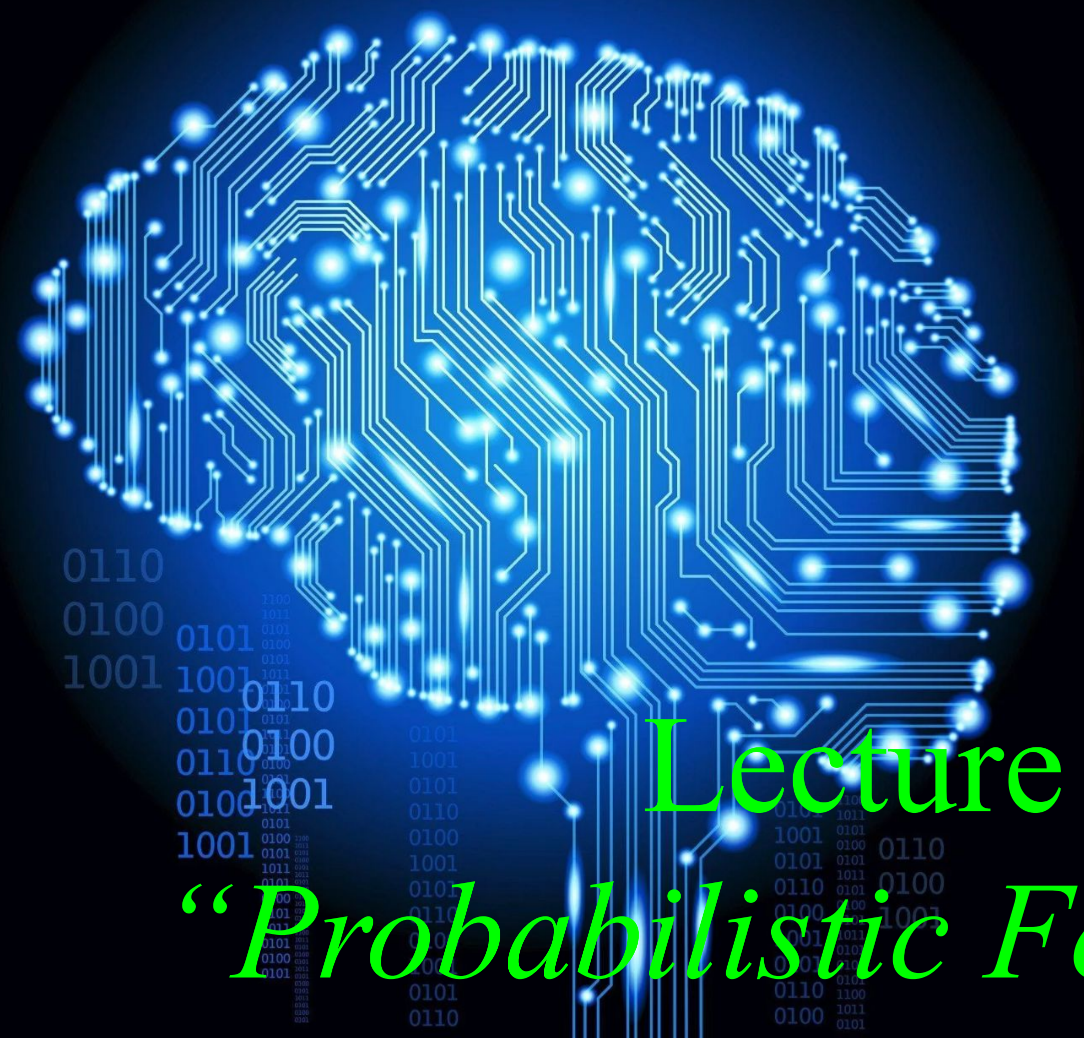
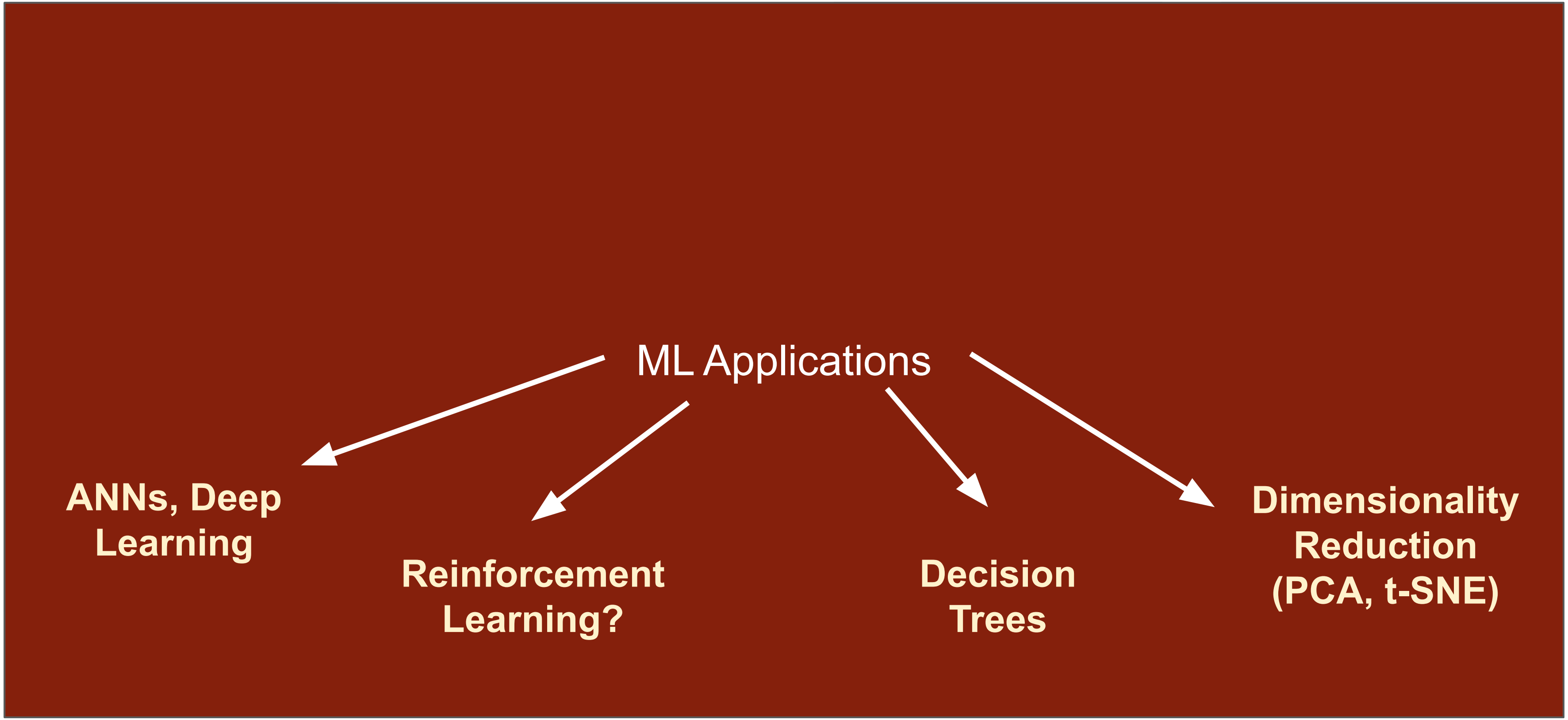


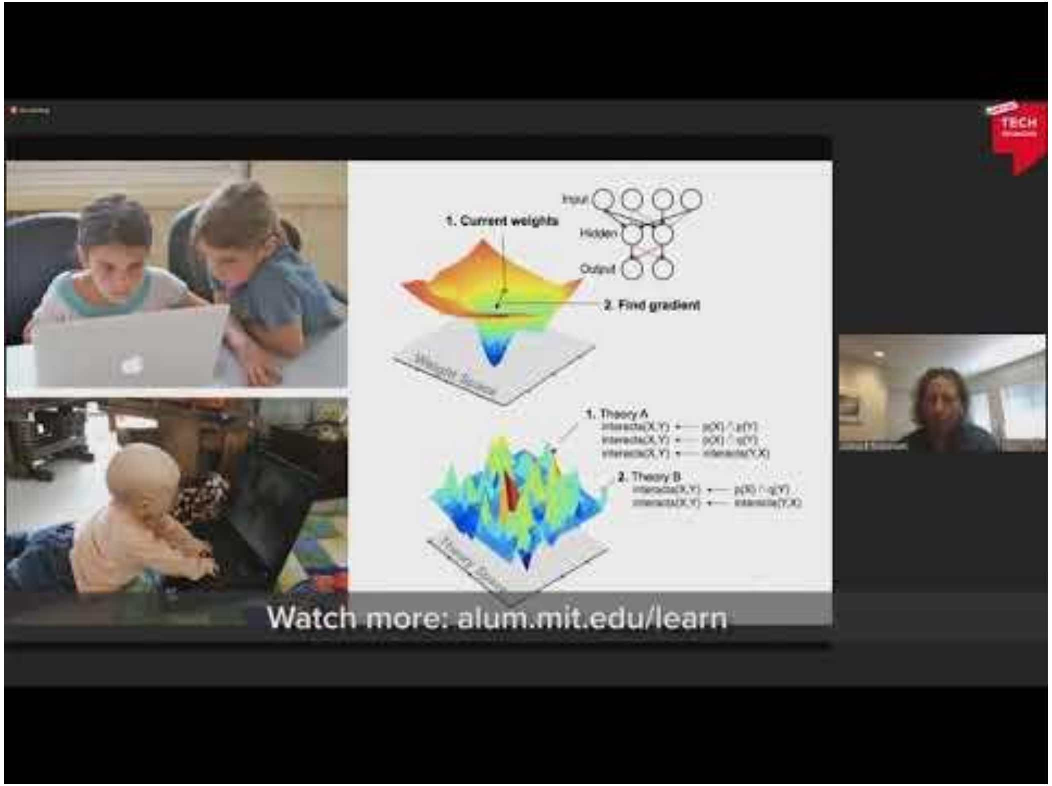
Lecture 02:

"Probabilistic Foundations"

What have we done? Where are we? Where are we going?



Before we start...



https://probmods.org/chapters/generative-models.html

What is “random”?

A coin flip?

A random survey of students in the class?

numpy.random.random()?

Today: we’ll talk about a mathematical formalism for randomness: **events and probability**

Sample Spaces

Sample space $\mathcal{S}$ is set of all observable possible events.

- Coin flips: $\mathcal{S} = \{h, t\}$
- An individual's height: $\mathcal{S} = \mathbb{R}^{\geq 0}$
- (An individual's height, weight): $\mathcal{S} = \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$

The probability of an observation falling somewhere in the sample space is 1.

$$Pr(\mathcal{S}) = 1$$

An event is a subset of the sample space.

- Observe heads: $\{h\}$
- Observe height of at least 170 cm: $[170, \infty]$
- Observe height between 170 and 190 cm and weight between 65 and 72 kg: $[170, 190] \times [65, 72]$

An event is assigned a probability between 0 and 1.

$$Pr(\{h\}) = 0.5$$

Random Variables

- Random variable (r.v.) = A mapping from the event space to a number or vector
  - Notation: X, Y, Z, etc.
- "Realizations" = observed pieces of data from random variable
  - Notation: x,y,z, etc.

- Set of possible realizations
  - Notation: $\mathscr{X}$ for X

- Realizations are observed as per probabilities specified by the **distribution** of X
  - realizations of the same X are independent and identically distributed (i.i.d)

Discrete Random Variable (R.V.)

- Discrete random variables take values from countable set
  - I.e. coin flip X, $\mathcal{X} = \{0, 1\}$

- **Probability mass function (PMF)**: for a discrete X, $p_x(x)$ gives Pr(X=x)
  - Here we need the sum of all probabilities to add to 1

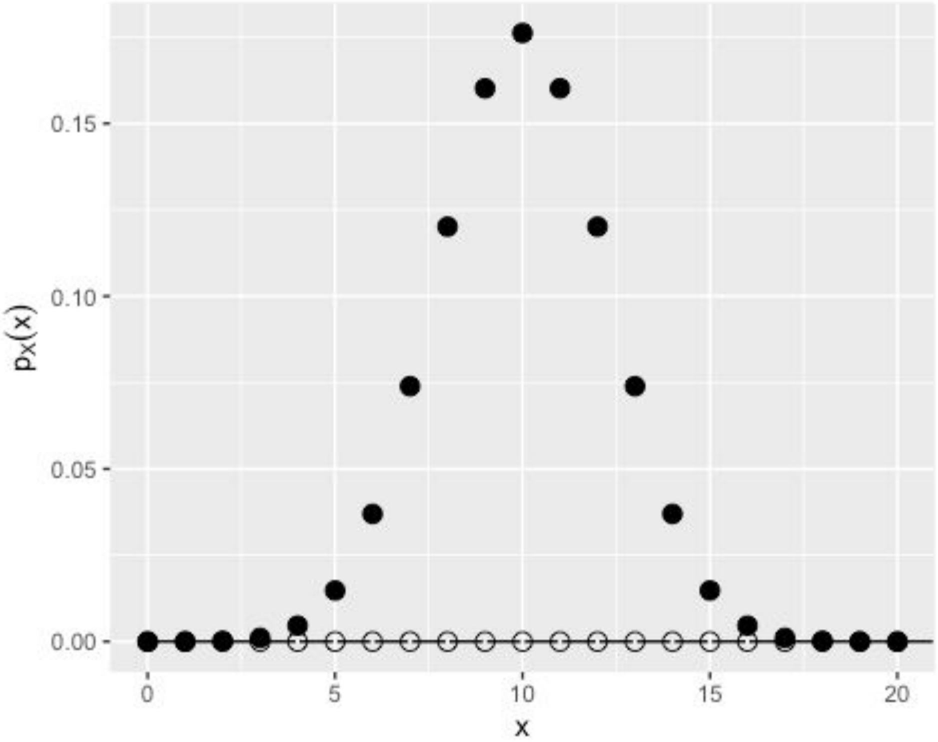
$$\sum_{x \in \mathcal{X}} p_X(x) = 1$$

- **Cumulative distribution function (CDF)**: for discrete X, $P_X(x)$ gives $Pr(X \leq x)$
  - Here we need P to be nondecreasing

$$P_X(b) = \sum_{x \leq b} p_X(x)$$

Probability Mass Function (PMF)

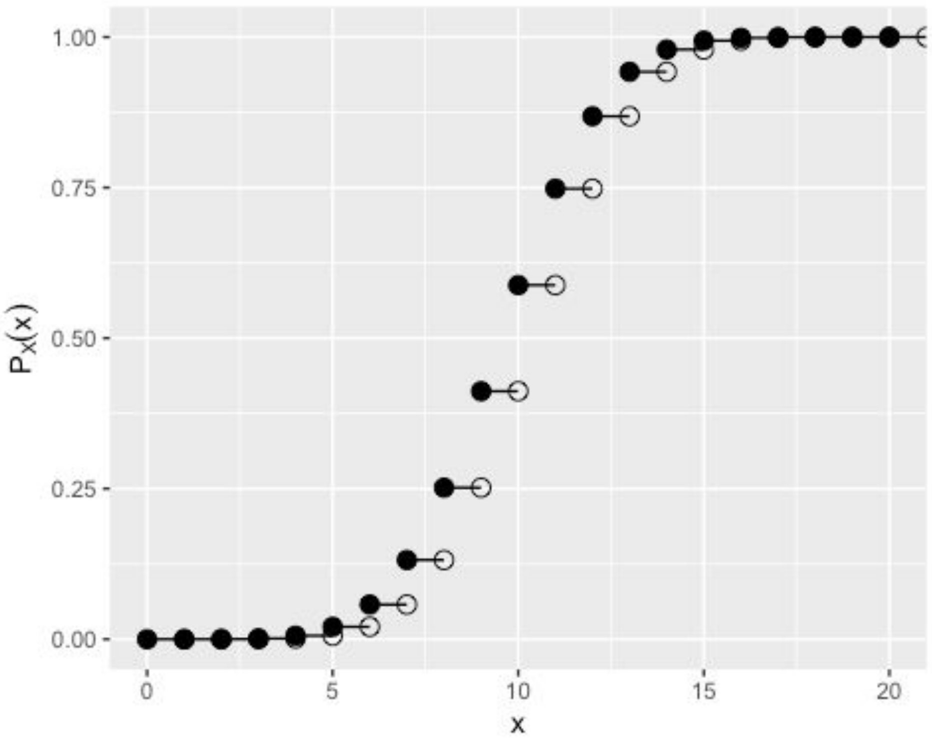
- ex: X is the number of heads counted in 20 coin flips



$$\mathscr{X} = \{0, 1, 2, 3, .., 18, 19, 20\}$$

Cumulative Distribution Function (CDF)

- ex: X is the number of heads counted in 20 coin flips



$$Pr(a < X \leq b) = P_X(b) - P_X(a)$$

Continuous Random Variables

- Continuous random variables take values in intervals of $\mathbb{R}$. Pr(X=x)= 0 for all x. Thus there is no probability mass function.

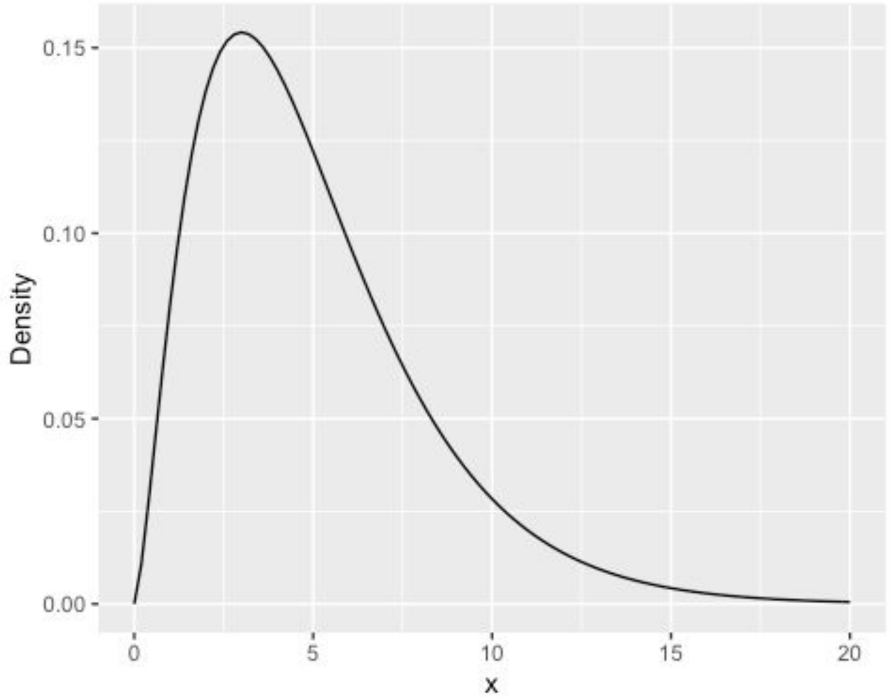
- **Probability Density Function (PDF)**: for a continuous X, we define f_X such that:

$$Pr(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{dx} \text{ and } \forall x\, f_X(x) > 0,\ \int_{-\infty}^{\infty} f_X(x)\,\mathrm{dx} = 1$$

- **Cumulative Distribution Function (CDF)**: for a continuous X, we define F_X such that:

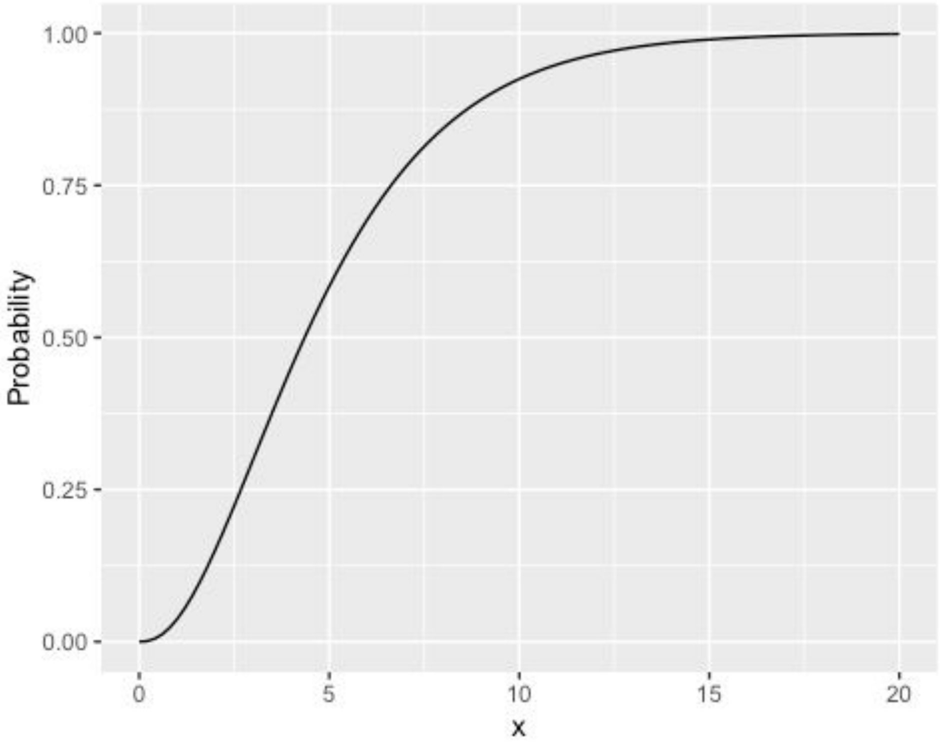
$$F_X(x) = \int_{-\infty}^{x} f(x)\,\mathrm{dx}$$ and F_X gives $Pr(X \leq x) = Pr(X \in (-\infty, x))$

Probability Density Function (PDF)

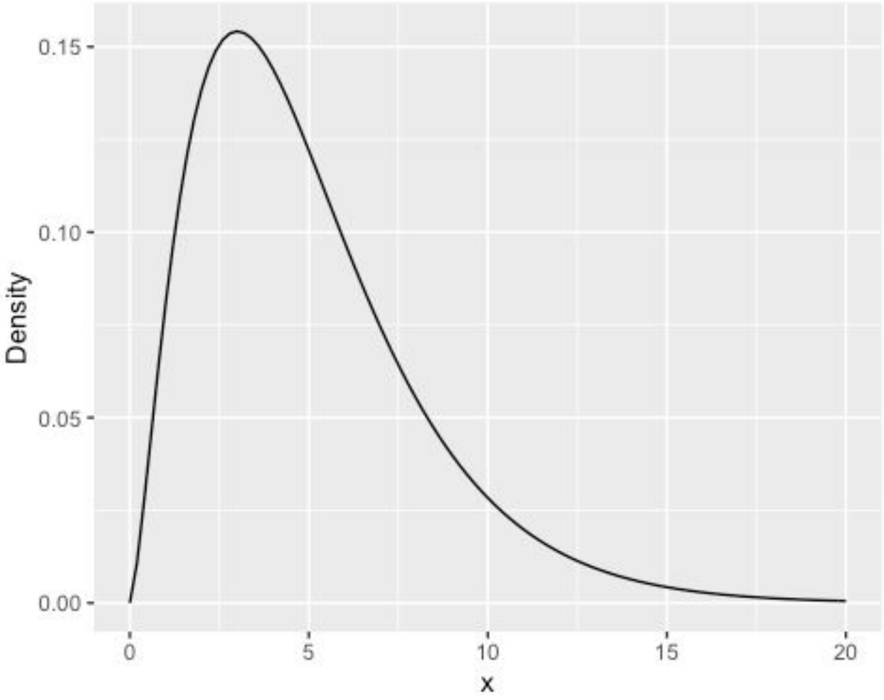


$$Pr(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

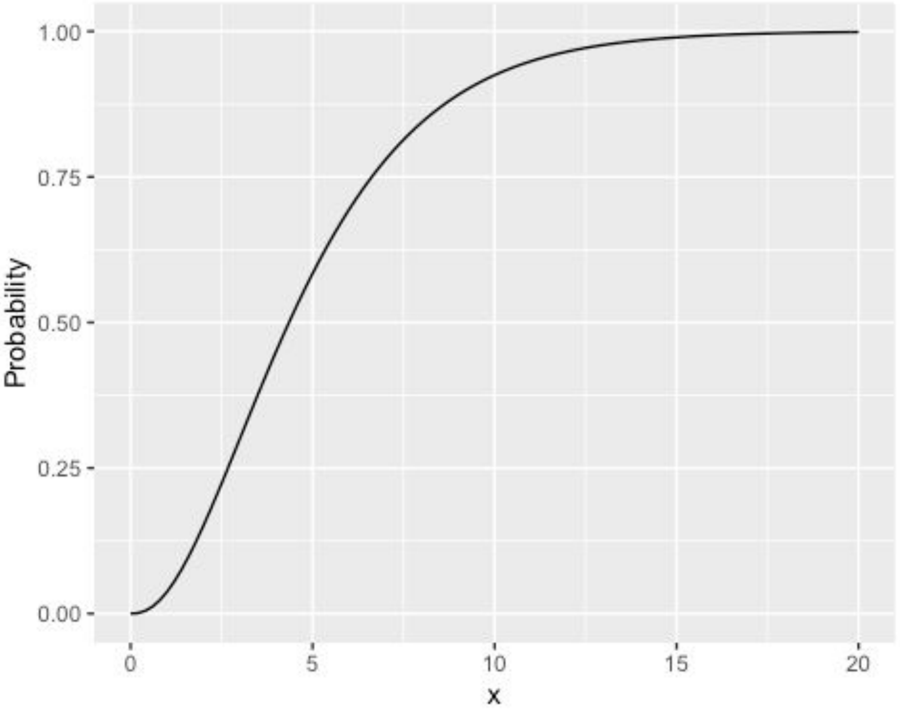
Cumulative Distribution Function (CDF)



$$Pr\left(x_1 < X \leq x_2\right) = F_X\left(x_2\right) - F_X\left(x_1\right)$$



Density
0.15
0.10
0.05
0.00
0
5
10
15
20
x



Probability
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
x

Joint Distributions

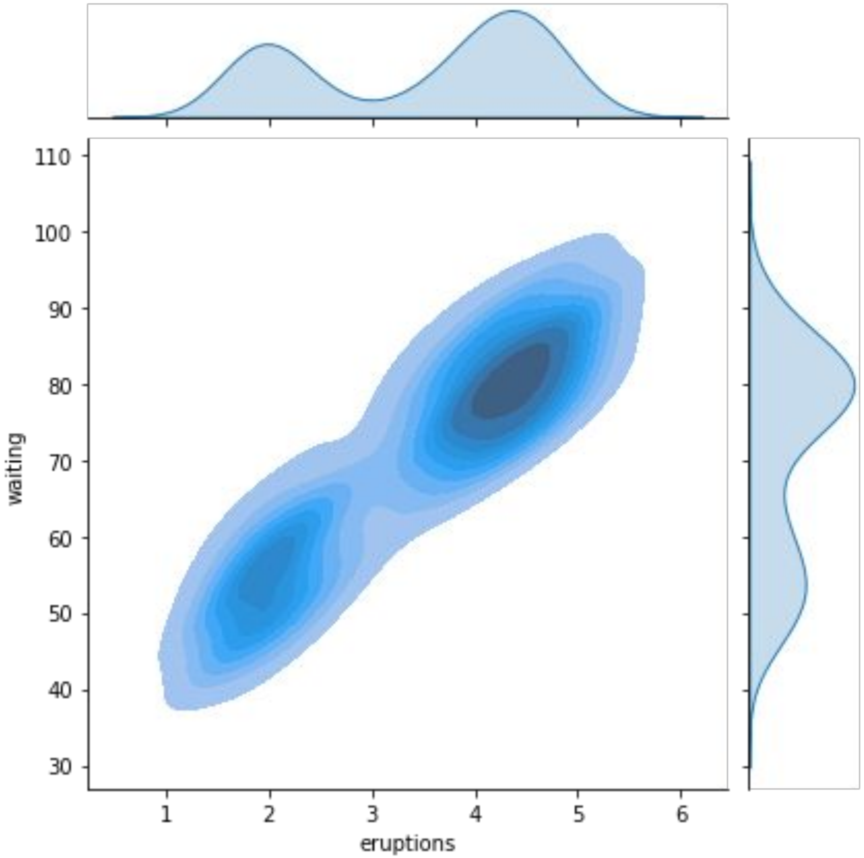
Random variables X and Y have a *joint distribution* if their realizations come together as a pair. (X,Y) is a random vector. Realizations are (x1,y1),(x2,y2),...

Joint CDF: $$Pr(X \leq b, Y \leq d) = F_{X,Y}(b,d)$$

Joint PDF: $$Pr\left[(X,Y) \in \mathscr{A} \subseteq \mathscr{X} \times \mathscr{Y}\right] = \int_{\mathscr{A}} f_{X,Y}(x,y)\,\mathrm{dxdy}$$

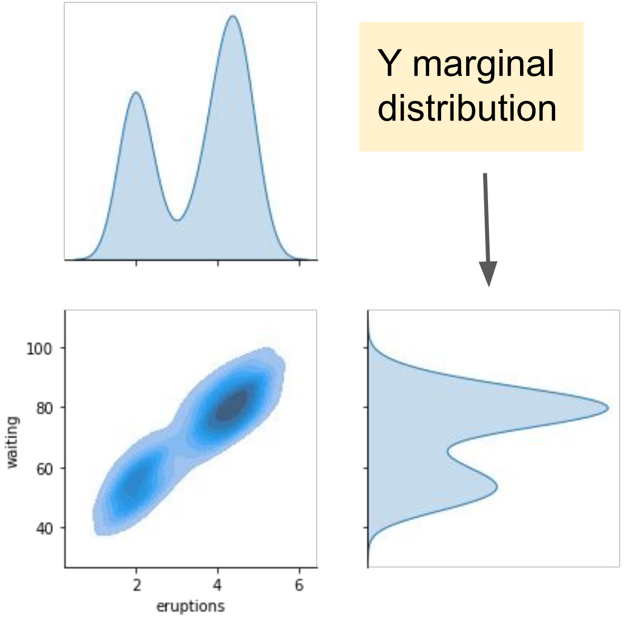
Example - Old Faithful

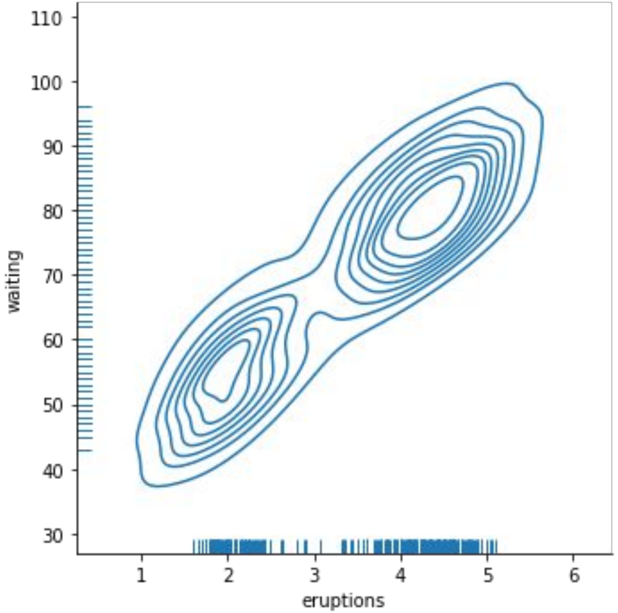
eruptions	waiting
3.6	79
1.8	54
3.333	74
2.283	62
4.533	85
2.883	55
4.7	88
3.6	85
1.95	51



Marginal Distribution

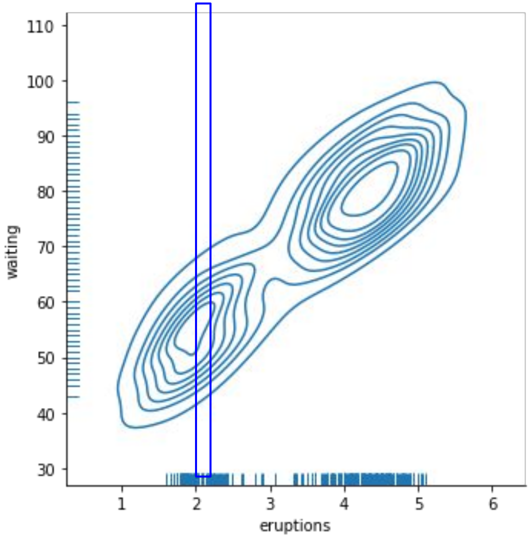
- Ignore X → Given that (X,Y) is random vector, what is the distribution of Y?

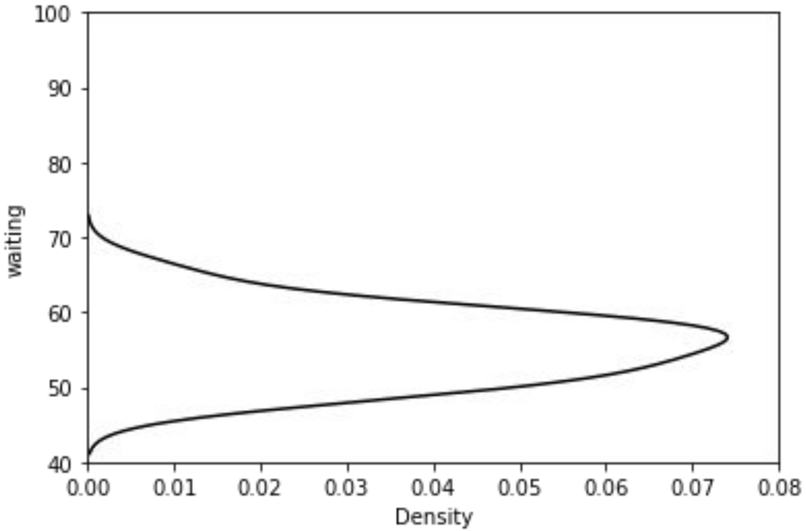




Conditional Distributions

- Given that (X,Y) is a random vector, let's say that we only look at Y values where $X \in [2,2.1]$. We would write this new random variable as $Y \mid X \in [2,2.1]$.
- The distribution describing this random variable is called the *conditional distribution of Y given* $X \in [2,2.1]$. Note: we do not have to use an interval for X (i.e. X=5).





Expected Value

We denote the expected value of a discrete random variable X as:

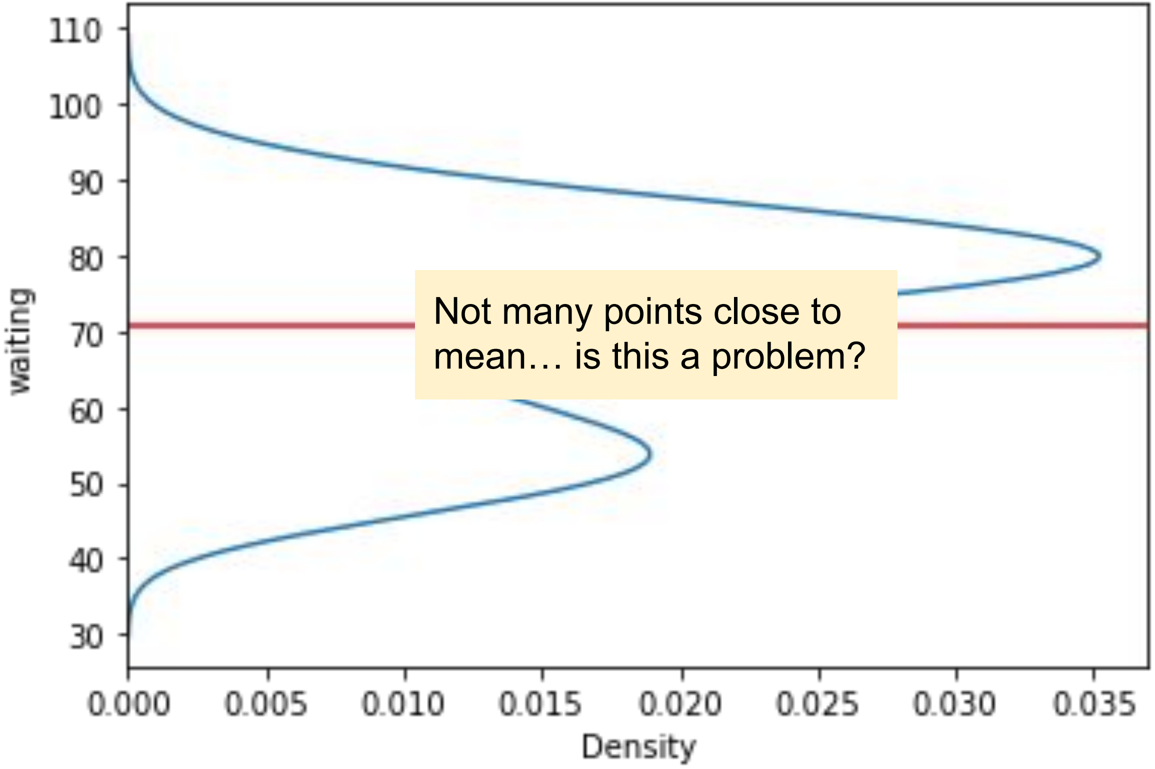
$$E[X] = \sum_{x \in \mathcal{X}} x \cdot p_X(X = x)$$

We denote the expected value of a continuous random variable X as:

$$E[Y] = \int_{y \in \mathcal{Y}} y \cdot f_Y(Y = y)\,\mathrm{dy}$$

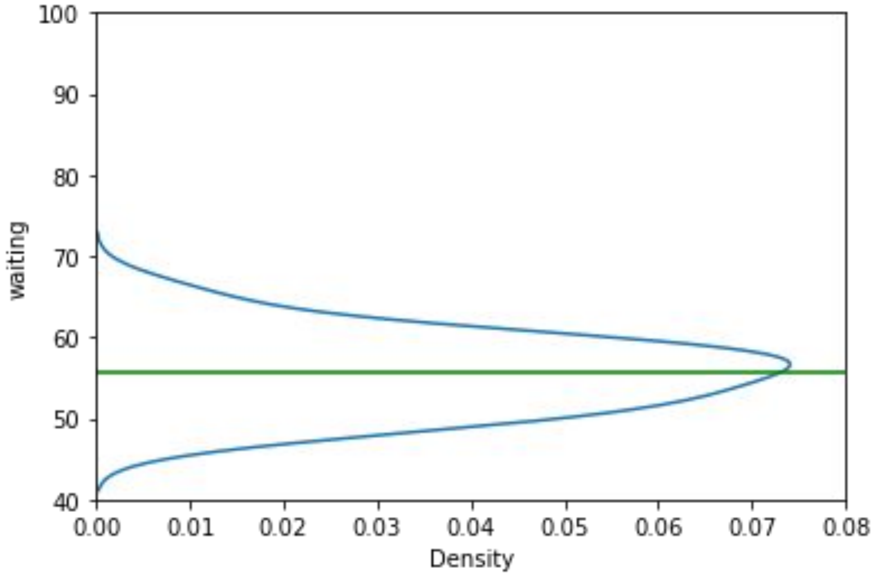
Often we call E[X] the *mean of X* (μ or μ_X). It is the measure of the location of the distribution.

E(Y) for marginal distribution

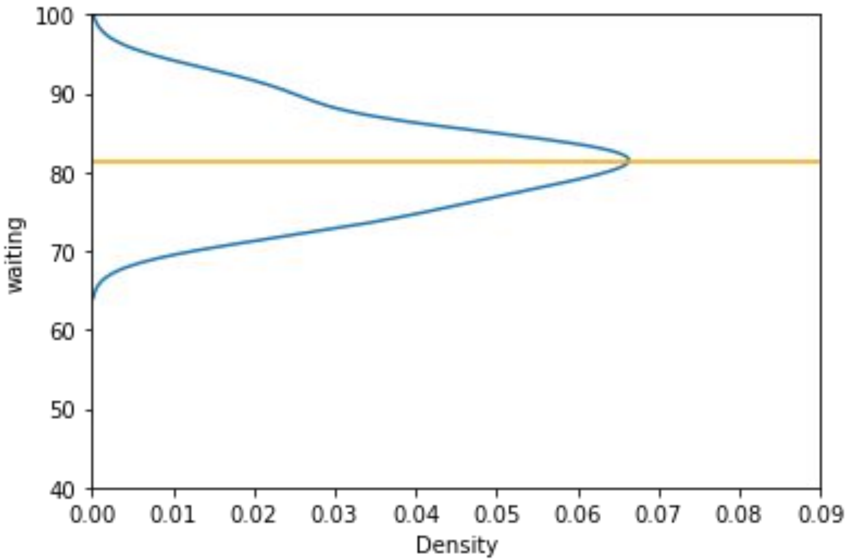


70.8970588235294

What if we know eruption time?



mean: 55.6



mean: 81.33333333333333

Regression Revisited

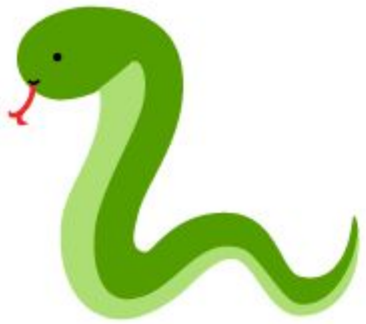
We can restate regression as "an estimation of conditional expected values"

$$\hat{y} = b_0 + b_1 x$$

$$E\left[Y \mid X = x\right]$$

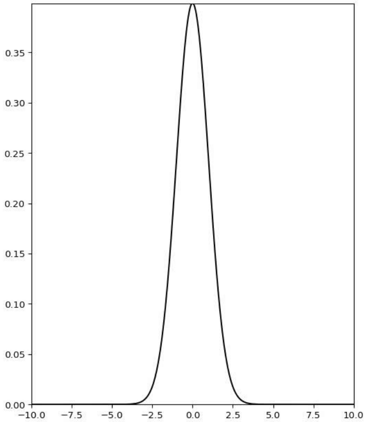
$$E\left[Y \mid X = x\right] = b_0 + b_1 x$$

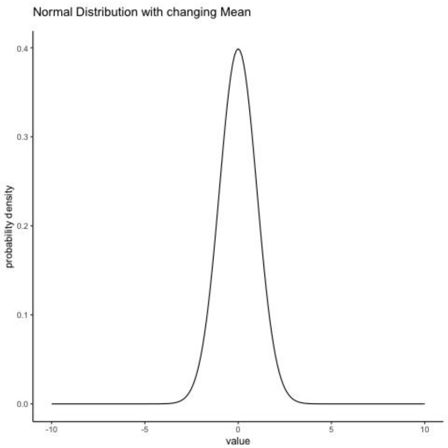
Let’s try it in Python...



Probabilistic Model Estimation

- If we can find a reasonable description of the distribution of some data, we can use that description to infer structure from the data and also **make predictions**.
- Let's consider, as an example, a Gaussian (normal) distribution, which is defined by two parameters: μ_Y (location/mean of distribution) and σ_Y^2 (variance of distribution)

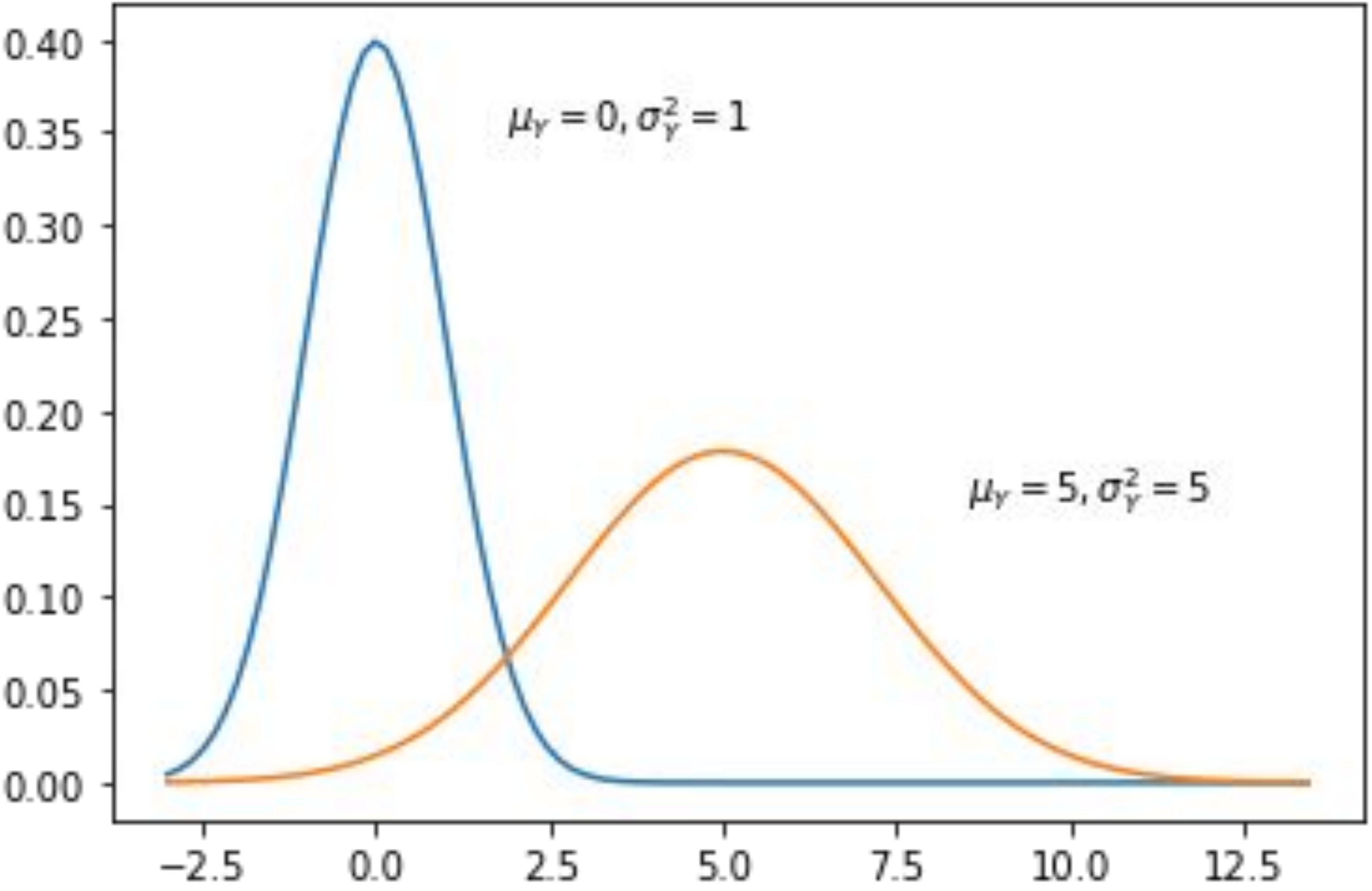




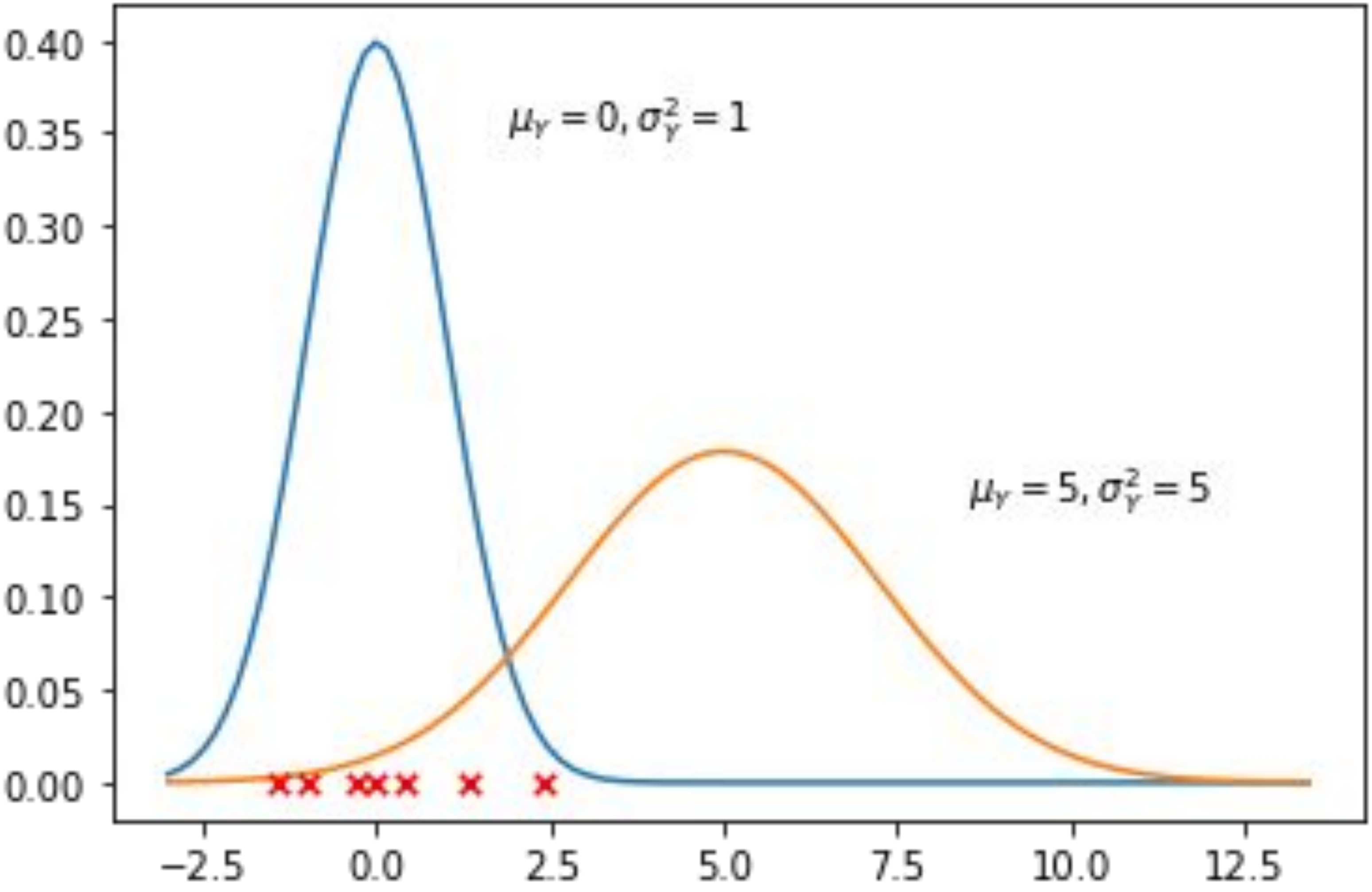
Variance (scale)

Mean (location)

$$f_Y(y) = \frac{1}{\sqrt{2\pi(\sigma_Y)^2}} e^{-\frac{(y-\mu_Y)^2}{2(\sigma_Y)^2}}$$



$\mu_Y = 0, \sigma_Y^2 = 1$
$\mu_Y = 5, \sigma_Y^2 = 5$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5



$\mu_Y = 0, \sigma_Y^2 = 1$
$\mu_Y = 5, \sigma_Y^2 = 5$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5

Likelihood

- Consider a family of distributions with parameters $\Theta \rightarrow$ which distribution in that family is a good match to the data we observe?

- If we have *independent and identically distributed (i.i.d)* data, the probability of seeing all realizations is the product of the probability of each realization

$$\mathcal{L}\left(\theta; y_1, y_2, \ldots, y_n\right) = \prod_i p_Y\left(\theta; y_i\right) \quad (\text{discrete})$$

$$\mathcal{L}\left(\theta; y_1, y_2, \ldots, y_n\right) = \prod_i f_Y\left(\theta; y_i\right) \quad (\text{continuous})$$

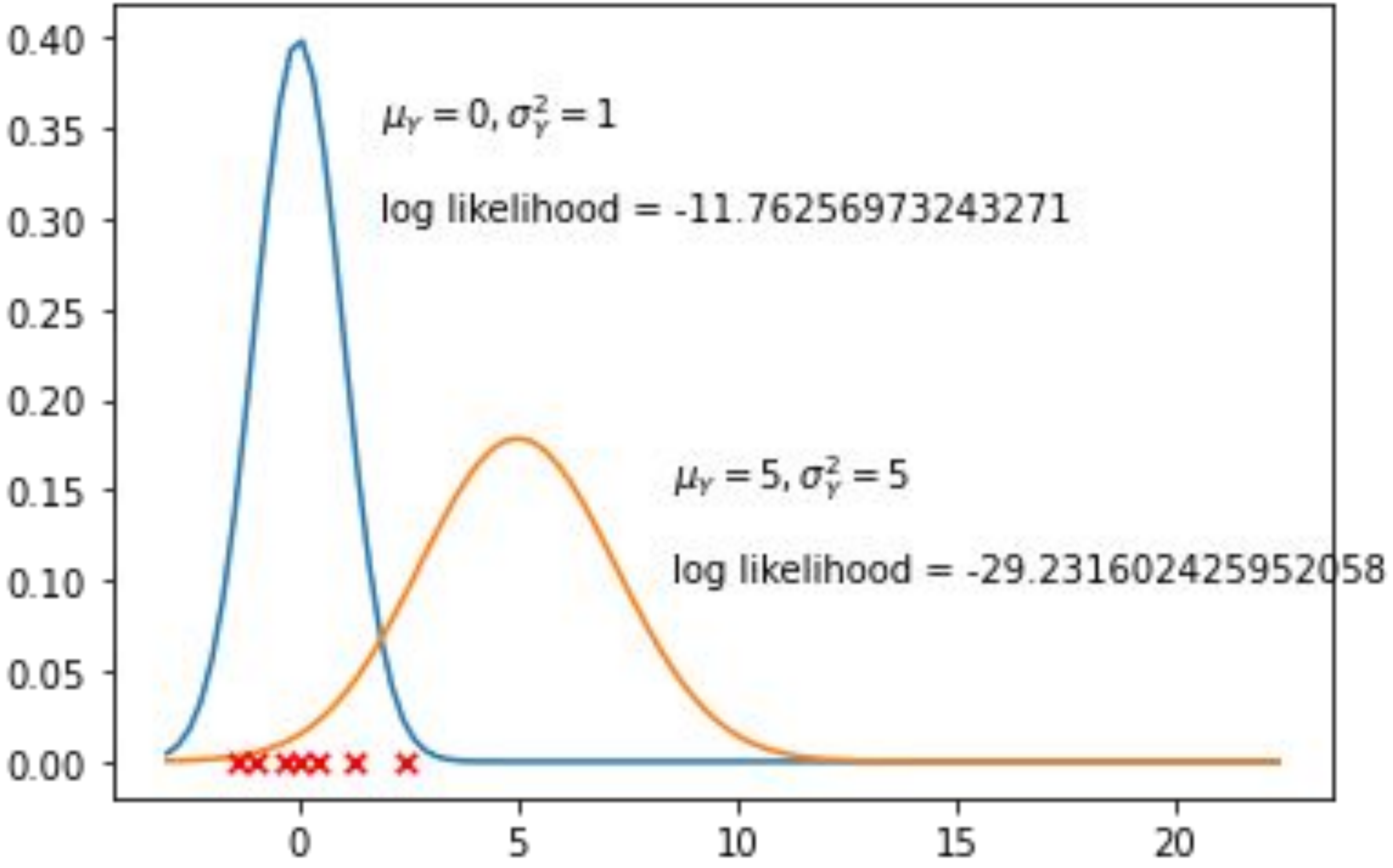
* A collection of random variables is **independent and identically distributed** if each random variable has the same probability distribution as the others and all are mutually independent.

Log Likelihood

- In a practical sense, products of near-zero probabilities are often lost from rounding errors.
- A workaround, therefore, is to consider log likelihood. If one maximizes log likelihood, then they are also maximizing likelihood.
- We note that products of probabilities are turned into sums of log-probabilities here.

$$\ell\left(\theta; y_1, y_2, \ldots, y_n\right) = \sum_i \log\left(p_Y\left(\theta; y_i\right)\right) \quad (\text{discrete})$$

$$\ell\left(\theta; y_1, y_2, \ldots, y_n\right) = \sum_i \log\left(f_Y\left(\theta; y_i\right)\right) \quad (\text{continuous})$$


$\mu_Y = 0, \sigma_Y^2 = 1$
log likelihood = -11.76256973243271
$\mu_Y = 5, \sigma_Y^2 = 5$
log likelihood = -29.23160242595205 8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
5
10
15
20

Maximum Likelihood Principle

1. Identify a set of potential distributions which can describe the data. For example, we could consider the set of all normal distributions.

2. Find the specific distribution in the previously mentioned set which maximizes the (log) likelihood of the data.

3. Using the found distribution, we can then infer and make predictions.

Normal Log Likelihood

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} \rightarrow \log\left(f_Y(y)\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_Y^2\right) - \frac{\frac{1}{2}\left(y-\mu_Y\right)^2}{\sigma_Y^2}$$

$$\ell\left(\mu_Y, \sigma_Y^2; y_1, y_2, \ldots, y_n\right) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_Y^2\right) - \frac{1}{2}\frac{\left(y_i-\mu_Y\right)^2}{\sigma_Y^2}\right]$$

$$\ell\left(\mu_Y, \sigma_Y^2; y_1, y_2, \ldots, y_n\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma_Y^2\right) - \frac{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu_Y\right)^2}{\sigma_Y^2}$$

Which $\boldsymbol{\mu}_Y$ gives highest likelihood?

Which $\boldsymbol{\sigma}_Y^2$ gives highest likelihood?

Maximum Likelihood Estimation (MLE)

Which $\boldsymbol{\mu}_{\mathrm{Y}}$ gives highest likelihood?

$$\frac{\partial \ell}{\partial \mu_Y} = \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} \left(y_i - \mu_Y \right)$$

$$\frac{\partial \ell}{\partial \mu_Y} = 0 \Leftrightarrow \mu_Y = \frac{\sum_{i=1}^{n} y_i}{n}$$

Which $\boldsymbol{\sigma}_{\mathrm{Y}}^2$ gives highest likelihood?

$$\frac{\partial \ell}{\partial \sigma_Y^2} = -\frac{n}{2\sigma_Y^2} + \frac{\sum_{i=1}^{n} \left(y_i - \mu_Y \right)^2}{2\sigma_Y^4}$$

$$\frac{\partial \ell}{\partial \sigma_Y^2} = 0 \Leftrightarrow \sigma_Y^2 = \frac{\sum_{i=1}^{n} \left(y_i - \mu_Y \right)^2}{n}$$

Coming back to regression...

Recall that in the last lecture we built models of the following form:

$$\hat{y} = \boxed{b_0} + \boxed{b_1} x$$

Least Squares from MLE

- Maximum Likelihood Estimation (MLE) tells us which distribution to select (i.e. from a set of normal distributions) to fit your data. From this we can predict the best parameters.
- When applying the Maximum Likelihood Principle on a *model such that* Y is normally distributed, mean is $b_0 + b_1x$, and variance is $\boldsymbol{\sigma}_{\varepsilon}^{2}$, we essentially get OLS Regression

$$f_Y(y \mid X=x) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon}^{2}}} e^{-\frac{\left(y-\left(b_0+b_1x\right)\right)^2}{2\sigma_{\varepsilon}^{2}}}$$

mean $b_0 + b_1x$

variance $\boldsymbol{\sigma}_{\varepsilon}^{2}$

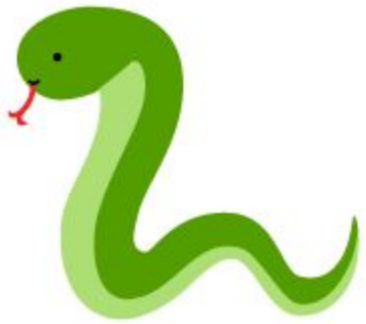
$$\ell\left(b_0, b_1, \sigma_{\varepsilon}^{2}; \left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_n, y_n\right)\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma_{\varepsilon}^{2}\right) - \frac{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \left(b_0 + b_1x_i\right)\right)^2}{\sigma_{\varepsilon}^{2}}$$

Maximum Likelihood Regression

1. Choose the form of the function
2. Choose the form of the distribution
3. Use the function/distribution to determine the likelihood of the data
4. Use an optimizer to find parameters Θ which maximize the likelihood.

Why use MLR? It serves as a foundation which can be adjusted (i.e. by regularization [later lesson]).
Also see “Generalized Linear Models”: https://en.wikipedia.org/wiki/Generalized_linear_model

Let's try it in Python...



In Summary

- Probabilities and Events
- Random Variables
  - Discrete and continuous
- Distributions
  - PMF, PDF, CDF
  - Joint, Marginal, Conditional
- Expected values
- Likelihood, Log Likelihood
- Maximum Likelihood Regression