Probability
&
Maximum Likelihood Estimation

Sample Space of an Experiment

- A random experiment is one whose outcome cannot be predicted with certainty.
  e.g. Toss of a coin, roll of a dice, etc.

- **Sample space** (S) represents the set of all possible outcomes of the experiment.

  e.g. Toss of a coin = {H, T}
  Roll of a single dice = {1, 2, 3, 4, 5, 6}
  Roll of two dice = {11, 12, …, 16,
  21, 22, ..., 26,
  ...,
  61, 62,..., 66 }

Sample Space and Event

- An event (E) is a subset of S (sample space).
- Example:
  - Tossing a coin:
    S = {H, T} E = {H} is an event.
  - Tossing a coin twice:
    S = {HH, HT, TH, TT}
    E = {HH, HT} is an event which describes the set which has the first outcome as heads
  - Tossing a dice:
    S = {1, 2, 3, 4, 5, 6}
    E = {2, 4, 6} is an event which describes the set in which the number is even.

Random Variable

- Suppose to each outcome in the sample space, we associate a value.
  e.g. for heads, +1 and for tails -1
  for even rolls of dice +1 and for odd rolls of dice -1
- Such a variable is known as a random variable.
- Formal definition:
  A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.
- It is generally denoted by X.
  If X takes a value x, it is written as: X = x

Random Variable

- If we restrict the random variable to the Boolean set, it may be defined as a function f:

f is a function defined over S as follows:

$$f: S \rightarrow \{0, 1\}$$

f maps every value in S to either 0 (failure) or 1 (success)

Random Variable

- If we restrict the random variable to the Boolean set, it may be defined as a function f:

f is a function defined over S as follows:

$$f: S \rightarrow \{0, 1\}$$

f maps every value in S to either 0 (failure) or 1 (success)

Examples of Random Variables:

- Suppose each experiment is tossing of 4 coins. You do this experiment multiple times
  Results:
  {HHHT}
  {HTHT}
  …
  If you count the number of heads in each experiment, it would be a random variable (X)
  X could have values from 0 to 4.
- Each element of the sample space would map to one value of X

Examples of Random Variables:

- Suppose each experiment is sampling 100 people from a population and measure their heights.

  If you measure the average of the heights of the 100 samples, you would get a random variable (X). It would be a continuous variable.

  Each element of the sample space would map to a value of X

Random Variable

- Each random event A has a **probability P(A)** associated with it.
- It defines the fraction of the sample space in which A is true.

$$\mathrm{P}(A) = \frac{\text{Number of events in which A is true}}{\text{Total number of events in S}}$$

e.g. in case of toss of a dice, probability of even number:

$$P(A) = \frac{3}{6} = 0.5$$

Useful Theorem

- $0 <= P(A) <= 1$, $P(True) = 1$, $P(False) = 0$,
  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

➔ $P(A) = P(A \wedge B) + P(A \wedge \sim B)$

$A = [A \text{ and } (B \text{ or } \sim B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)]$

$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P((A \text{ and } B) \text{ and } (A \text{ and } \sim B))$

$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P(A \text{ and } B \text{ and } A \text{ and } \sim B)$

Definition of Conditional Probability

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

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\,P(B)$$

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

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

Applying Bayes Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

A = you have the flu, B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

what does all this have to do with function approximation?

After all, that's what we are looking for

Function approximation by probability

- Our aim is to approximate the class separating function
  F: X -> Y
  X is the set of attributes and Y is the class.
- In case of probabilistic reasoning, we approximate it with the conditional probability
  P(Y | X)
  we find the probability of each class, given the data i.e.
  P(Y = 1| X) and P(Y = 0 | X).
- How do we find the probability of class given the data?
- Need a joint distribution

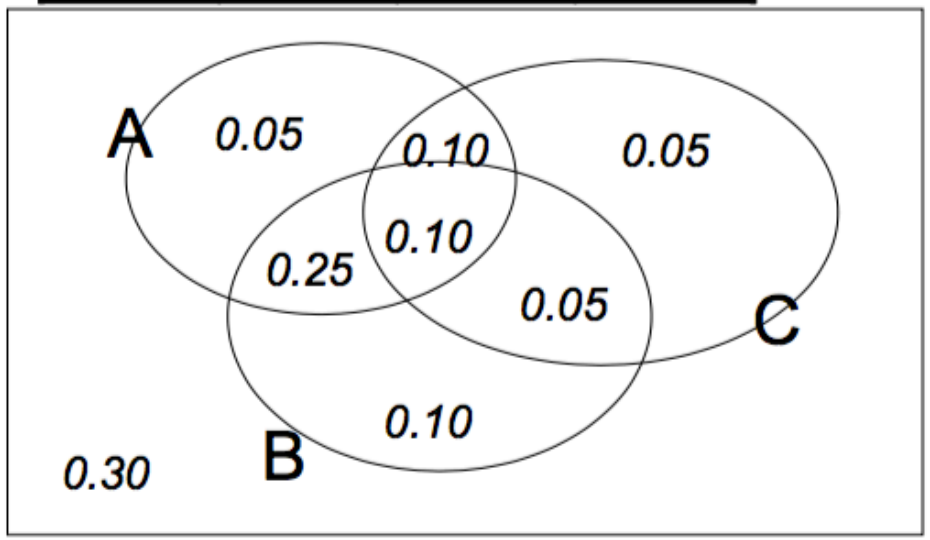
The Joint Distribution

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

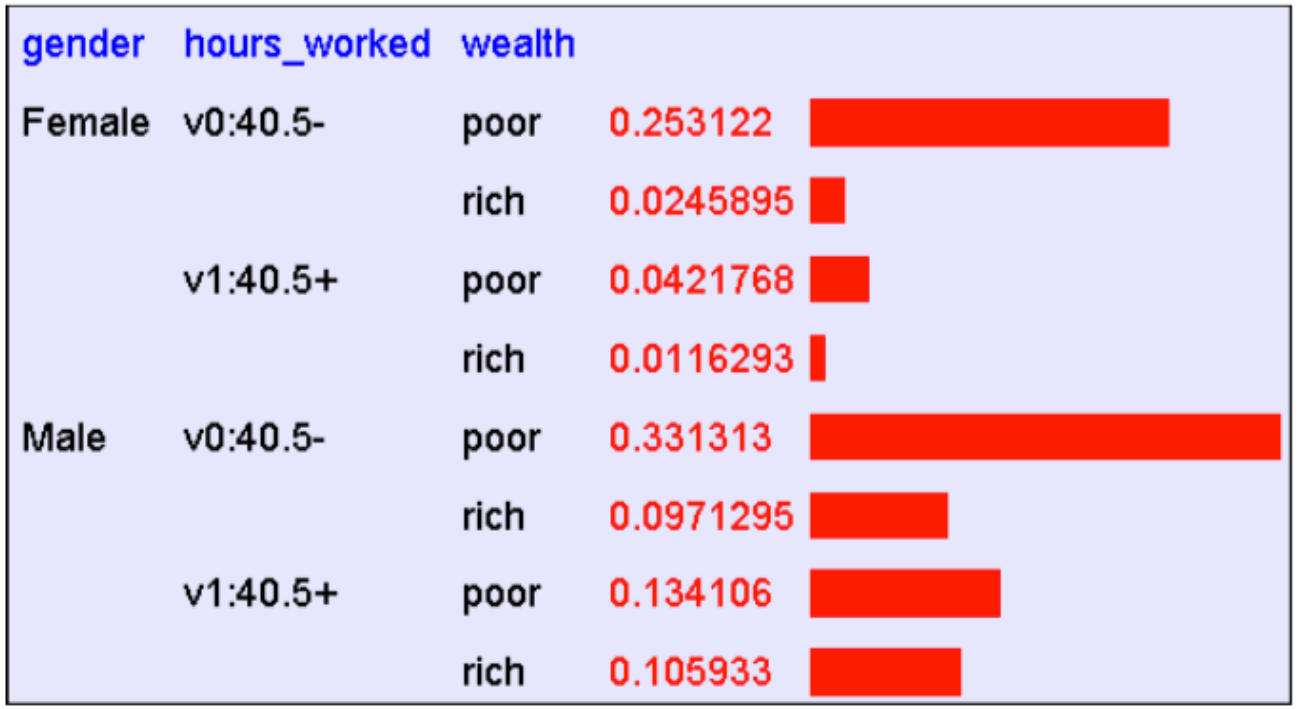
Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933



One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

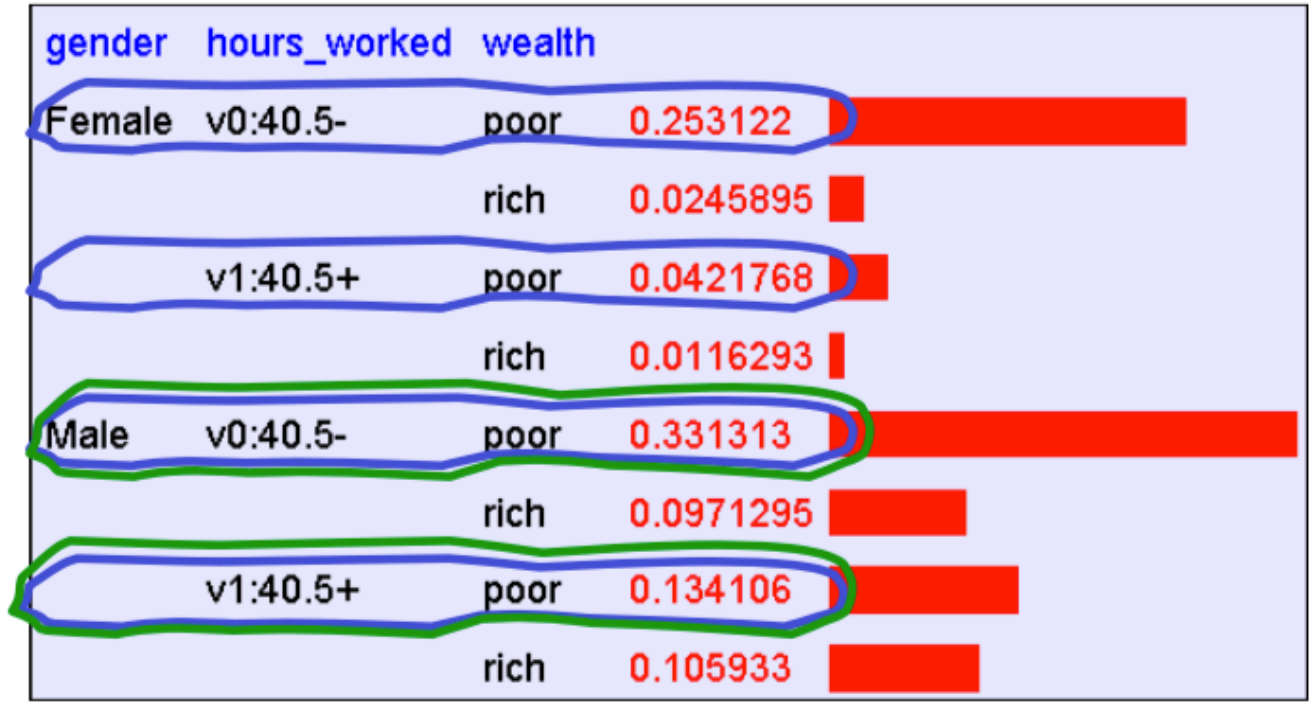
Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint



$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

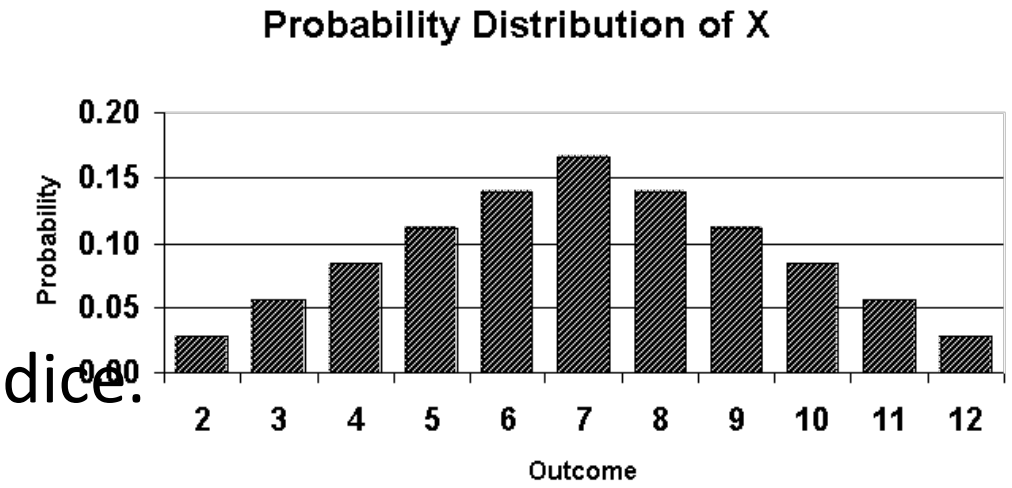
Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

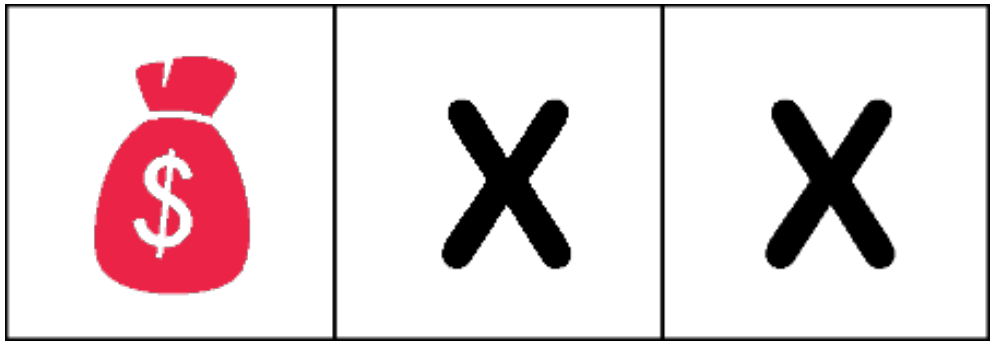
e.g., P(W=rich | G = female, H = 40.5-) =

Let's get back to random variables

- Each random variable X has a domain ***D(x)*** associated with it. It is the set of outcome values possible.
  e.g. for a dice D(x) = {1, 2, 3, 4, 5, 6}
  for a Boolean outcome D(x) = {0, 1}.
- For a pair of dice, D(x) = {11, …., 66}.
  Suppose X is the sum of the outcome of the dice.
  X can range from 2 to 12.
  Each value of X has a different probability.
- A plot of the random variable and probability values is called the probability distribution plot and the function is called probability distribution function (pdf) or probability mass function (pmf).



Example:



- Imagine there are 3 doors – one of them has a treasure, the other 2 nothing.
- Each one is equally likely to be selected. Once you open a door, you don't open it again.
- Define random variable X as the number of doors needed to open before finding the treasure.
  X can have values {1, 2}. Find the probability distribution of X.

  P(X = 1) = 1/3
  P(X = 2) = (1/3) * (1/2)
  -- Assume a person is smart enough to figure it out after two attempts -- ☺

Example 2:

Consider a group of five potential blood donors—$a, b, c, d,$ and e—of whom only a and b have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv $Y =$ the number of typings necessary to identify an O+ individual. Then the pmf of Y is

$$p(1) = P(Y = 1) = P(a \text{ or } b \text{ typed first}) = \frac{2}{5} = .4$$

$$p(2) = P(Y = 2) = P(c, d, \text{or } e \text{ first, and then } a \text{ or } b)$$

$$= P(c, d, \text{or } e \text{ first}) \cdot P(a \text{ or } b \text{ next} \mid c, d, \text{or } e \text{ first}) = \frac{3}{5} \cdot \frac{2}{4} = .3$$

$$p(3) = P(Y = 3) = P(c, d, \text{or } e \text{ first and second, and then } a \text{ or } b)$$

$$= \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{2}{3}\right) = .2$$

$$p(4) = P(Y = 4) = P(c, d, \text{and } e \text{ all done first}) = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right) = .1$$

$$p(y) = 0 \quad \text{if } y \neq 1, 2, 3, 4$$

Binary Variables

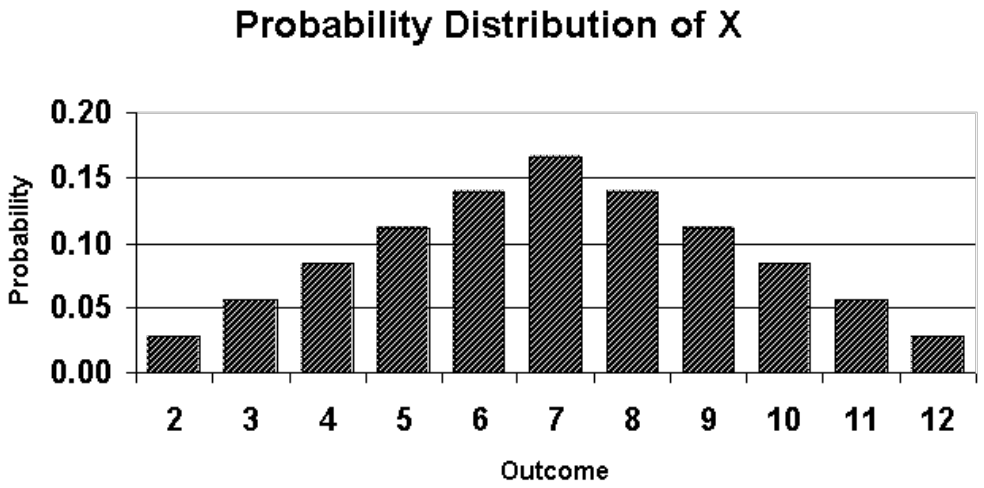
- Let's focus on the case where the random variable X can only take two values {0, 1}.
- X is said a Boolean random variable.
- For every outcome in the sample space, you associate 0 (failure) or 1 (success).
- Example:
  - Coin toss – Heads = 1, Tails = 0
  - Lottery - Winning Number = 1, Rest of the numbers = 0
  - …

Expected Value and Variance

- **Expected value** of a discrete random variable under P:

$$E_P(X) = \sum_{x \in D(x)} x\, P(x)$$

take each value and multiply by its probability



- **Variance** of the random variable under P:

$$var_P(X) = \sum_{x \in D} \left(x - E_P(x)\right)^2 P(x)$$

Shortcut formula:

$$var_P(X) = E_P(X^2) - [E_P(X)]^2$$

Expected Value and Variance

- **Expected value** of a continuous random variable under a continuous probability function f:

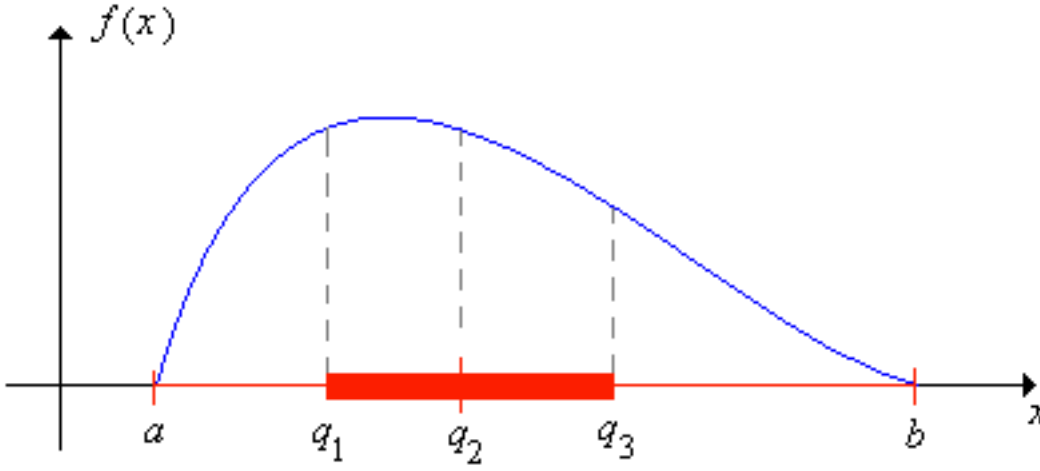
$$E_f(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

- **Variance** of the random variable under P:

$$var_f(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Shortcut formula:

$$var_f(X) = E_f(X^2) - \left[E_f(X)\right]^2$$



Continuous Probability Density Function

Example:

You draw one card from a standard deck of playing cards. If you pick a heart, you will win$10. If you pick a face card (K, Q, J), which is not a heart, you win $8. If you pick any other card, you lose $6. Do you want to play? Explain.

Solution

Let X be the random variable that takes on the values 10, 8 and –6, the values of the winnings. First, we calculate the following probabilities:

$$P(X=10)=\frac{13}{52},\ P(X=8)=\frac{9}{52},\text{ and }\ P(X=-6)=\frac{30}{52}.$$

The expected value of the game is

$$\begin{aligned}E(X) &= P(X=10)*10+P(X=8)*8-P(X=-6)*6\\ &=\frac{13}{52}*10+\frac{9}{52}*8-\frac{30}{52}*6\\ &=\frac{130+72-180}{52}\\ &=\frac{22}{52}\end{aligned}$$

Since the expected value of the game is approximately \$.42, it is to the player's advantage to play the game.

Binary Variables

- Consider coin flipping which has 2 outcomes (heads = 1 and tails = 0)
  If you know probability of heads, you know the distribution.
  Let's say

$$p(x = 1|\,\mu) = \mu \qquad p(x = 0|\,\mu) = 1 - \mu$$

- Probability of heads in **one** coin flip makes the Bernoulli Distribution (Bern):

$$Bern(x\,|\,\mu) = \mu^x(1-\mu)^{1-x}$$

Not convinced?
Plug in x = 0 and x = 1 in this equation to see

Easy to show that

$$E_{Bern}(x) = \mu$$

$$var_{Bern}(x) = \mu(1-\mu)$$

What do you notice?
If I know μ, I know everything about the distribution.

Binomial Distribution

- Now imagine you throw that same coin **N times** and you want to find the probability of **m heads** where m can be from 0 to N.

$$p(m\ heads|\ N, \mu)$$

- This is called the Binomial distribution.

$$Bin(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

It is easy to show that:

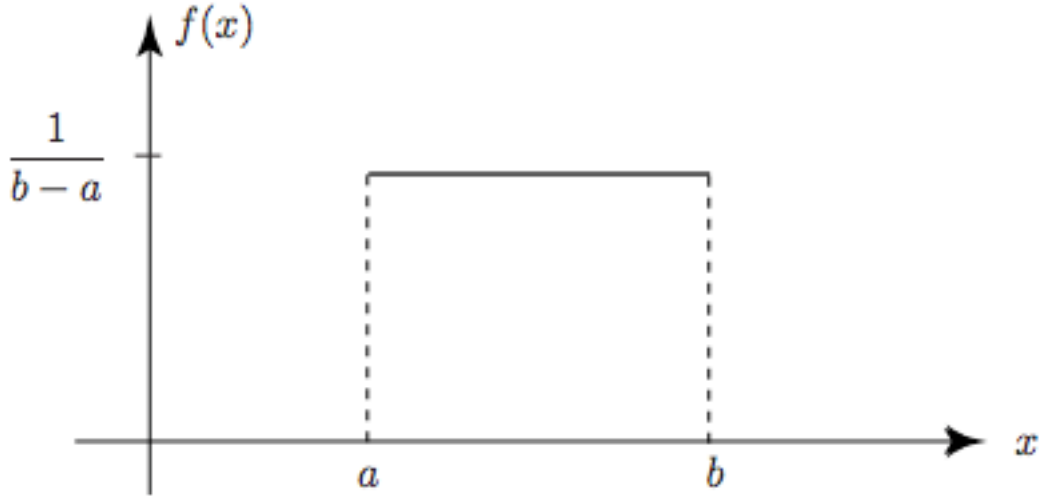
$$E[m] = \sum_{m=1}^{N} m\ Bin(m \mid N, \mu) = N\ \mu$$

$$var[m] = \sum_{m=1}^{N} (m - E[m])^2\ Bin(m \mid N, \mu) = N\ \mu(1-\mu)$$

What do you notice?
If I know N and μ, I know everything about the distribution.

Uniform Probability Distribution

- Continuous Distribution:
  Density plot shown on right



The function $f(x)$ is defined by:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \int_a^b x\frac{1}{b-a}\mathrm{d}x = \frac{1}{2(b-a)}\left[x^2\right]_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$

What do you observe?
Just need a and b to know everything about the distribution.

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \int_a^b x^2.\frac{1}{b-a}\mathrm{d}x - \left(\frac{b+a}{2}\right)^2 = \frac{1}{3(b-a)}\left[x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

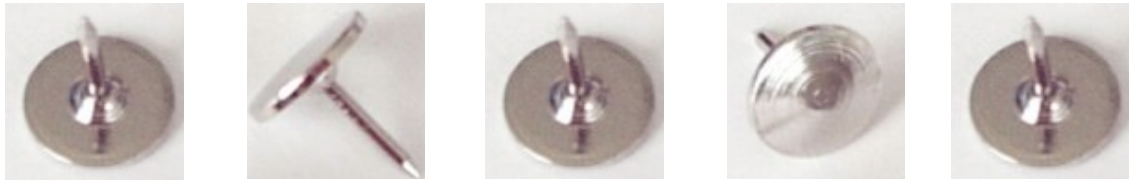
Parameter Estimation

- The big question is – how do we estimate parameters?
- If someone is tossing a coin or rolling a dice, how do we find their parameters (the probability of heads or probability of getting each numbers)?
- This is related to machine learning, where we estimate a function from data.
- Approach is the same – generate some data from the distribution and use the data to estimate parameters.

Your first consulting job

Billionaire in Dallas asks:

- He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
- You say: Please flip it a few times:



- You say: The probability is:
  - P(H) = 3/5
- **He says: Why???**
- You say: Because…

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are *i.i.d.*:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- **Learning:** finding θ is an optimization problem
  - What's the objective function?

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- **MLE:** Choose θ to maximize probability of *D*

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your first parameter learning algorithm

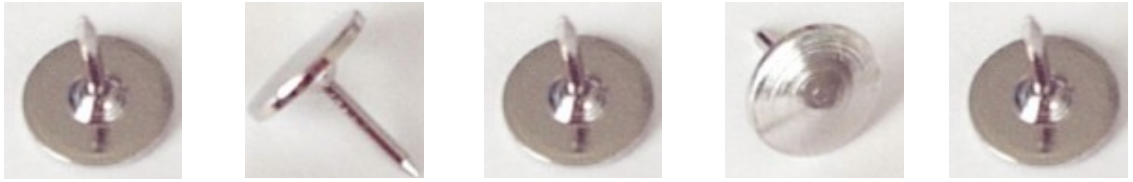
$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

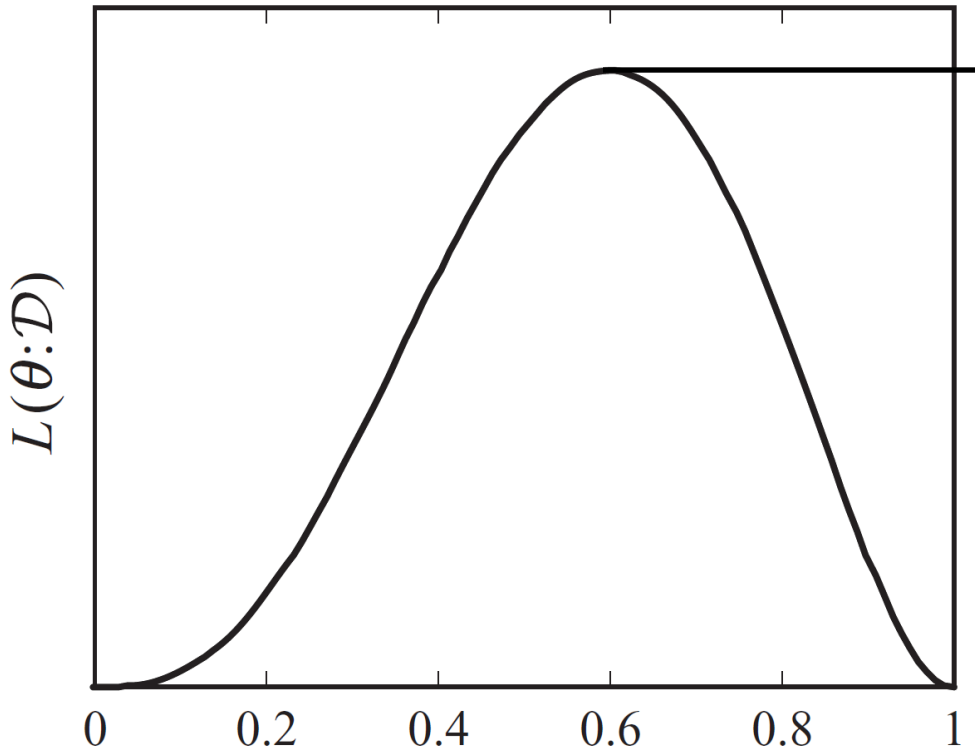
- Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right] \\ &= \frac{d}{d\theta}\left[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right] \\ &= \alpha_H \frac{d}{d\theta}\ln\theta + \alpha_T \frac{d}{d\theta}\ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0\end{aligned}$$

$$\boxed{\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Data





$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Parameter Estimation: Summary

- ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\},\ m$ heads $(1),\ N - m$ tails (0)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

Parameter Estimation

- Learning scenario:
  - You have data that is coming from some **fixed distribution characterized by θ.**
  e.g. Bernoulli, Binomial, Gaussian, Poisson.
  - You get some data D and its class.
  - You want to evaluate *which value of θ most likely generated this data*.
  Find most likely parameter that generated this data.

  => known as Maximum Likelihood Estimate (MLE)

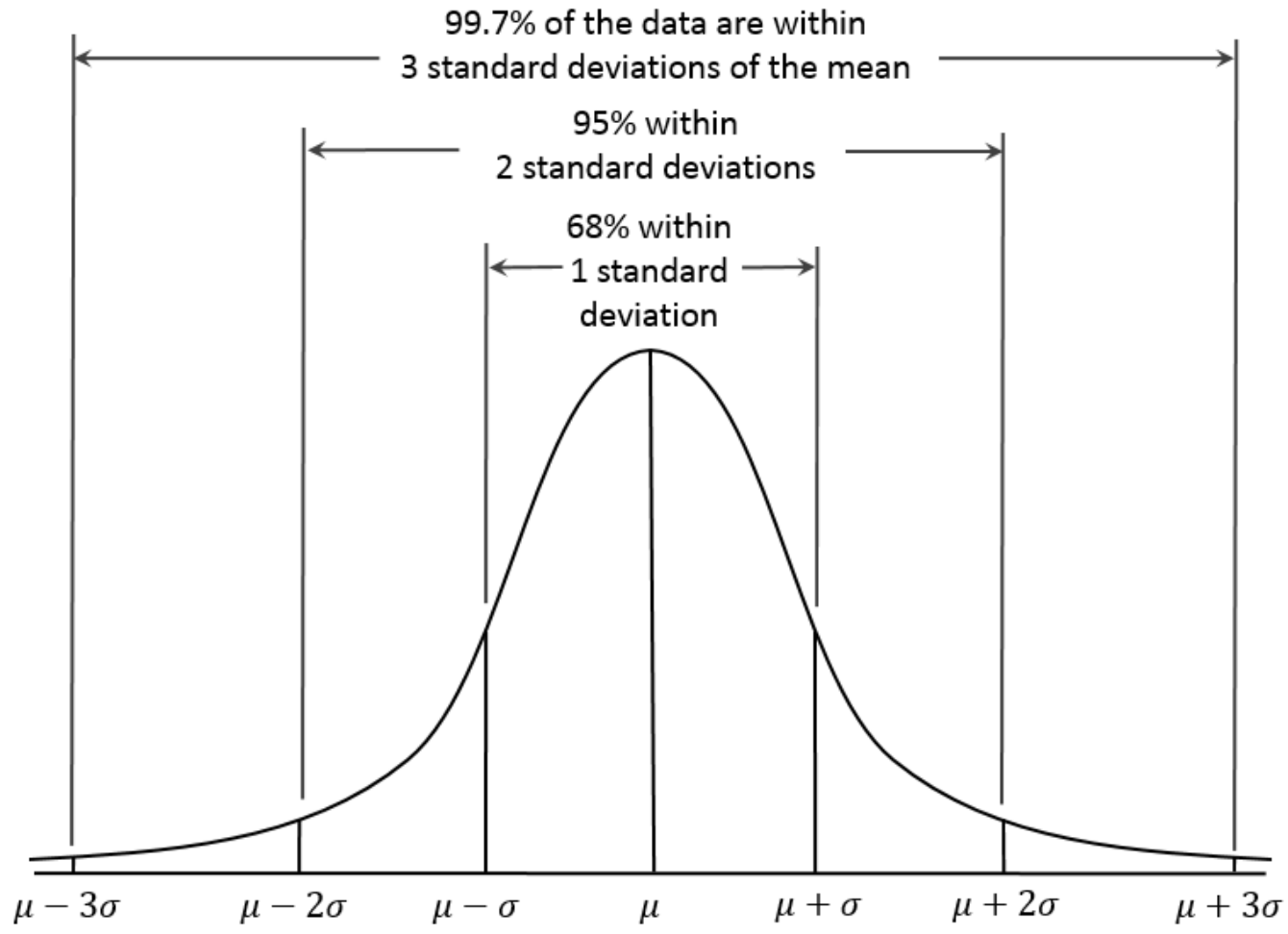
Why we should care

- Maximum Likelihood Estimation is a very very very very fundamental part of data analysis.
- "MLE for Gaussians" is training wheels for our future techniques
- Learning Gaussians is more useful than you might guess…

Gaussian Distribution

- PDF is described by:

$$f(x) = \frac{}{\sqrt{2}}$$



- Central Limit Theorem:
  Sample repeatedly n items (n > 30) from any distribution and take their means. The plot of the means will be a Gaussian (normal) distribution.

Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R \sim$ (i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ

(you do know σ^2)

Simplifying assumption

MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

Something we will cover in the future:

MAP: Which μ maximizes $p(\mu|x_1, x_2, \ldots x_R, \sigma^2)$?

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ (you do know σ^2)
- MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus
3. Set $\partial LL/\partial\theta=0$ for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you've found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

*This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new.

Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \qquad \text{(by i.i.d)}$$

$$= \qquad \text{(monotonicity of log)}$$

$$= \qquad \text{(plug in formula for Gaussian)}$$

$$= \qquad \text{(after simplification)}$$

Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i \mid \mu, \sigma^2)$$ (by i.i.d)

$$= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i \mid \mu, \sigma^2)$$ (monotonicity of log)

$$= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{R} -\frac{(x_i - \mu)^2}{2\sigma^2}$$ (plug in formula for Gaussian)

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$ (after simplification)

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \quad \text{s.t.} \quad 0 = \frac{\partial \mathrm{LL}}{\partial \mu} =$$

= (what?)

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \quad \text{s.t.} \quad 0 = \frac{\partial \mathrm{LL}}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$- \sum_{i=1}^{R} 2(x_i - \mu)$$

$$\text{Thus} \quad \mu = \frac{1}{R} \sum_{i=1}^{R} x_i \qquad \mu^{mle} = \frac{1}{R} \sum_{i=1}^{R} x_i$$

A General MLE strategy

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum

Now, let us assume we don't know both the parameters of the Gaussian.
Can we estimate that?

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu) \quad \Rightarrow \quad 0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

$$\frac{\partial LL}{\partial \sigma^2} = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2 \qquad 0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu) \Rightarrow \mu = \frac{1}{R}\sum_{i=1}^{R}x_i$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2 \Rightarrow \text{what?}$$

$$\mu^{mle} = \frac{1}{R}\sum_{i=1}^{R}x_i$$

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Unbiased Estimators

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

$$E[\mu^{mle}] = E\left[\frac{1}{R}\sum_{i=1}^{R} x_i\right] = \mu$$

μ^{mle} is unbiased

Biased Estimators

- An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R}x_i - \frac{1}{R}\sum_{j=1}^{R}x_j\right)^2\right] \neq \sigma^2$$

σ^2_{mle} is biased