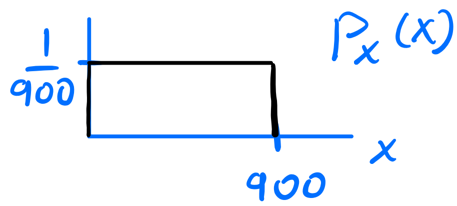
$$p(x,y) = p_X(x)\, p_{Y|X}(y|x)$$ product rule

where $p_{Y|X=x}: \mathcal{Y} \to [0,1]$ is a pmf

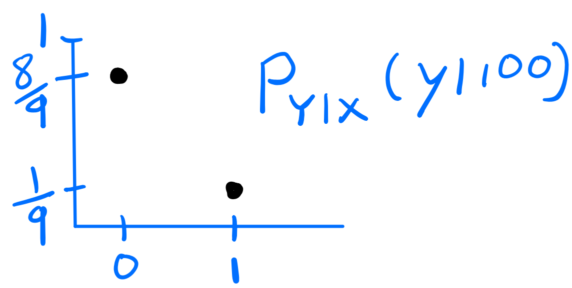
$p_X(x): \mathcal{X} \to [0,\infty)$ is a pdf

Ex: $X \in \mathcal{X} = [0,900]$, $Y \in \mathcal{Y} \in \{0,1\}$ Barolo

pdf: $p_X = \text{Uniform}(0,900)$
$= \frac{1}{900}$



$P_{Y|X=x} = \text{Bernoilli}\left(\frac{x}{900}\right)$



pmf: $P_{Y|X}(y|x) = \begin{cases} \frac{x}{900} & \text{if } y=1 \\ 1-\frac{x}{900} & \text{if } y=0 \end{cases}$

Defn of Bernoilli$\left(\frac{x}{900}\right)$

$$\mathbb{P}(X \in [0,50], Y=1) = \int_0^{50} \left(\sum_{y \in \{1\}} p(y,x) \right) dx$$

$$= \int_0^{50} \left(\sum_{y \in \{1\}} P_{Y|X}(y|x)\, P_X(x) \right) dx$$

$$= \int_0^{50} p_{Y|X}(1|x)\, p_X(x)\, dx$$

$$= \int_0^{50} \frac{x}{900} \frac{1}{900}\, dx$$

$$= \frac{1}{810000} \left. \frac{x^2}{2} \right|_0^{50}$$

$$= \frac{1}{810000} \frac{2500}{2}$$

$$= \frac{1}{648}$$

0.154%

Representing Random Features, Labels, and Datasets

Random variables:

$$D = (Z_1, Z_2, \ldots, Z_n) \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n = \mathcal{Z}^n \quad \text{since } \mathcal{Z} = \mathcal{Z}_1 = \cdots = \mathcal{Z}_n$$

$$Z_i = (\vec{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$$ each Z_i is a feature-label pair

$$\vec{X}_i = (X_{i,1}, \ldots, X_{i,d})^T \in \mathbb{R}^d = \mathcal{X}$$ $\vec{X}_i$ is a feature vector

Distributions:

$\mathbb{P}_D$: distribution for D, $\mathbb{P}_{Z_i}$: marginal distribution for Z_i

assumptions:

1. $(\vec{X}_i, Y_i) = Z_i$ are indendent for all $i \in \{1, \ldots, n\}$

2. $\mathbb{P}_{Z_1} = \mathbb{P}_{Z_2} = \cdots = \mathbb{P}_{Z_n} = \mathbb{P}_Z$ all Z_i have the same distribution

"$(\vec{X}_i, Y_i)$ are independent and identically distributed (i.i.d)"

$$\mathbb{P}_D(Z_1 \in E_1, \ldots, Z_n \in E_n) = \mathbb{P}_{Z_1}(Z_1 \in E_1) \cdots \mathbb{P}_{Z_n}(Z_n \in E_n)$$

$$= \mathbb{P}_Z(Z_1 \in E_1) \cdots \mathbb{P}_Z(Z_n \in E_n)$$

Equivalently:

$$D = ((\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$$

where $(\vec{X}_i, Y_i) \sim \mathbb{P}_{\vec{X}, Y}$ are idependent for all $i \in \{1, \ldots, n\}$

"sampled/distributed according to"

D contains n independent samples of $(\vec{X}_i, Y_i)$ "feature-label" pairs all coming from the same distribution $\mathbb{P}_{\vec{X},Y}$

Functions of Random Variables

A function of a r.v. is a r.v.

Ex: (X is a fair six-sided dice)

$X \in \{1, 2, 3, 4, 5, 6\} = \mathcal{X}$ with $p(x) = \frac{1}{6}$

$f(X) = X^2 \in \underbrace{\{1^2, 2^2, 3^2, 4^2, 5^2, 6^2\}}_{\text{outcome space for } f(X)} = \mathcal{Y}$ is a r.v.

Notice $f: \mathcal{X} \to \mathcal{Y}$

Sometimes we give the r.v. a new symbol

$Y = f(X) = X^2$

$P_Y(y) = P_{f(X)}(y) = \frac{1}{6}$ where $y \in \{1, 2^2, 3^2, 4^2, 5^2, 6^2\}$

In this case $p_Y(x^2) = p(x)$ where $x \in \mathcal{X}$

ex: $p_Y(9) = p_Y(3^2) = p(3) = \frac{1}{6}$

Ex: (X is the payout from a slot machine)

$$X \in [-10, 10] \quad \text{with } p(x) = \frac{1}{20}, \quad \mathbb{P} = \text{Uniform}(-10, 10)$$

$$Y = f(X) = X^2 \in [0, 100] = \mathcal{Y}$$

$$p_Y(y) = \frac{1}{20\sqrt{y}}$$ much more complicated

In general p_Y is complicated and we will not need to know how to calculate it

The Predictor and Learner are functions of r.v.

Ex: (Predictor)

$$\vec{X} = (X_1, X_2)^T \in \mathbb{R}^2 = \mathcal{X} \quad \text{with } \mathbb{P}_{\vec{X}}$$

predictor: $f: \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \mathbb{R}$

$f(\vec{X}) = 3 + 6X_1 + 2.5X_2$ is a r.v. with values in $\mathcal{Y}$

and has some distribution $\mathbb{P}_{f(\vec{X})}$

Ex: (Learner)

$$D = ((\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n \quad \text{with } \mathbb{P}_D$$

Learner: $\mathcal{A}: (\mathcal{X} \times \mathcal{Y})^n \Rightarrow \{f \mid f: \mathcal{X} \Rightarrow \mathcal{Y}\} = \mathcal{F}$

$\mathcal{A}(D) = f$ is a r.v. with values in $\mathcal{F}$

and has some distribution $\mathbb{P}_{\mathcal{A}(D)}$

example

if $D = ((7, 6), (12, 2.5))$ where $n = 2$, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$

then f_D can be $f(x) = 2.5 + 6x$

This means we can talk about things like:

- What is the probability the Predictor $f(\vec{x})$ outputs some value y
- What is the probability the Learner $\mathcal{A}(D)$ outputs some predictor f

Expectation and Variance

<u>Expected Value of a r.v.:</u> average value of the r.v. if you sample from its distribution infinitely many times.

The r.v. <u>must</u> take values in $\mathbb{R}$.

It is <u>not</u> always the value we expect to see most frequently (that is the mode)

$X \in \mathcal{X}$ is a r.v. with pmf or pdf p

$$\boxed{\mathbb{E}[X] \stackrel{\text{def}}{=} \begin{cases} \sum_{x \in \mathcal{X}} x\, p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} x\, p(x)\, dx & \text{if } X \text{ is continuous} \end{cases}}$$

<u>Ex:</u> (fair six-sided dice)

$X \in \{1,2,3,4,5,6\} = \mathcal{X}$ and $\mathbb{P} = \text{Uniform}(n=6)$

thus $p(x) = \frac{1}{6}$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$= 3.5$$

This is <u>not</u> an number you can roll on a dice!

Ex: (Unfair coin)

$X \in \{0,1\}$ and $\mathbb{P} = \text{Bernoulli}(\alpha)$

thus $p(1) = \alpha,\ p(0) = 1-\alpha$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x) = 0 \cdot (1-\alpha) + 1 \cdot \alpha$$
$$= \alpha$$

This is <u>not</u> a result of a coin flip (unless $\alpha = 1$ or $\alpha = 0$)

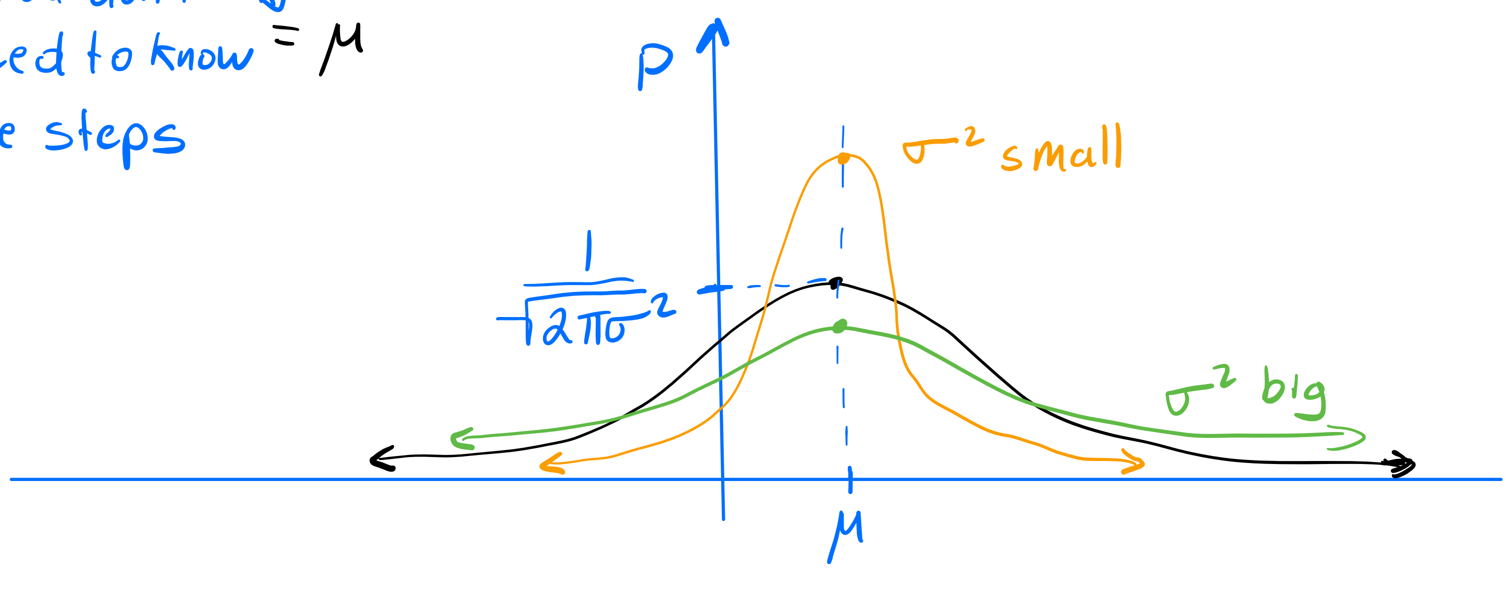
Ex: (Normal distribution)

$X \in \mathbb{R} = \mathcal{X}$ and $\mathbb{P} = \mathcal{N}(\mu, \sigma^2)$

thus $p(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{1}{2\sigma^2}(x-\mu)^2\right)$

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

You don't need to know the steps $= \mu$



Expected value of functions of r.v.:

$X \in \mathcal{X}$ is a r.v. with pmf or pdf p

The function $f : \mathcal{X} \to \mathcal{Y}$ must have $\mathcal{Y} = \mathbb{R}$

$$\boxed{\mathbb{E}[f(X)] \overset{\text{def}}{=} \begin{cases} \sum_{x \in \mathcal{X}} f(x) p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} f(x) p(x)\, dx & \text{if } X \text{ is continuous} \end{cases}}$$

Ex: (X is the payout from a slot machine)

$X \in [-10, 10]$ with $p(x) = \frac{1}{20}$, $\mathbb{P} =$ Uniform$(-10, 10)$

$Y = f(X) = X^2 \in [0, 100] = \mathcal{Y}$

$p_Y(y) = \frac{1}{20\sqrt{y}}$ much more complicated

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x) p(x)\, dx = \int_{-10}^{10} x^2 \frac{1}{20}\, dx$$

$$= \frac{x^3}{3} \cdot \frac{1}{20} \Big|_{-10}^{10}$$

$$= \left(\frac{1000}{3} - \left(\frac{-1000}{3}\right)\right) \cdot \frac{1}{20}$$

$$= \frac{2000}{60} = 33.333$$

It turns out

$$\mathbb{E}[f(X)] = \mathbb{E}[Y] = \int_{\mathcal{Y}} y\, p_Y(y)\, dy$$

exercise $\rightarrow$ $= 33.333$

Usually we don't know $p_Y = p_{f(X)}$

So we work with p

Variance of a r.v.: How much the r.v. varies from its expected value on average

$X \in \mathcal{X}$ is a r.v. with pmf or pdf p

$$\boxed{\mathrm{Var}[X] \stackrel{\text{def}}{=} \mathbb{E}\Big[\underbrace{\big(X - \mathbb{E}[X]\big)^2}\Big] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2}$$

this is just a function of the r.v. X

Ex: (Unfair coin)

$X \in \{0, 1\}$ and $\mathbb{P} = \text{Bernoulli}(\alpha)$

thus $p(1) = \alpha,\ p(0) = 1 - \alpha$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x) = 0 \cdot (1-\alpha) + 1 \cdot \alpha$$
$$= \alpha$$

$$\text{Var}[X] = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$
$$= \sum_{x \in \mathcal{X}} \left(x - \mathbb{E}[X]\right)^2 p(x)$$
$$= (0-\alpha)^2 \cdot (1-\alpha) + (1-\alpha)^2 \cdot \alpha$$
$$= \alpha^2 - \alpha^3 + \alpha - 2\alpha^2 + \alpha^3$$
$$= \alpha - \alpha^2$$
$$= \alpha(1-\alpha)$$

or $\text{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2$

$$= \sum_{x \in \mathcal{X}} x^2 p(x) - \alpha^2$$
$$= 0^2 \cdot (1-\alpha) + 1^2 \cdot \alpha - \alpha^2$$
$$= \alpha(1-\alpha)$$

<u>Ex:</u> (Normal distribution)

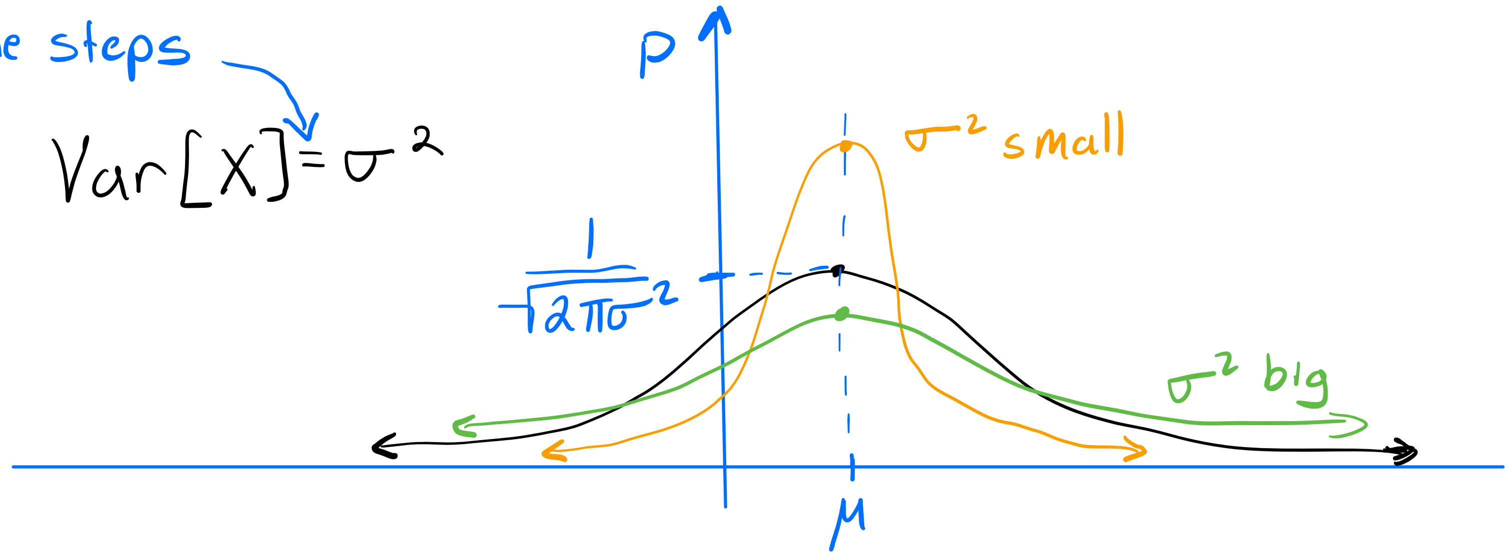
$X \in \mathbb{R} = \mathcal{X}$ and $\mathbb{P} = \mathcal{N}(\mu, \sigma^2)$

thus $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = \mu$$

You don't need to know the steps

$$\mathrm{Var}[X] = \sigma^2$$



Multivariate Expected Value:

$Z = (X, Y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ is a r.v.

$f : \mathcal{X} \times \mathcal{Y} \Rightarrow \mathbb{R}$

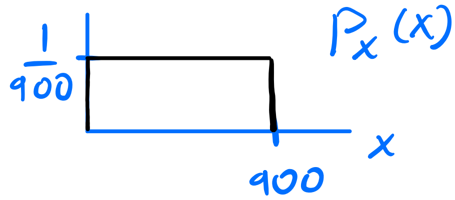
$$\mathbb{E}[f(X,Y)] = \begin{cases} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y) p(x,y) & \text{if } X, Y \text{ are discrete} \\ \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y) p(x,y) \, dy \, dx & \text{if } X, Y \text{ are continuous} \\ \int_{\mathcal{X}} \left(\sum_{y \in \mathcal{Y}} f(x,y) p(y|x) \right) p(x) \, dx & \text{if } Y \text{ is discrete and } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} \left(\int_{\mathcal{Y}} f(x,y) p(y|x) \, dy \right) p(x) & \text{if } Y \text{ is continuous and } X \text{ is discrete} \end{cases}$$

you can always use:

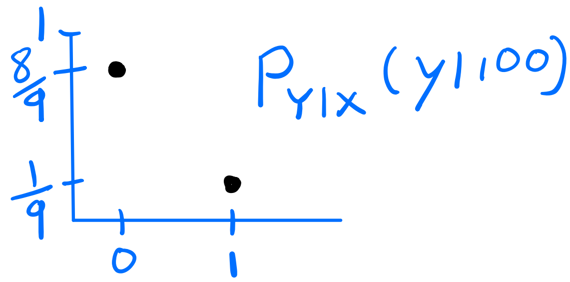
$$p(x,y) = p(y|x)\,p(x) = p(x|y)\,p(x)$$

Ex: $X \in \mathcal{X} = [0,900]$, $Y \in \mathcal{Y} \in \{0,1\}$ ← Barolo

pdf: $p_X = \text{Uniform}(0,900)$
$= \frac{1}{900}$



$$P_{Y|X=x} = \text{Bernoilli}\left(\frac{x}{900}\right)$$



pmf: $$P_{Y|X}(y|x) = \begin{cases} \frac{x}{900} & \text{if } y=1 \\ 1-\frac{x}{900} & \text{if } y=0 \end{cases}$$

Defn of Bernoilli$\left(\frac{x}{900}\right)$

$$f(X,Y) = \left(\frac{X}{900} - Y\right)^2$$

$$\mathbb{E}[f(X,Y)] = \int_{\mathcal{X}} \left(\sum_{y \in \mathcal{Y}} f(x,y)\,p(y|x) \right) p(x)\,dx$$

$$= \int_0^{900} \left(\sum_{y \in \{0,1\}} \left(\frac{x}{900} - y\right)^2 p(y|x) \right) p(x)\,dx$$

$$= \int_0^{900} \left(\left(\frac{x}{900} - 0\right)^2 \left(1 - \frac{x}{900}\right) + \left(\frac{x}{900} - 1\right)^2 \left(\frac{x}{900}\right) \right) \frac{1}{900}\,dx$$

$$= \frac{1}{900} \int_0^{900} \frac{x}{900} \left(1 - \frac{x}{900}\right) dx$$

$$= \frac{1}{900} \left(\frac{x^2}{1800} \Big|_0^{900} - \frac{x^3}{3 \cdot 900^2} \Big|_0^{900} \right)$$

$$= \frac{1}{6}$$

Conditional Expected Value:

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is a r.v.

$p = p_{Y|X}$ is a conditional pmf or pdf

$f : \mathcal{Y} \to \mathbb{R}$

$$\boxed{\mathbb{E}[f(Y) \mid X = x] = \begin{cases} \sum_{y \in \mathcal{Y}} f(y) p(y \mid x) & \text{if } Y \text{ is discrete} \\ \int_{\mathcal{Y}} f(y) p(y \mid x) & \text{if } Y \text{ is continuous} \end{cases}}$$

Useful Properties

Let X, Y be r.v. and $c \in \mathbb{R}$ be a constant

1. $\mathbb{E}[cX] = c\,\mathbb{E}[X]$
2. $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$
4. $\mathrm{Var}[c] = 0$
5. $\mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$

If X and Y are independent:

6. $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$
7. $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$