Lecture #1

Notes for 6822

Read: Murphy 2.1-2.4, 2.6, 2.8
3.1-3.2
Murphy 4.2

Lecture#1

Definitions

short definition (UWash)

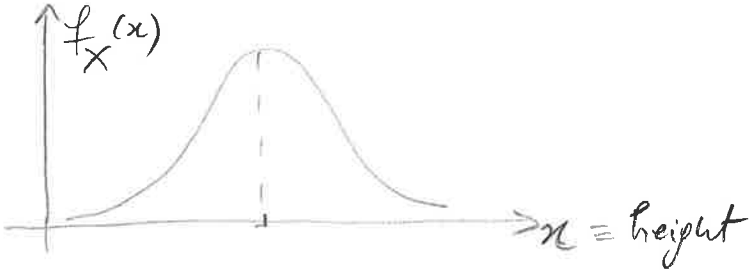
> **Random variable** (A variable that takes on different values determined randomly)

Long definition (Yates)

- Consists of:
  (1) an experiment with a probability measure P[·] defined on a sample space S
  (2) a function that assigns a real number to each outcome in the sample space of the experiment.
- when we observe one of these numbers, we refer to the observation as a random variable; X with set of possible values in $S_X \triangleq$ range of X.

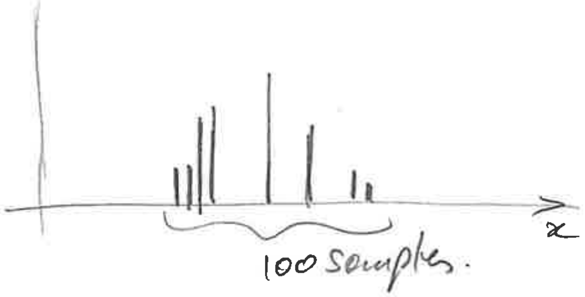
Example: Height of a person from the US.

> **Probability Distribution Function** (for continuous random variables)



> **Sampling from a distribution**

Consider a sample of 100 heights of people from the US drawn randomly from the distribution of all heights



Using Yates def. of random variables. In this case the "function" to assign values to the r.v. is the outcome itself.

Independence

- Let X, Y be random variables
  - X is the outcome of first roll of a 6-sided dice
  - Y is the outcome of second roll of the dice.
- They take values $S_X = S_Y = \{1,2,3,4,5,6\}$ with equal probability

  P[·] says each side is $\frac{1}{6}$ probability.

- An event is a statement about the world that holds or not:

  (Yates: Event is a set of outcomes of an experiment)

  I like it better! Experiment is "Rolling Dice" in this case.

  Note: set could be also just one outcome of the experiment in a particular case.

Define 3 events:

$$\begin{cases} A = \{X \in \{3,4\}\} \\ B = \{X = 1\} \\ C = \{Y \in \{3,4\}\} \end{cases}$$

- Events are assigned probability:

$$P(A) = P[X \in \{3,4\}] = \frac{2}{6} = \frac{1}{3}$$
$$P[B] = \frac{1}{6}$$
$$P[C] = \frac{1}{3}$$

- For any events U, V, we have:

$$\boxed{P(U \cup V) = P(U) + P(V) - P(U \cap V)}$$

- Any events U, V are independent if and only if $\boxed{P(U \cap V) = P(U) \cdot P(V)}$

  also notation: $\boxed{P[UV] = P[U] \cdot P[V]}$

Q: Are A, B independent?

$P[A] = \frac{1}{3}$ $\quad P[B] = \frac{1}{6}$ $\Rightarrow P[A] \cdot P[B] = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}$

$P[A \cap B] = 0$ so, they are not independent.

How about B, C? yes

How about A, C. yes

- Conditional Probability $\quad \boxed{P(U|V) = \frac{P(U \cap V)}{P(V)}}$

$$P(X \leq 4 \mid X \geq 3) = \frac{P(3 \leq X \leq 4)}{P(X \geq 3)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

- if U, V independent then:

$$\boxed{P(U|V) = \frac{P(U \cap V)}{P(V)} \downarrow \frac{P(U)\cancel{P(V)}}{\cancel{P(V)}} = P(U)}$$

Mean, Variance of a Discrete random variable

Mean: $\mu_X = E[X] = \sum_x x \cdot P(X=x)$ Expected value of X is weighted by the probability of seeing it!

Variance: $Var[X] = \sigma_X^2 = E[(X-\mu_X)^2]$ The expected squared Deviation from its mean.

(captures spread in your data)

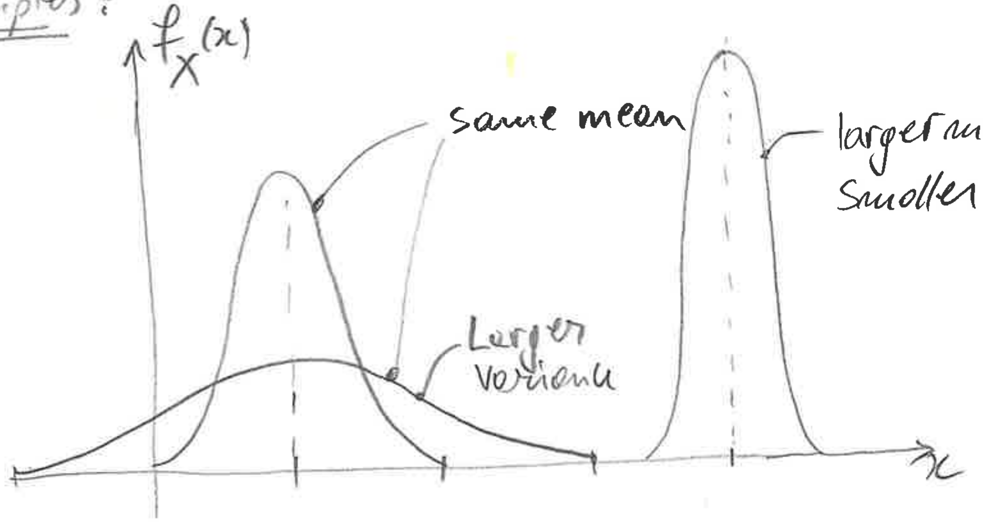
Median: M the value of X that separates the higher half of its range from the lower half.

$$\boxed{P[X \leq M] = \frac{1}{2}}$$

(Yates definition: Median M is a number that satisfies:

$$\boxed{P[X<M] = P[X>M]}$$

Examples:



Mean - is a prediction of the value of the random variable

- answers question: "what do I expect the height of a random person to be?"

Variance - captures the spread of the Data.

- captures the error in the prediction using Mean.
- "How much do people's height deviate?".

Appendix $\boxed{A}$:

Note: it is interesting to note that is coin is indeed "special" or hacked such that it lands Heads 70% of time, then the $\hat{\theta}_{MLE}$ should capture/model that!

$\boxed{Q}$: what is difference between Likelihood and Probability in this context?

→ Probability describes the likelihood of an event occurring

→ Likelihood describes the plausibility of a parameter value given the observed Data! Use in statistical inference to estimate model parameters.!

Sometimes, people use probability to mean likelihood ?! Should be clear from context that it is about likelihood.

$\boxed{Q}$: What exactly do we mean by "probability of observed data"? ≡ Likelihood actually ?!

It means that the model is desired to capture the happening of say $\{HHTHT\ldots HT_n\}$ the best way; we fine-tune or calibrate θ (or search for θ) that will increase the chances (will actually maximize) the chances for the model to say that that particular sequence will happen! (well actually, just the fact that $\boxed{k}$ flips are Heads matter out of $\boxed{n}$)

Maximum Likelihood Estimation | Part I: Discrete | MLE

- A popular discrete random variable is: Bernoulli Random Variable X has the probability mass function (PMF) in the form:

$$P_X(x) = \begin{cases} 1-p, & x=0 \\ p, & x=1 \\ 0, & \text{otherwise} \end{cases}$$

where $\boxed{p}$ is a parameter $0<p<1$.

- Problem: Have a special coin, flip it, what's the probability it will be Heads?
  - Collect Data: HH => Probability 100% Heads 0% Tails.
  - collect more Data: HHTHT => Probability 60% Heads 40% Tails
  - Collect even more Data, say 10000 times and it comes out to be: Probability 70% Heads 30% Tails.
  - ....

- Let's do some math:
- Coin flip can be modeled by a Bernoulli Distribution. (← My Hypothesis)

outcomes of experiment ≡ events.

Data: sequence $D = (\underbrace{\text{H T H H T}, \dots}_{})$ $\boxed{K}$ Heads out of $\boxed{n}$ flips

Number of Heads is a Binomial random variable! (see Yates)

Hypothesis: $\begin{cases} P(H) = \boxed{\theta} \\ P(T) = 1-\theta \end{cases}$ and Flips are i.i.d. $\begin{cases} \text{- independent events} \\ \text{- identically distributed according to Bernoulli Distribution.} \end{cases}$

Assume Basically model this as a Bernoulli

is the "p" in Bernoully.

Likelihood:

should use;

$$P(D|\theta) = P(\text{H T H H T}, \dots)$$

$$P(D|\theta) = P(H) \cdot P(T) \cdot P(H) \cdot \dots \quad \leftarrow \text{by independence.}$$

$$\boxed{P(D|\theta) = \theta^k \cdot (1-\theta)^{n-k}}$$

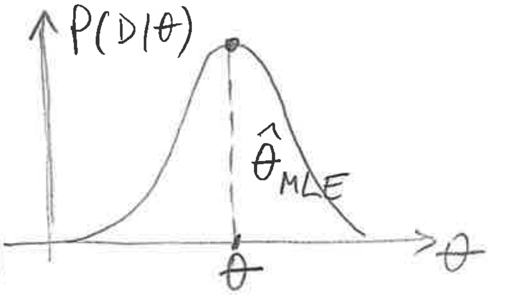
Should be called likelihood in this context.

- Maximum Likelihood Estimation (MLE): Choose $\boxed{\theta}$ that maximizes the "probability" of observed Data":

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

$$= \arg\max_{\theta} \underbrace{\log P(D|\theta)}_{\triangleq \ell(\theta)}$$

: log because it's easy (easier than differentiating products)



Appendix $\boxed{B}$:

→ Binomial Random Variable (Discrete):

X is a binomial if the PMF has the form:

$$p_X(x) = \begin{cases} \binom{n}{x} p^{x}(1-p)^{n-x} & , x = 0, 1, 2, \ldots, \boxed{n} \\ 0 & , \text{otherwise.} \end{cases}$$

where $\begin{cases} 0 < p < 1 \\ n \geq 1 \end{cases}$

and

$$C_n^x \equiv C(n,x) \equiv \boxed{\binom{n}{x}} = \frac{n!}{x!\,(n-x)!}$$ is the Binomial coefficient

commonly used notation "n choose x"

if $\boxed{n=1}$ this becomes the Bernoulli ⟹ $\begin{cases} \text{for } x=0 \Rightarrow & (1-p) \\ \text{for } x=1 \Rightarrow & n \cdot p \cdot (1-p)^{n-1} \overset{n=1}{=} p \end{cases}$

[Yates]: whenever we have a sequence on $\boxed{n}$ independent trials and each w/ a success probability $\boxed{p}$, the number of successes is a binomial variable.

Mean: $\mu_X = E[X] = n \cdot p$ $\Rightarrow E\left[\frac{k}{n}\right] = \frac{\cancel{n} \cdot \theta^*}{\cancel{n}} = \theta^*$ ⟶

$\theta^* \equiv$ my $\boxed{p}$

Variance: $Var[X] = np(1-p)$ $\Rightarrow Var\left[\frac{k}{n}\right] = \frac{\cancel{n}\theta^*(1-\theta^*)}{n^{\cancel{2}}} = \frac{\theta^*(1-\theta^*)}{n}$

used. over next page

[Yates] Use Theorem: if $Y = aX$, $Var[Y] = a^2 Var[X]$
pp 73

- In our Problem:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log\left(\theta^{k} \cdot (1-\theta)^{n-k}\right)$$

- Take derivative and set to zero; find θ that satisfies that equation:

$$\boxed{\frac{d}{d\theta} \log P(D|\theta) = 0} \quad MLE$$

- So: $\hat{\theta}_{MLE} = \arg\max_{\theta} \log\left[\theta^{k} \cdot (1-\theta)^{n-k}\right]$

$$= \arg\max_{\theta} \left[\underbrace{k \cdot \log\theta + (n-k) \cdot \log(1-\theta)}\right]$$

- Use: $\boxed{\frac{d}{dx} \log u(x) = \frac{u'}{u}}$ $\boxed{u=\theta,\ u=(1-\theta)}$ to find derivative of , which we set to zero:

$$\frac{d}{d\theta}\left[k \log\theta + (n-k)\log(1-\theta)\right] = \boxed{\frac{k}{\theta} - \frac{n-k}{1-\theta} = 0}$$

$$k - \cancel{k\theta} = n\theta - \cancel{k\theta}$$

$$k = n\theta \quad \Rightarrow \quad \boxed{\hat{\theta}_{MLE} = \frac{k}{n}} = \frac{3}{5} \qquad 60\%$$

↑ for {HHTHT} this run of experiments

This is our first learning algorithm !!!

How good is MLE ?

→ We treat MLE $\boxed{\hat{\theta}_{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates data $\mathcal{D}=(HHTTH\ldots)$ of fixed size $\boxed{n}$.

What can we say about random variable $\hat{\theta}_{MLE}$?

Use fact that K as number of successes is Binomial! So it has mean np

- Unbiased (good property of MLE for Binomial)

Definition: $\boxed{\text{Bias}(\hat{\theta}_{MLE})} \triangleq E_{\boxed{\mathcal{D}\sim P_{\theta^*}}}[\hat{\theta}_{MLE}] - \theta^* = E\left[\frac{k}{n}\right] - \theta^* = \frac{\theta^* \cdot n}{n} - \theta^* \to 0$

↑ "True Predictor"

- Expectation describes how the estimator $\hat{\theta}_{MLE}$ behaves on average

Data $\mathcal{D}$ is comes from a true distribution with parameter θ^*

Appendix $\boxed{C}$

What does "bias" mean?

> For an estimator $\hat{\theta}$ of parameter θ:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \underset{\uparrow \text{true}}{\theta^*}$$

> Unbiased estimator: $\mathbb{E}[\hat{\theta}] = \theta^*$

On average, across repeated samples, the estimator hits the true parameter θ

> Biased estimator: $\mathbb{E}[\hat{\theta}] \neq \theta^*$

On average, it systematically overshoots or undershoots.

MLE and Bias

> MLE is not guaranteed to be unbiased.

- for some params / distributions, MLE is unbiased.
- for others, it's biased, but often bias bias shrinks as sample size $\boxed{n} \to \infty$

> Shrinking bias is why MLE estimators are called "consistent estimators" (they converge to the true parameter with enough data!)

- $\boxed{\text{How many flips do we need ?}}$

→ Consider running many experiments with $\theta^* = \frac{3}{5}$ (it is how the special coin is) and observe many <u>instances</u> of the random variable:

$$\hat{\theta}_{MLE} = \frac{k}{n}$$

Experiment 1: $\boxed{n=5}$ flip coin five times, get 2 heads:

$$\hat{\theta}_{MLE} = \frac{2}{5}$$

Experiment 2: $\boxed{n=50}$ flip coin fifty times, get 30 heads:

$$\hat{\theta}_{MLE} = \frac{30}{50} = \frac{3}{5}$$

<u>Question</u>: they are both unbiased, which one is right? why?

$\boxed{\text{Variance}}$ goes down with larger n:

$$\sigma_{\hat{\theta}_{MLE}} = \sqrt{Var(\hat{\theta}_{MLE})} \underset{\substack{\uparrow \\ \text{Appendix B}}}{=} \sqrt{\frac{\cancel{n}\theta^*(1-\theta^*)}{n^{\cancel{2}}}} = \sqrt{\frac{\theta^*(1-\theta^*)}{n}} \xrightarrow[n\to\infty]{} 0$$

So, more data ⟹ better performance!!!

$\boxed{\text{Summary}}$:

<u>Observe</u>: $X_1, X_2, \ldots, X_n$ drawn iid from $f(x;\theta)$ for some $\boxed{\text{"true"}}$ $\theta = \theta^*$

Likelihood function:

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

<u>Choice!</u>

Log-Likelihood function:

$$\ell_n(\theta) = \log[L_n(\theta)] = \sum_{i=1}^{n} \log(f(X_i;\theta))$$

<u>Easier</u>!

Maximum Likelihood Estimator (MLE):

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$$

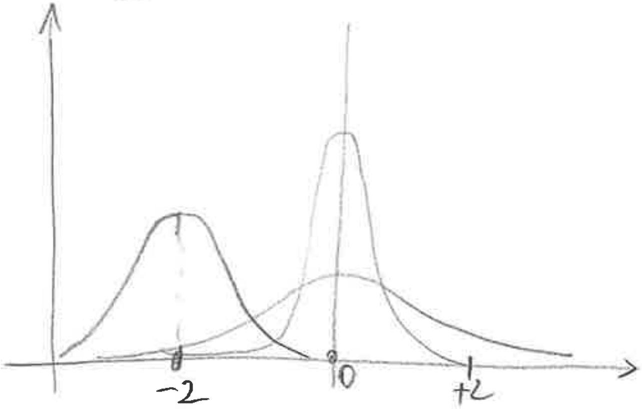
Part II: Continuous

→ What if we are "measuring" a continuous variable?

A: Let me tell you about $\boxed{\text{Gaussians}}$...

$$f_X(x) \equiv \boxed{P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

(PDF) w/ parameters ($\boxed{\mu}$, $\boxed{\sigma}$.)



Given a set of iid samples from Gaussian Distribution, find $\theta = [\mu, \sigma]$

→ Some properties of Gaussians

- Affine transformation (multiply by scalar and add a constant)

$$X \sim N(\mu, \sigma^2)$$

$$Y = aX + b \quad \Rightarrow \quad Y \sim N(a\mu + b, a^2\sigma^2)$$

- Sum of Gaussians

$$X \sim N(\mu_X, \sigma_X^2)$$

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

$$Z = X + Y \quad \Rightarrow \quad Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

- **MLE for Gaussian**

- Probability of iid samples $D=\{x_1,x_2,\dots,x_n\}$ (e.g., temperature)

$$P(D|\mu,\sigma)=P(x_1,x_2,\dots,x_n|\mu,\sigma)\overset{\text{iid}}{=}$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\prod_{i=1}^{n}\cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\quad : \text{Likelihood.}$$

- Log-Likelihood (LL) of data:

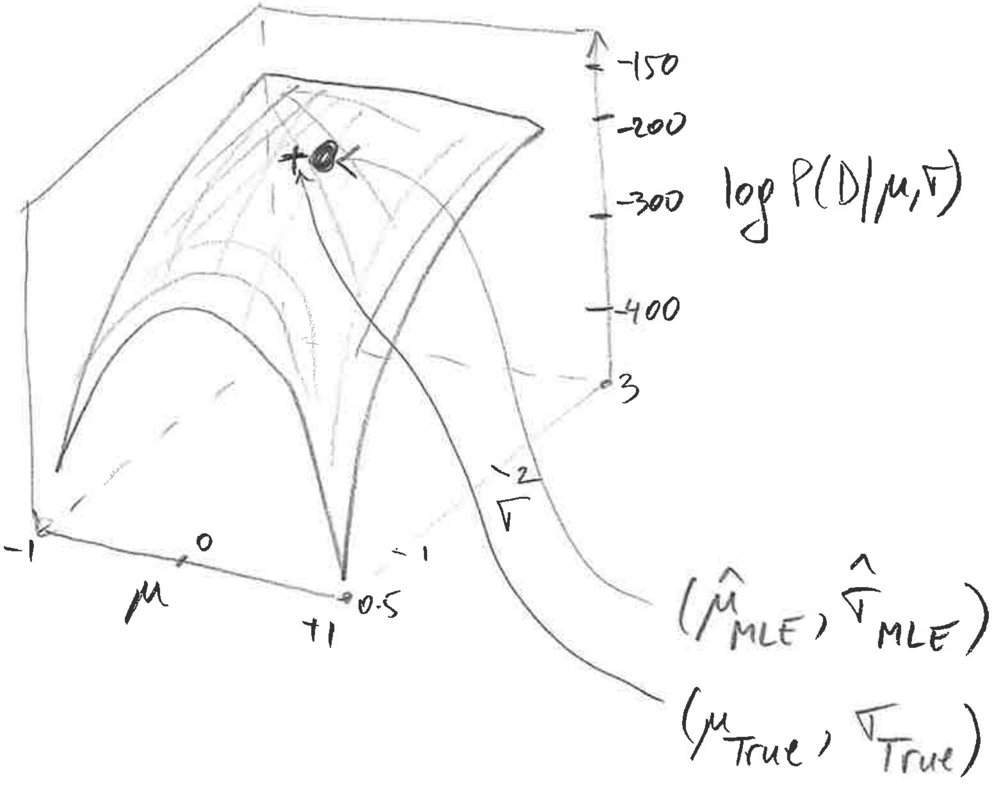
$$\log P(D|\mu,\sigma)=-n\cdot\log(\sigma\sqrt{2\pi})-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- Generate $D=\{x_1,x_2,\dots,x_{n=100}\}$ with $n=100$ samples from Gaussian Distribution.

$x_i\sim N(\mu,\sigma^2)$

$\mu=0\ ,\ \sigma^2=1$

$$\log P(D|\mu,\sigma)=-n\log(\sigma\sqrt{2\pi})-\sum_{i=1}^{100}\frac{(x_i-\mu)^2}{2\sigma^2}$$



→ Like in the Discrete case, to find maximum, set partial derivative to zero:

$$\frac{\partial}{\partial\mu}\log P(D|\mu,\sigma) = \frac{\partial}{\partial\mu}\left[-n\cancel{\log(\sigma\sqrt{2\pi})} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

to find max.

$$= \cancel{-}\sum_{i=1}^{n}\frac{\cancel{+2}\cdot(x_i-\mu)}{\cancel{2}\sigma^2} = \frac{1}{\sigma^2}\cdot\left(-n\mu+\sum_{i=1}^{n}x_i\right) \overset{\downarrow}{=} 0$$

$$\Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}x_i}$$

This is your second learning algorithm !!!

→ Do the same for Variance:

$$\frac{\partial}{\partial\sigma}\log P(D|\mu,\sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right] =$$

Use: $\boxed{\frac{d}{dx}\log x = \frac{1}{x}}$ $\boxed{\frac{d}{dx}\frac{1}{x^2} = -\frac{2}{x^3}}$

$$= -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{\cancel{2}\cdot(x_i-\mu)^2}{\cancel{2}\sigma^3} = \frac{-n\sigma^2+\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{MLE})^2} \;!$$

Learning Gaussian Parameters:

→ MLE

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{MLE})^2$$

→ MLE for Variance is Biased!

$$E[\hat{\sigma}^2_{MLE}] \neq \sigma^2$$

→ Unbiased variance estimator:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu}_{MLE})^2$$

Note: see derivation in additional notes!

Summary: Data $\boxed{\mathcal{D}}$

Observe: $X_1, X_2, \dots, X_n$ drawn iid from $f(x;\theta)$ for some "true" $\theta = \theta^*$

Likelihood function: $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

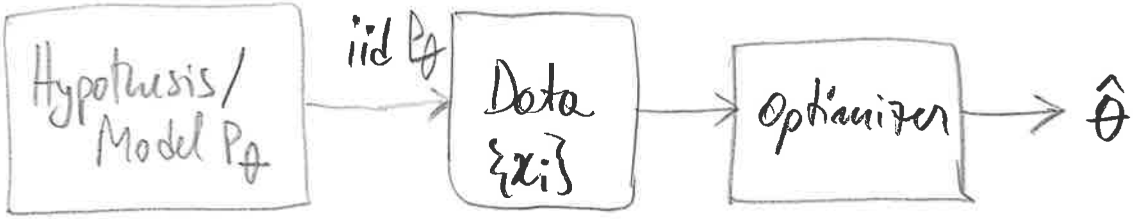
Log-Likelihood function: $\ell_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

MLE: $\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

- As the number of observations $n \to \infty$, we have $\hat{\theta}_{MLE} \to \theta^*$
- The MLE is a "recipe" that begins with a model for data $f(x;\theta)$.

Recap

- Learning is:
  - Collect some data (e.g., coin flips)
  - Choose a hypothesis class or model (e.g., Bernoulli)
  - Choose a loss function (e.g., Data likelihood)
  - Choose an optimization procedure (e.g., set derivative to zero to obtain MLE)


Hypothesis/ Model P_θ
iid P_θ
Data $\{x_i\}$
Optimizer
$\hat{\theta}$