What was the pace of the last lecture for you?

Video credit: Clay Bavor

Happy Wednesday!

- Focus videos on Linear algebra and probability theory out Thusday by 10am
  - **LA:** SVD, Eigen-decomposition, matrix calculus, norms
  - **Prob:** ?

- Open office hours on Thursday, 7pm to 8pm
  - https://primetime.bluejeans.com/a2m/live-event/rjsfkuku

- **Project seminar 1**, available Thursday, Aug 27th at 5pm
  - Seminar series information available on the class website

- Quiz 1, Friday, Aug 28th 6am until Aug 29th 6am
  - Linear algebra and probability

CS4641B Machine Learning

Lecture 04: Probability theory

Rodrigo Borela ▸ rborelav@gatech.edu

These slides are based on slides from Le Song , Sam Roweis, Chao Zhang and Mahdi Roozbahani

Outline

- Probability distributions
- Joint and conditional probability distributions
- Bayes' rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Complementary reading: Bishop PRML – Chapter 1, Sections 1.2 through 1.2.4 and Appendix B

Outline

- Probability distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Why probability theory?

- It is often intractable to obtain data about an entire population of items (e.g. measuring every person's height at a given age)
- We use **statistical inference** to estimate information about the distribution in the population, such as the mean (average) and variance (how much spread there is in the data) from samples
- With samples of finite size or noise in measurements comes **uncertainty** about the true mean/variance about the population
- **Probability theory allows us to quantify different forms of uncertainty**

Probability world views

- **Frequentist (classical) definition:** long-term frequencies of repeatable random events (e.g. result of flipping a coin).

- **Bayesian definition**: more general concept, in which the probabilities represent the uncertainty in any event or hypothesis (not just one that can be repeated a number of times), for example the probability of becoming an opera singer by the end of the semester.

<u>In this class we will work with the frequentist (classical) approach</u>

Three key ingredients in probability theory

- A sample space is a collection of all possible outcomes
- Random variables X represent outcomes in the sample space
- Probability of a random variable to happen $p(x) = p(X = x),\ p(x) \geq 0$

Probability

- A **sample space S** is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite)
  - Example, S may be the set of all possible outcomes of a dice roll: $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$
  - Example, S may be the set of all possible nucleotides of a DNA site: $(A \quad C \quad G \quad T)$
  - Example, S may be the set of all possible time-space positions of an aircraft on a radar screen.
- An **event A** is any subset of S
  - Seeing “1” or “6” in a dice roll; observing a “G” at a DNA site

Types of variables

- Discrete variable
  - Example: Coin flip (integer)
  - Discrete probability distribution (e.g. Bernoulli)
  - Probability mass function
  - Probability value

$$\sum_{x \epsilon A} p(x) = 1$$

- Continuous variable
  - Example: Temperature (real number)
  - Continuous probability distribution (e.g. Gaussian)
  - Probability density function
  - Density or likelihood value

$$\int_x p(x)dx = 1$$

What distribution to model my data with?

- Is my variable discrete or continuous?
- How can I define the stochastic process generating the data?
- How much information do I have about the data?
- Can I visualize my data? If so, can I represent it as a parametric distribution or should I opt for a non-parametric distribution?
- What does the literature on this type of data suggests?

Discrete probability functions

- Bernoulli distribution (single trial is conducted)

$$\begin{cases} 1-\theta & for\ x = 0 \\ \theta & for\ x = 1 \end{cases}$$

- Binomial distribution (k number of successes, $n - k$ number of failures)

$$P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$\binom{n}{k}$ the total number of ways of selecting k distinct combinations of n trials irrespective of order

Continuous probability functions

- Uniform density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Exponential density function

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \text{for } x \geq 0$$

- Gaussian density function

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

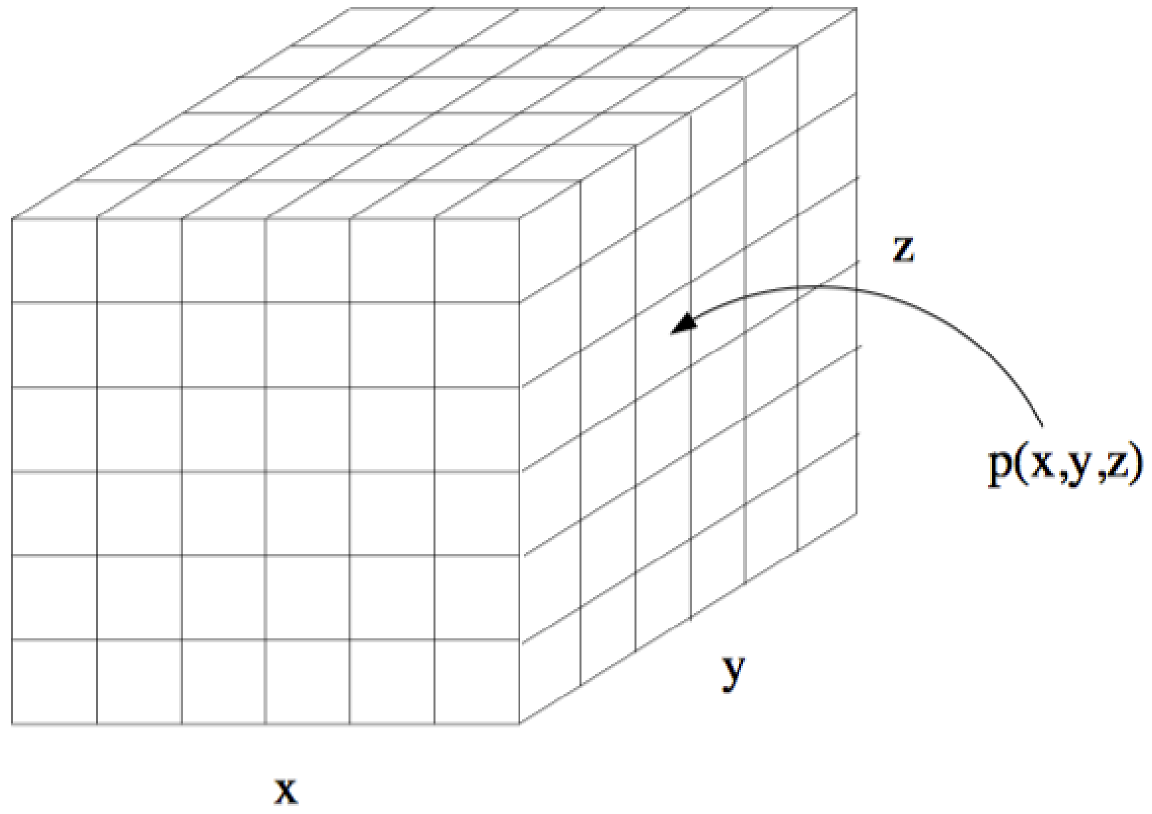
Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Joint distribution

- **Key concept**: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which values the others are taking
- We call this a join ensemble and write:

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$



Marginal distribution

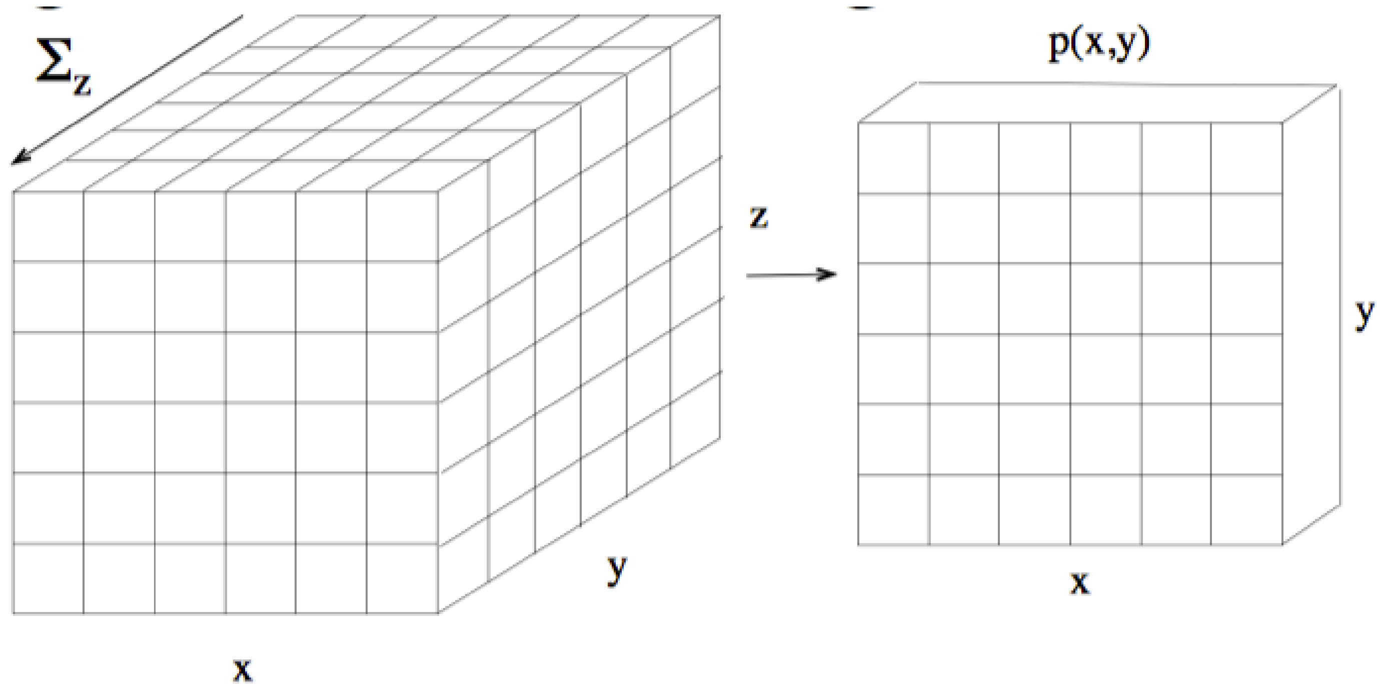
- We can “sum out” part of a joint distribution to get the marginal distribution of a subset of variables:

$$p(x) = \sum_y p(x, y) \text{ (discrete variables)}$$

or

$$p(x) = \int p(x, y) dy \text{ (continuous variables)}$$

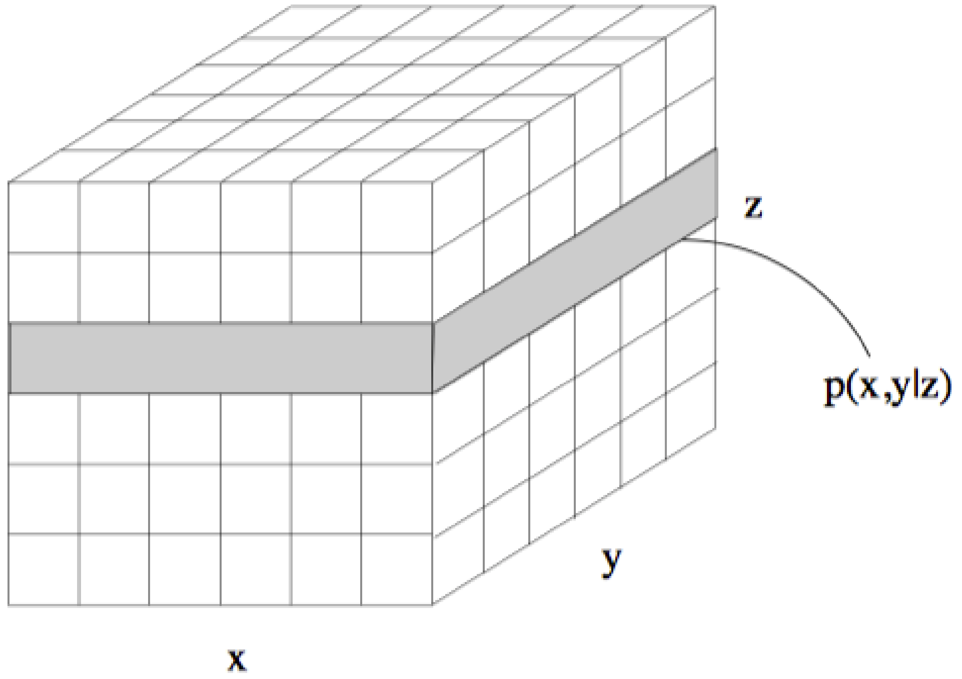
- This is like adding slices of the table together



Conditional distribution

- If we know that some event has occurred, it changes our belief about the probability of other events
- This is like taking a "slice" through the joint table

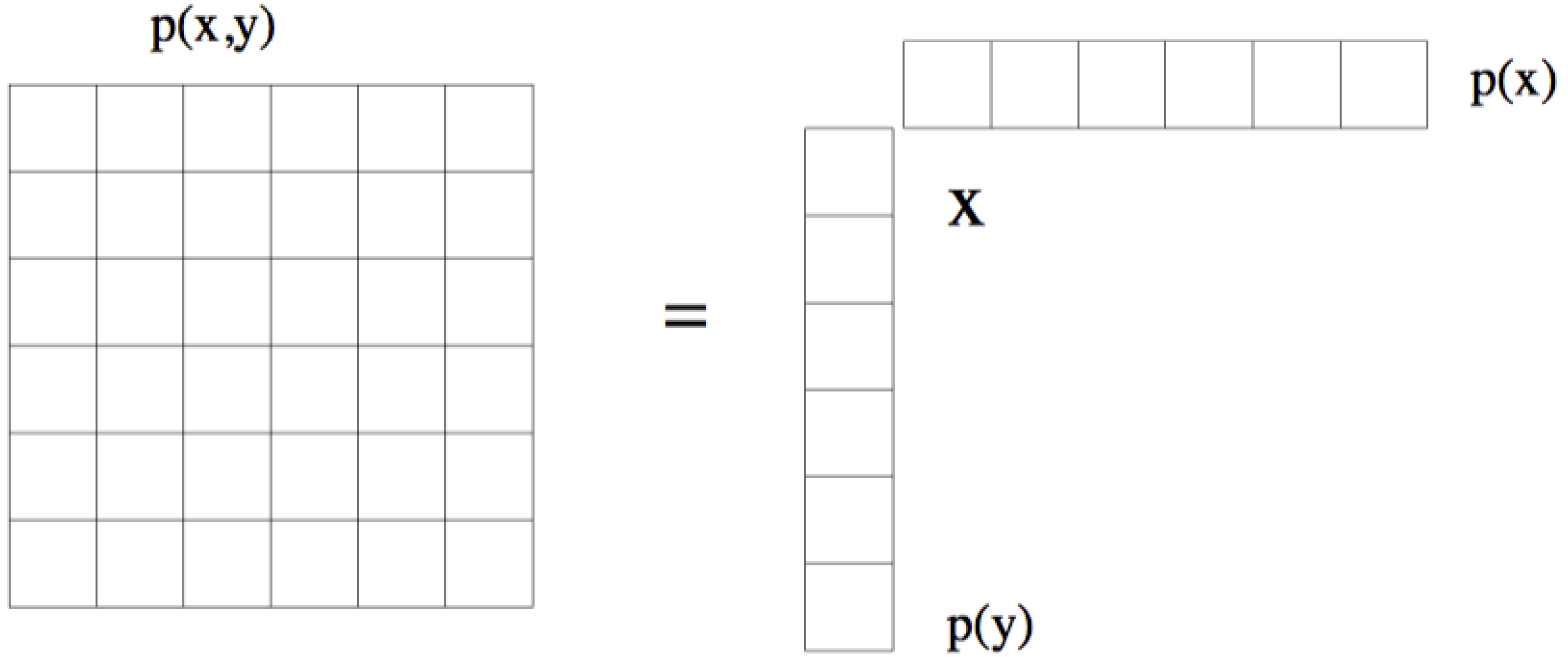
$$p(x|y) = \frac{p(x,y)}{p(y)}$$



Independence and conditional independence

- Two variables are independent iff their joint factors are

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



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

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

Example: conditional independence

- $p(virus|coffee) = p(virus),$ iff virus is independent of drinking coffee
- $p(flu|virus, coffee) = p(flu|virus),$ iff flu is independent of drinking coffee, given the virus
- $p(headache|flu, virus, coffee) = p(headache|flu, coffee),$ iff headache is independent of virus, given drinking coffee and the flu

- We can write the joint distribution:

$$p(headache, flu, virus, coffee) = p(h|f, v, c)p(f|v, c)p(v|c)p(c)$$

- Assuming the above independence:

$$p(headache, flu, virus, coffee) = p(h|f, c)p(f|v)p(v)p(c)$$

Example: Joint, conditional and marginal

- X and Y are random variables
- $N \ = \ total\,number\ of\ trials$
- $n_{ij} = number\ of\ occurrence$

X = Throw a die

Y = Flip a coin

Example: Joint, conditional and marginal

X

Y	$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	c_j
$y_{j=1} = head$	$n_{ij} = 3$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 5$	$n_{ij} = 1$	$n_{ij} = 5$	20
$y_{j=2} = tail$	$n_{ij} = 2$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 1$	15
c_i	5	6	6	7	5	6	$N = 35$

Definitions (discrete variables)

- Marginal probability

$$p(X = x_i) = \frac{c_i}{N}$$

- Join probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Definitions (discrete variables)

- Sum rule

$$p(X = x_i) = \sum_{j=1}^{L} p\left(X = x_i, Y = y_j\right) \rightarrow p(X) = \sum_{Y} P(X, Y)$$

- Product rule

$$p\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p\left(Y = y_j \middle| X = x_i\right)p(X = x_i)$$

$$p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y)$$

Example: Joint, conditional and marginal

X

Y		$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	c_j
	$y_{j=2} = tail$	$n_{ij} = 3$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 5$	$n_{ij} = 1$	$n_{ij} = 5$	20
	$y_{j=1} = head$	$n_{ij} = 2$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 1$	15
	c_i	5	6	6	7	5	6	$N = 35$

Joint probability: $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$, $p(X = 1, Y = tail) = \frac{3}{35}$, $p(X = 5, Y = head) = \frac{4}{35}$

Conditional probability: $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$, $p(Y = head | X = 3) = \frac{4}{6}$, $p(X = 6 | Y = tail) = \frac{5}{20}$

Marginal probability: $p(X = x_i) = \frac{c_i}{N}$, $p(X = 6) = \frac{6}{35}$, $p(Y = head) = \frac{15}{35}$

Example: Joint, conditional and marginal

X

Y		$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	c_j
	$y_{j=2} = tail$	$n_{ij} = 3$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 5$	$n_{ij} = 1$	$n_{ij} = 5$	20
	$y_{j=1} = head$	$n_{ij} = 2$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 1$	15
	c_i	5	6	6	7	5	6	$N = 35$

Sum rule:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

$$p(X = 6) = p(X = 6, Y = tail) + p(X = 6, Y = head) = \frac{5}{35} + \frac{1}{35} = \frac{6}{35}$$

Product rule: $p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i) p(X = x_i)$

$$p(X = 1, Y = tail) = p(Y = tail | X = 1) p(X = 1) = \frac{3}{5} \cdot \frac{5}{35} = \frac{3}{35}$$

Bayes' rule (theorem)

- $p(X|Y) =$ Fraction of the worlds in which X is true given that Y is also true

- For example:
  - $H\ =$ "having a headache"
  - $F\ =$ "Coming down with the flu"
  - $P(headache|flu)\ =$ fraction of flu-inflicted worlds in which you have a headache. How to calculate?

- Definition

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)}$$

Bayes' rule

$$p(headache|flu) = \frac{p(headache, flu)}{p(flu)} = \frac{p(flu|headache)p(headache)}{p(flu)}$$

- Other cases:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X|Y)p(Y)+p(X|\neg Y)p(\neg Y)} \text{ (binary variables)}$$

$$p(Y|X, Z) = \frac{p(X|Y, Z)p(Y, Z)}{p(X, Z)} = \frac{p(X|Y, Z)p(Y, Z)}{p(X|Y, Z)p(Y, Z) + p(X|\neg Y, Z)p(\neg Y, Z)}$$

$$p(Y = y_i|X) = \frac{p(X|Y)p(Y)}{\sum_{y_i} p(X|Y=y_i)p(Y=y_i)} \text{ (multiple discrete states)}$$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Mean and variance

- Expectation: the mean value, center of mass, first moment:

$$E_x[g(x)] = \int_{-\infty}^{\infty} g(x) p_x(x) dx = \mu$$

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$

Mean and variance

- Mean (first moment)

$$E_x[X] = \int_{-\infty}^{\infty} x p_x(x) dx$$

- Properties
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$

- Variance (second central moment)

$$var(X) = E_x[(X - E_x[X])^2] = E_x[X^2] - E_x[X]^2$$

- Properties
  - $var(\alpha X) = \alpha^2 Var(X)$
  - $var(\alpha + X) = Var(X)$

For joint distributions

- Expectation

$$E[X + Y] = E[X] + E[Y]$$

- Covariance

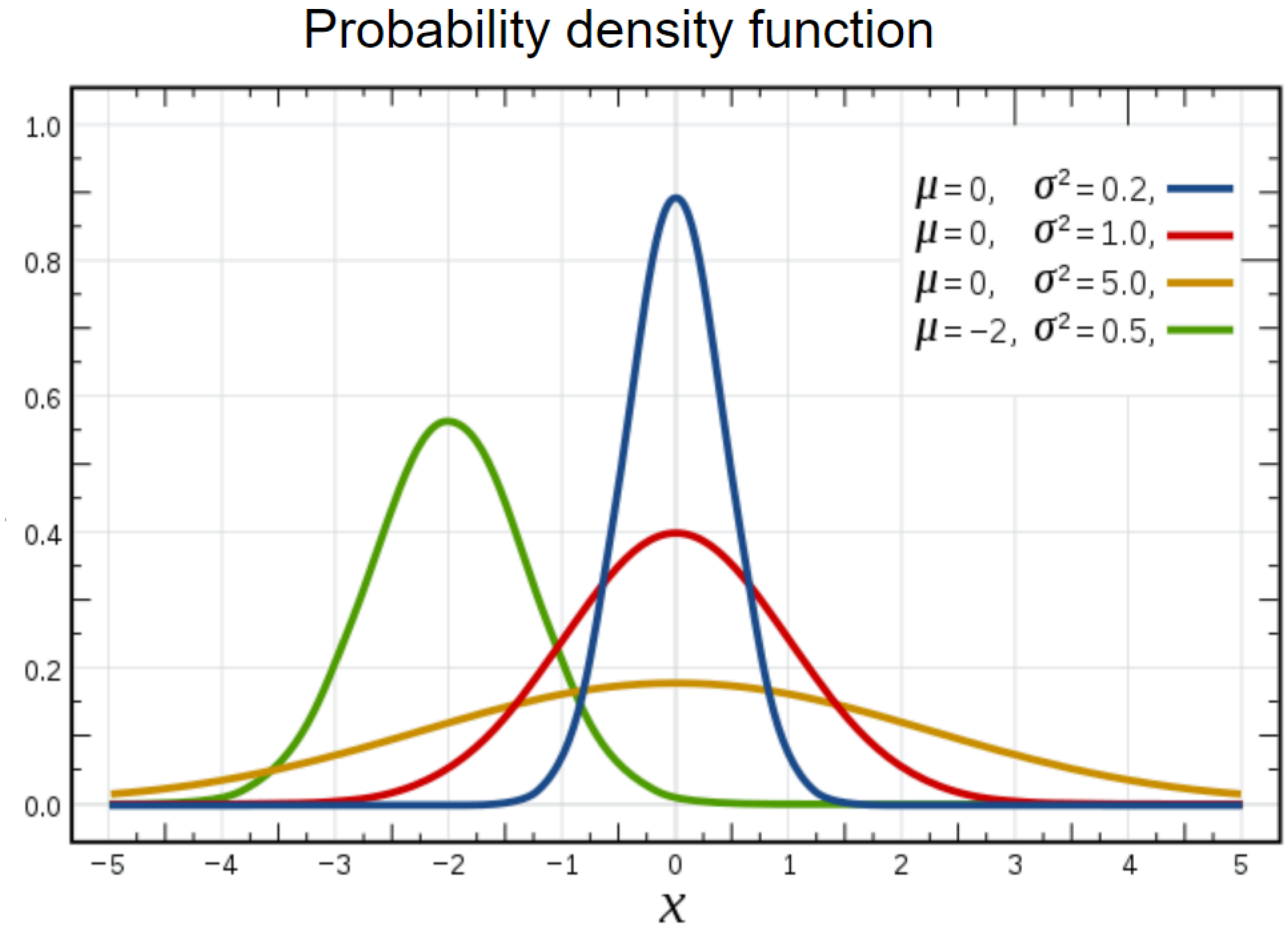
$$cov(X, Y) = E\left[(X - E_X[X])\left(Y - E_y(Y)\right)\right] = E[XY] - E[X]E[Y]$$

$$var(X + Y) = Var(X) + 2cov(X, Y) + Var(Y)$$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

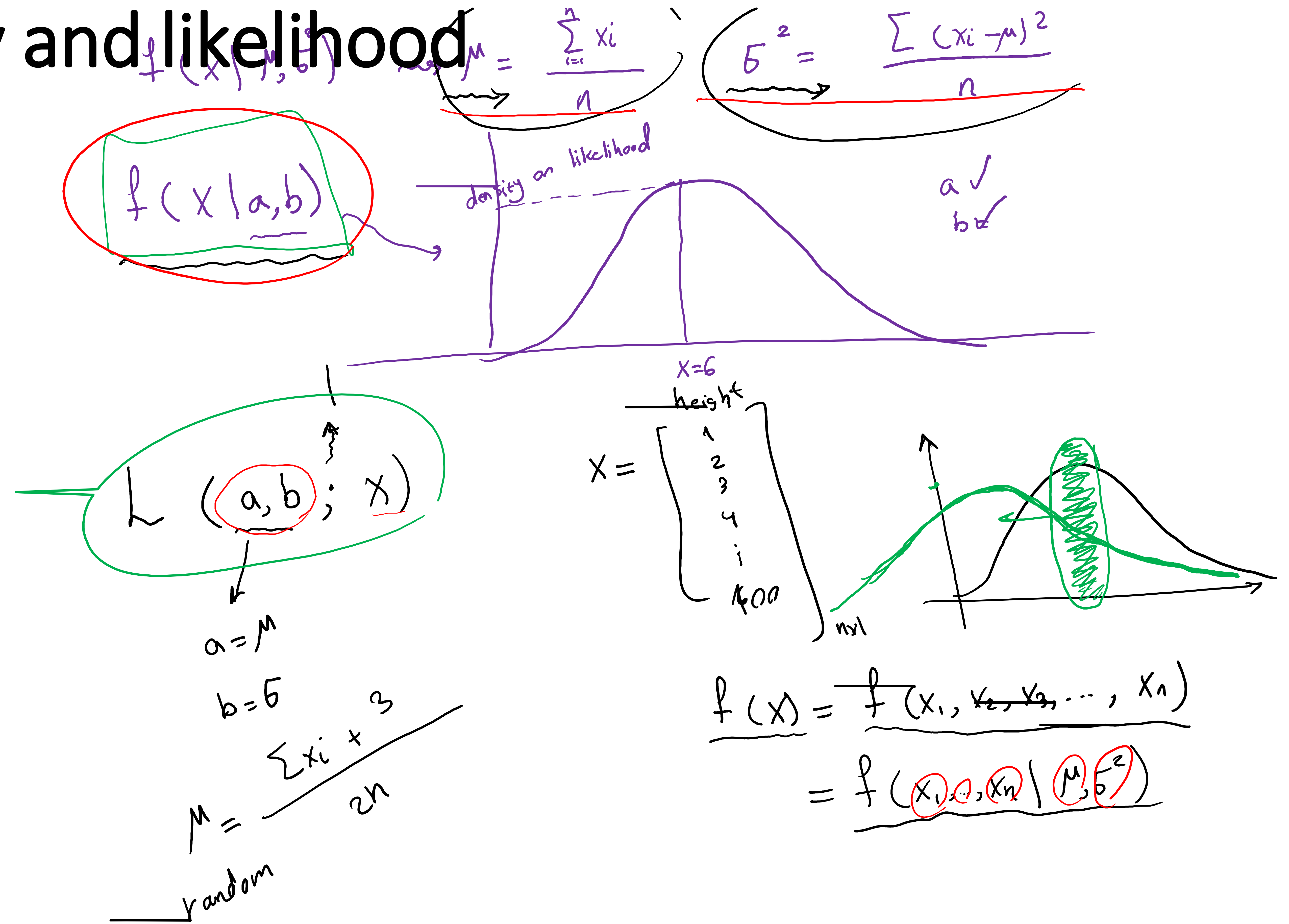
Gaussian distribution



Probability versus likelihood

Probability and likelihood

- …



Multivariate Gaussian Distribution

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

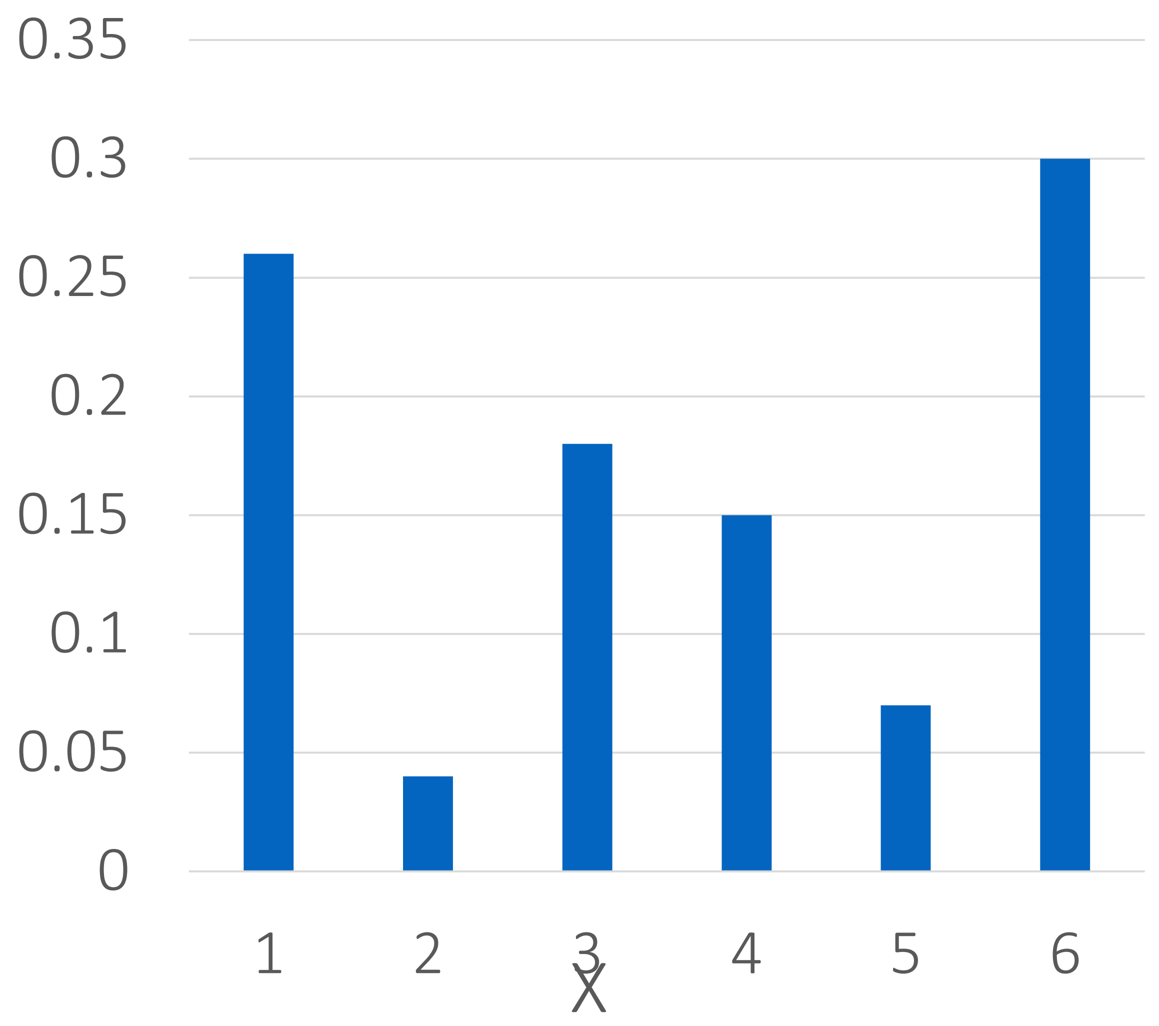
- Moment parametrization:
  - $\boldsymbol{\mu} = E(\mathbf{X})$
  - $\boldsymbol{\Sigma} = cov(\mathbf{X}) = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}]$
- Tons of applications (Mixture of gaussians, Bayesian linear regression, PPCA, Kalman filter)

Properties of Gaussian distribution

- The linear transform of a Gaussian r.v. is a Gaussian. Remember that no matter how x is distributed
  - $E(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}E(\mathbf{x}) + \mathbf{b}$
  - $cov(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}cov(\mathbf{x})\mathbf{A}^T$
- This means that for Gaussian distributed quantities
  - $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow \mathbf{Ax}+\mathbf{b} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right)$
- The sum of two independent Gaussian r.v. is a Gaussian
  - $Y = X_1 + X_2,\ X_1 \perp X_2 \rightarrow \mu_y = \mu_1 + \mu_2,\ \Sigma_y = \Sigma_1 + \Sigma_2$
- The multiplication of two Gaussian functions is another Gaussian function (no longer normalized)
  - $\mathcal{N}(a, A)\mathcal{N}(b, B) \propto \mathcal{N}(c, C)$
  - Where $C = (A^{-1} + B^{-1})^{-1},\ c = CA^{-1}a + CB^{-1}b$

Central limit theorem

Probability mass function of a **biased** dice



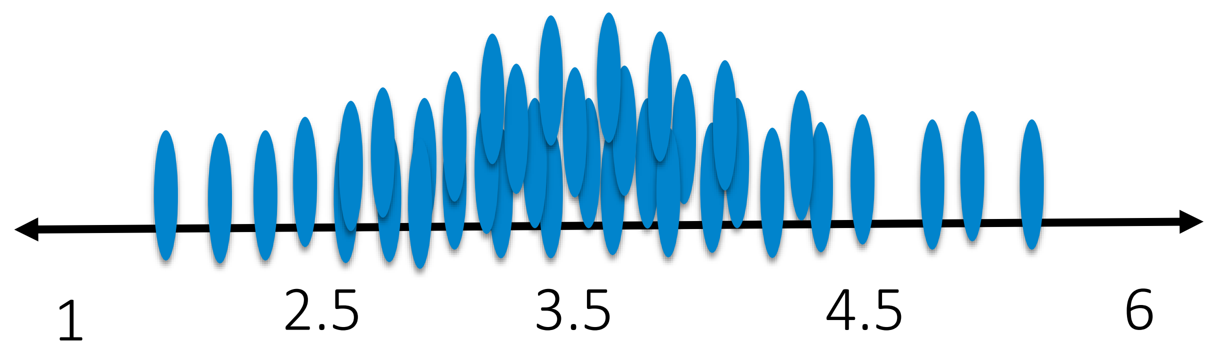
Let's say, I am going to get a sample from this pmf having a size of $\boldsymbol{n = 4}$

$$S_1 = \{1,1,1,6\} \Rightarrow E(S_1) = 2.25$$

$$S_2 = \{1,1,3,6\} \Rightarrow E(S_2) = 2.75$$

$$\vdots$$

$$S_m = \{1,4,6,6\} \Rightarrow E(S_m) = 4.25$$



According to CLT, it will follow a bell curve distribution (normal distribution)

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Maximum likelihood estimation

- Probability: inferring probabilistic quantities for data given fixed models (e.g. prob. of events, marginals, conditionals, etc.)
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression)

- Main assumption in MLE:

  <u>Independent and identically distributed random variables</u>

Maximum likelihood estimation

- For Bernoulli (i.e. flip a coin):

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} \qquad x_i \in \{0,1\}\ or\ \{head, tail\}$$

- **Objective function** (what we are trying to maximize):

$$L(\mathcal{D}|\,\theta) = p(X = x_1, X = x_2, X = x_3, \dots, X = x_n)$$

applying the i.i.d. assumption

$$= p(X = x_1)p(X = x_2) \dots p(X = x_n)$$

We can then rewrite:

$$L(\mathcal{D}|\,\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

Maximum likelihood estimation

$$L(\mathcal{D}|\,\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \times \theta^{x_2}(1-\theta)^{1-x_2} \ldots \times \theta^{x_n}(1-\theta)^{1-x_n}$$
$$= \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}$$

- We don't like multiplication, let's convert it into summation by taking the log:

$$L(\mathcal{D}|\,\theta) = p^{\sum x_i}(1-p)^{\sum(1-x_i)}$$

$$logL(\mathcal{D}|\,\theta) = l(\mathcal{D}|\,\theta) = \log(\theta)\sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i)$$

Maximum likelihood estimation

- How to optimize p?

$$\frac{\partial l(\mathcal{D}|\theta)}{\partial\theta} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{\sum_{i=1}^{n}(1 - x_i)}{1 - \theta} = 0$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$