

2022

INTRODUCTION TO NUMERICAL ANALYSIS

Lecture 3-4:
Modeling of Data:
Probability & Probability Distributions

Kai-Feng Chen
National Taiwan University

DESCRIBE YOUR DATA, STATISTICALLY

- Consider a set of data (collected from your experiments, or whatever source), you may want to describe / summarize / fit your data according to a model with some **adjustable parameters**.
- The model can be a collection of easy-handling functions, such as polynomials, or Gaussians, etc, and reuse the model somewhere else with some extrapolation or interpolation, or even use it to predict the next out coming data point.
- Or the model can be derived from the **underlying physics theory**, and the adjustable parameters are related to some physical parameters. We can fit to the data in order to provide an estimate of the parameter and probe to the underlying physics information.

DESCRIBE YOUR DATA, STATISTICALLY (II)

- A general approach is usually carried out by defining a “**merit function**” (or the loss function, if you prefer the same language as we introduced in the ML lecture) which represents the agreement between the model with a given set of parameters and the data. The **best-fit parameters** can be estimated by minimizing or maximizing the function.
- One of the common issues is the data is not exact in general. *It contains uncertainties*. These uncertainties have to be taken into account in the fit for a proper estimate.
- On the other hand, we may also want to know how accurate your measurement is, ie. the **uncertainties of the resulting best-fit parameters** are also needed to be calculated.

THE PROBABILITY

- **Probability** is the measure of the likelihood that an event will occur, quantified as a number between [0,1]:
 - 0 = impossibility; 1 = certainty.
 - The higher the probability of an event, the more likely it is that the event will occur.
- Classical probability is defined by

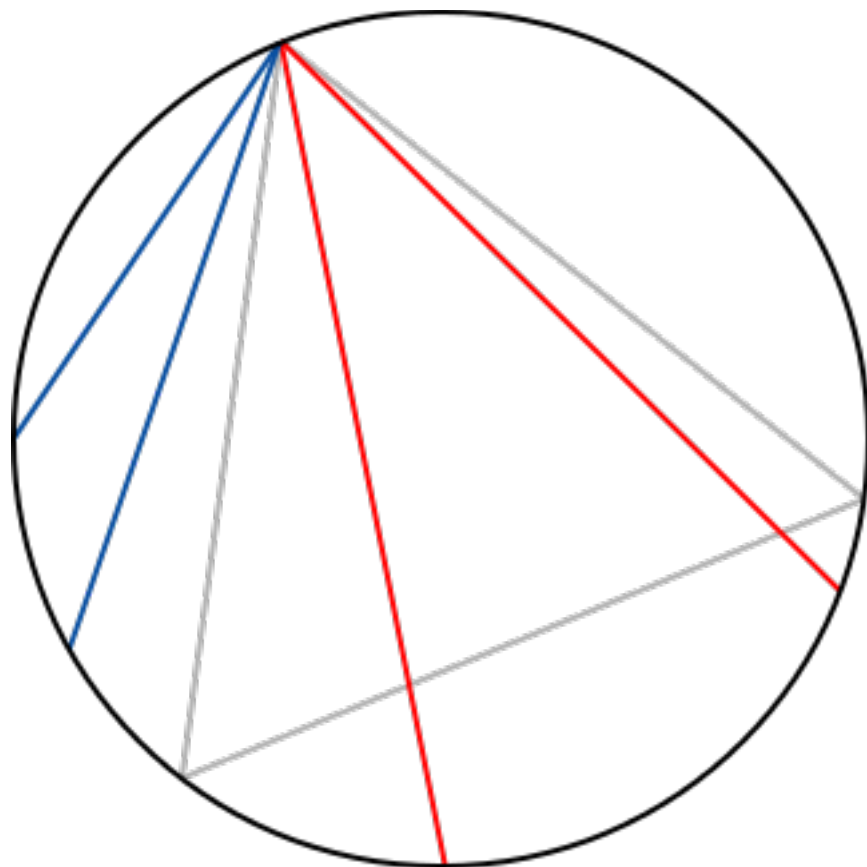
$$P = \frac{\text{Number of favorable cases}}{\text{Number of total cases}}$$

- Assuming all of the cases are *equally possible*.
- This only works for **discrete cases** rigorously.
- Problems in continuous cases (*to be discussed*).



CONTINUOUS CASES?

- A typical example of problematic probability definition in non-discrete cases, e.g. the **Bertrand's paradox**:
 - Remember our homework assignment — given a randomly chosen chord on a circle, what is the probability that the chord's length is larger than the side of the inscribed triangle?



The "random endpoints" method

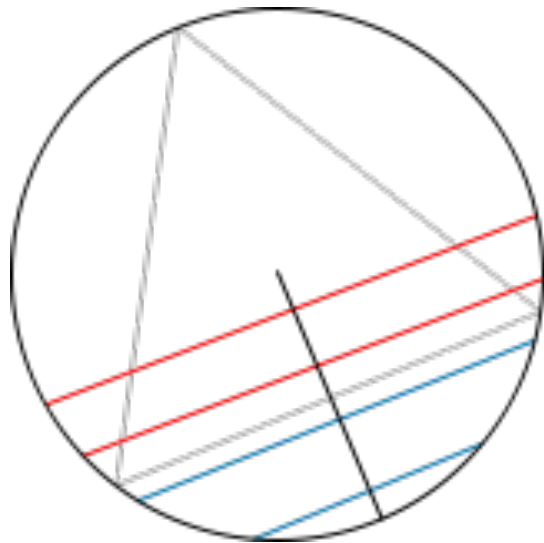
Choose two random points on the circumference of the circle and draw the chord joining them. The probability that a random chord is longer than a side of the inscribed triangle is **1/3**.

...is this always true?

CONTINUOUS CASES? (2)

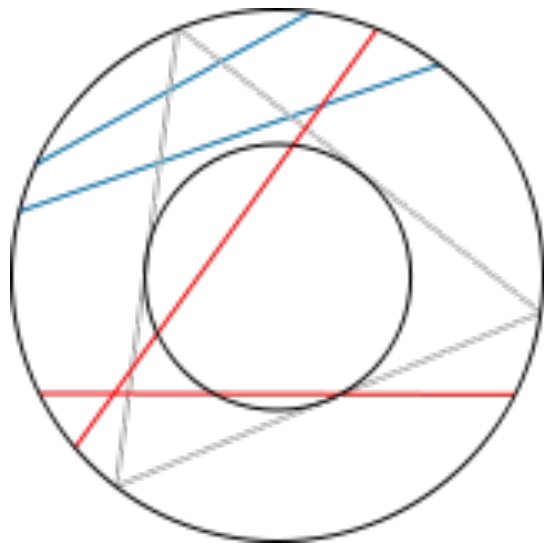
- If one considers slightly different methods, for example:

The "random radius" method



Choose a radius of the circle, choose a point on the radius and construct the chord through this point and perpendicular to the radius. The side of the triangle bisects the radius, therefore the probability is $1/2$.

The "random midpoint" method



Choose a point anywhere within the circle and construct a chord with the chosen point as its midpoint. The chord is longer if the chosen point falls within a circle of radius $1/2$. Thus the probability is $1/4$.

“Random choice” is not a well defined concept in this case; some classical probability concepts become arbitrary until we move to discuss the **probability density functions**.

FORMAL DEFINITION OF PROBABILITY

- **Mathematical probability**: define Ω to be the set of all possible *elementary events* X_i , which are exclusive (ie. occurrence of one of them implies none of others occurs). The probability of the occurrence of X_i , $P(X_i)$, to obey the **Kolmogorov axioms**:

(a) $P(X_i) \geq 0$ for all i

(b) $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$

(c) $\sum_{\Omega} P(X_i) = 1$

more complex probability expressions can be deduced for non-elementary events.

We require operational definitions which allows us to measure probabilities: **Frequentist probability** and **Bayesian probability**. Both of them satisfy the Kolmogorov axioms.

FREQUENTIST PROBABILITY

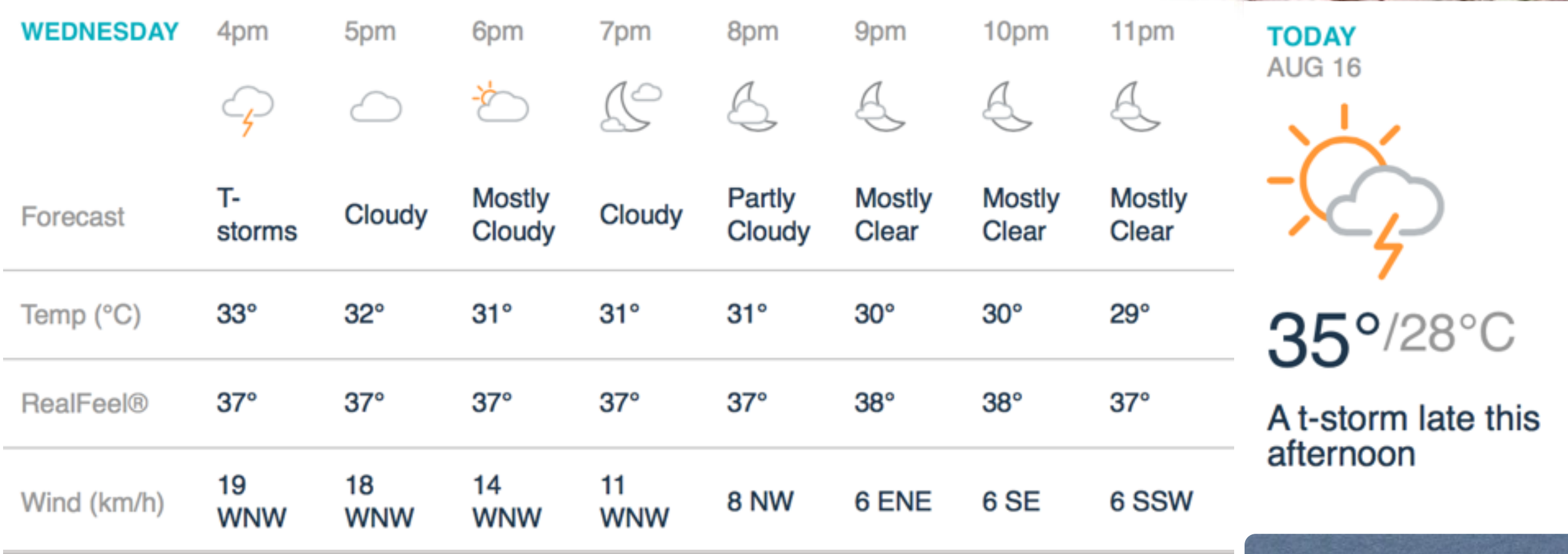
- **Frequentist probability** is in fact, defined along **experiments**.

Consider # of events of type X is n , and total # of events is N obtained from a series of experiments, then the frequentist probability that any single event will be of type X can be defined as

$$P(X) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

- Obviously this definition requires an infinite number of experiments, and it cannot be the real case! But as long as it is in principle possible always to perform one more experiments, a targeting accuracy can be obtained.
- However, this definition implies an important restriction: it can be only applied to *repeatable experiments*!

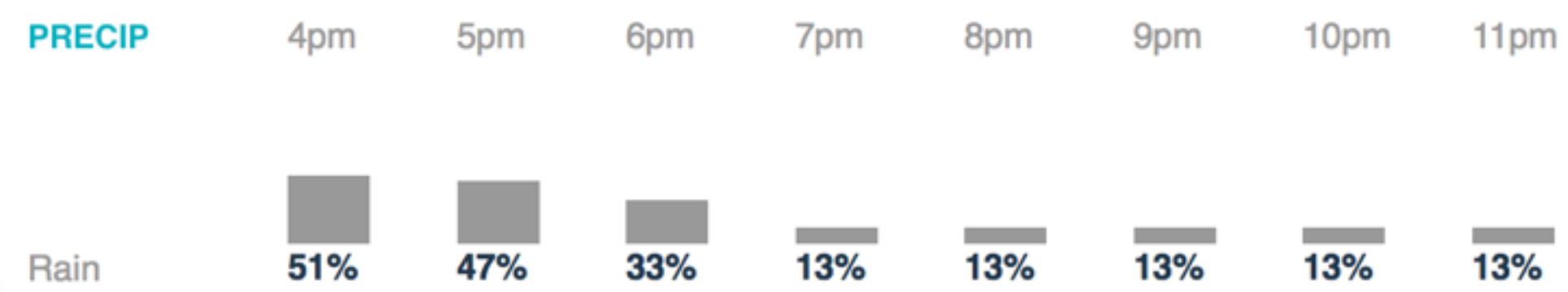
A FAMILIAR “BROKEN” CASE?



TODAY
AUG 16

35°/28°C

A t-storm late this afternoon



This is definitely **NOT** a **Frequentist probability**, since one cannot repeat the experiments!

Unless you want to talk about multiverse...

BAYESIAN PROBABILITY

- In order to define a probability that can be applied to non-repeatable experiments, we have to replace it by something else: the *degree of belief*, which is the basis of **Bayesian probability**.
- The idea is to determine how strongly a person believes that X will occur by determining how much he would be willing to bet on it, assuming that he wins of fixed amount of X does later occur and nothing if it fails to occur.
- $P(X)$ is defined as the largest amount he would willing to bet, divided by the amounts he stands to win.
- Although all these statement may sound strange, this definition does obey the **Kolmogorov axioms**.

BAYESIAN PROBABILITY (2)

- Bayesian probability is an interpretation of the concept of probability, which is interpreted as reasonable expectation representing **a state of knowledge** or as **quantification of a personal belief**.
- Properties of (*subjective*) Bayesian probability:
 - It is as much a property of observer as it is of the system being observed.
 - It depends on the state of the observer's knowledge, and will in general change as the observer obtains more knowledge.
- For example, **P(tomorrow is the end of world)** and **P(God does exist)** do exist, which cannot be defined in frequentist way!

50%

70%

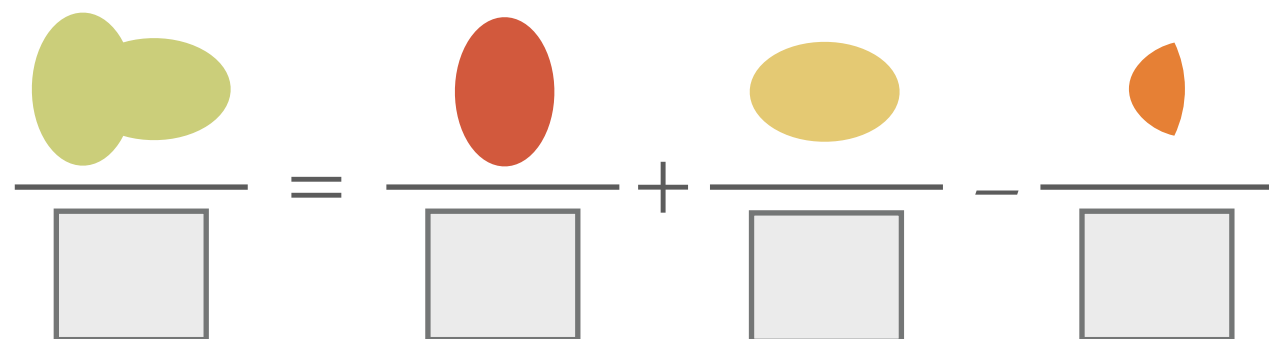
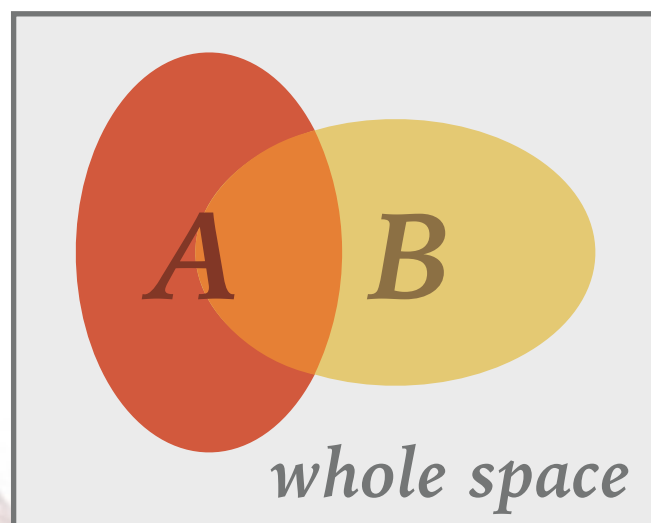
85%

Give me a probability you think you will
become a billionaire in 10 years.

PROPERTIES OF PROBABILITY

- For any probability satisfies **Kolmogorov axioms**, the following discussions do apply.
 - Consider a set A of elementary event X_i , we denote $P(A)$ as the probability that an X_i in set A occurs.
 - For two non-exclusive sets A and B , the probability of an event occurring in A or in B , or in both can be obtained by the addition law:

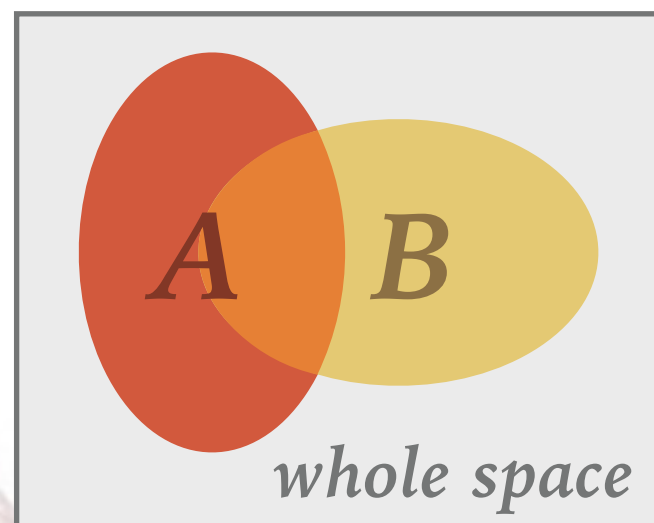
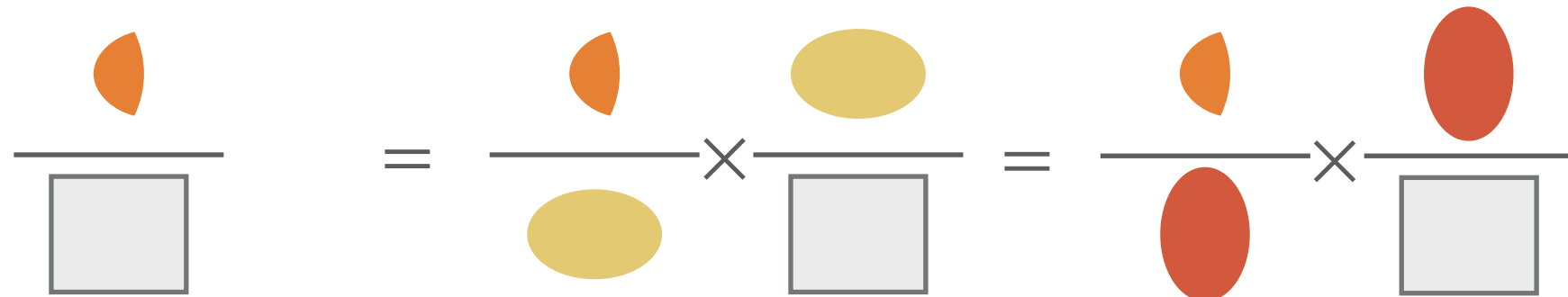
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



CONDITIONAL PROBABILITY

- Then the conditional probability, $P(A|B)$, the probability that an elementary event, known to belong to the set B , and is also a member of set A :

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$



Sets A and B are said to be independent (occurrence of B is irrelevant to the occurrence of A) if

$$P(A|B) = P(A)$$

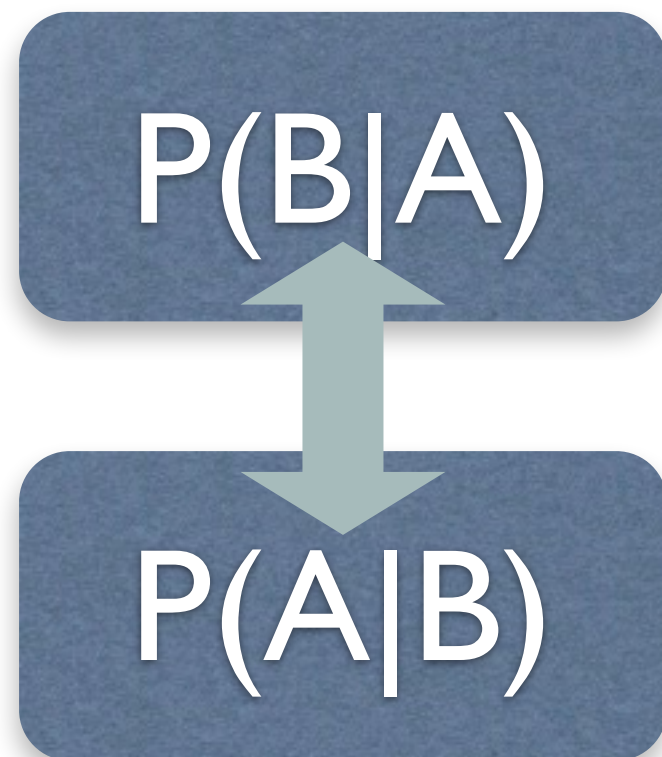
or

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

BAYES THEOREM

- **Bayes theorem** describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- A common usage is to invert conditional probabilities.

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$



the likelihood of observing event B given that A is true.

e.g. **A: typhoon is landing**
B: it is raining

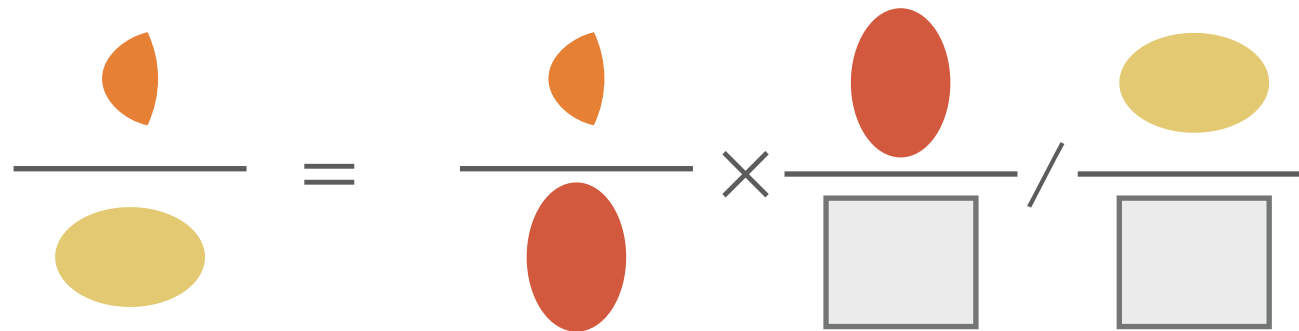
the posterior probability of A is true given observing B.

e.g. **B: it is raining**
A: typhoon is landing

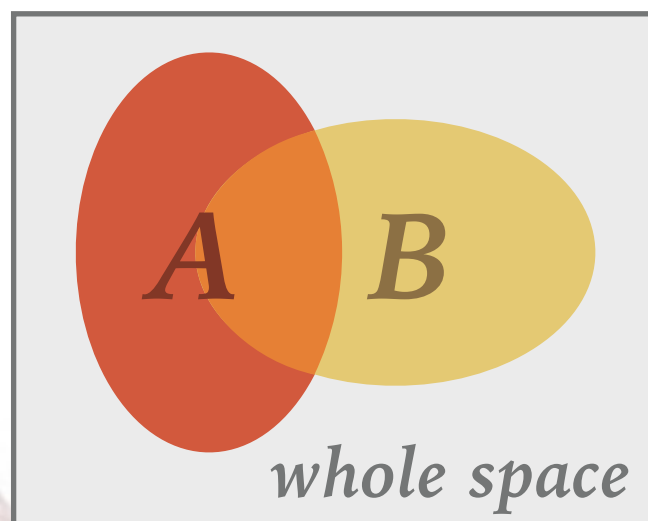
BAYES THEOREM FOR DISCRETE EVENTS

- The theorem which links $P(A|B)$ to $P(B|A)$ is the **Bayes theorem**, which follows the definition of conditional probability:

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$



This is kind of obvious from the formulation, but it may be totally straightforward if one considers a real case.



Remark: using the above Bayes theorem does not imply you are using a Bayesian probability. The Bayes theorem applies to Frequentist probability as well.

TERMINOLOGY

- When involving hypotheses testing (e.g. the idea / assumption is true), we are entering the Bayesian framework. Therefore the Bayes theorem can be written as

$$P(\theta_i | X^0) = P(X^0 | \theta_i) \times P(\theta_i) / P(X^0)$$

- $P(\theta_i | X^0)$: the **posterior probability** for hypothesis θ_i , given data X^0 have been observed.
- $P(X^0 | \theta_i)$: the probability of obtaining the observed data X^0 , given hypothesis θ_i , which must be known.
- $P(\theta_i)$: the **prior probability** and represents the knowledge or degree of belief before the experiment was performed.
- $P(X^0)$: normalization, since $\sum_i P(\theta_i | X^0) = P(X^0)$, but this may not be known. If this is the case, a weaker form is usually given by

$$P(\theta_i | X^0) \propto P(X^0 | \theta_i) \times P(\theta_i)$$

DRUG TEST RESULTS

PASSED

FOR INTERNAL USE ONLY!

EMPLOYEE INFORMATION

EMPLOYEE NAME: John Doe

STREET ADDRESS: 12345 Main Street

STATE: Anywhere

ZIP CODE: 123456

BAYES THEOREM EXAMPLE

(I)

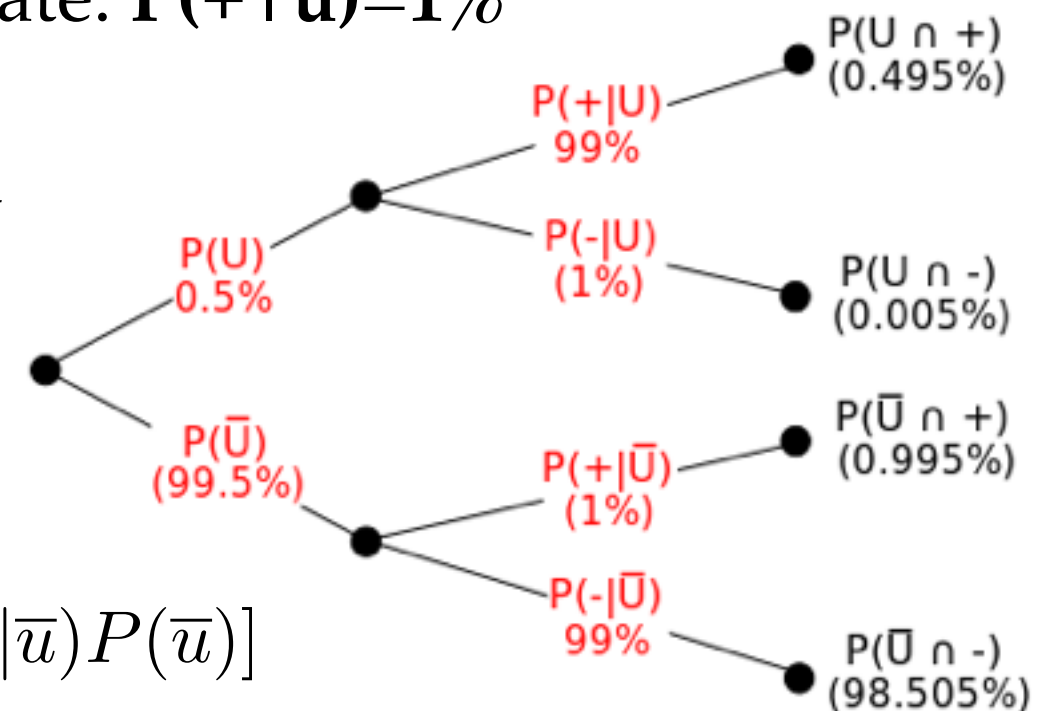
- Let's consider a classical problem – suppose a drug test is **99%** sensitive and **99%** specific:
 - **$P(+|u)=99\%$** positive results for drug users
.....with a failing detection rate: **$P(-|u)=1\%$**
 - **$P(-|\bar{u})=99\%$** negative results for non-drug users
.....with a false alarm rate: **$P(+|\bar{u})=1\%$**

- Suppose that **0.5% of people are users of the drug**. What is the probability **$P(u|+)$** , a randomly selected individual with a positive test is a user?

$$\begin{aligned}P(u|+) &= P(+|u)P(u)/P(+)
 &= P(+|u)P(u)/[P(+|u)P(u) + P(+|\bar{u})P(\bar{u})]
 &= 0.99 \times 0.005/[0.99 \times 0.005 + 0.01 \times 0.995]\end{aligned}$$

$\approx 33\%$

It's as low as 33% in fact!



BAYES THEOREM EXAMPLE

(II)

- Suppose that **5% of people are users of the drug**. What is the probability $P(u|+)$ now?

$$\begin{aligned}P(u|+) &= P(+|u)P(u)/P(+)\end{aligned}$$
$$= P(+|u)P(u)/[P(+|u)P(u) + P(+|\bar{u})P(\bar{u})]$$
$$= 0.99 \times 0.05 / [0.99 \times 0.05 + 0.01 \times 0.95]$$

$\approx 84\%$ Now the chance of getting a drug user is much higher!

Point # 1

The chance of a positive test = a user depends on the absolute rate of drug users!

Point # 2

The probabilities discussed here can be defined a frequentist probability.

All these sound very reasonable, but in reality it is not so easy to know the absolute rate of drug users $P(u)$!

BAYES THEOREM EXAMPLE

(III)

- **Bayes theorem gives us a nice method to access the underlying probability.** But let's practice a "no-so-natural" example:
- Somebody gave you a "magic coin", and claimed that it only shows head in coin tosses. But you only allowed to perform the experiment (toss it!) for 3 times and all tosses indeed show the head, e.g.
 - **$P(\text{all 3 heads} \mid \text{magic coin}) = 1$**
 - **$P(\text{all 3 heads} \mid \text{normal coin}) = (0.5)^3 = 0.125$**
 - **$P(\text{not 3 heads} \mid \text{magic coin}) = 0$**
 - **$P(\text{not 3 heads} \mid \text{normal coin}) = 1 - 0.125 = 0.875$**
- Question: given the experimental result, what is the probability of this coin is really a magic coin? ie. what is $P(\text{magic coin} \mid \text{all 3 heads})$?



BAYES THEOREM EXAMPLE

(IV)

■ **Answer: we do not know!** $P(\text{magic coin} \mid \text{all 3 heads})$ cannot be calculated since we do not know **the prior $P(\text{magic coin})$** !

- If you really believe in magic, e.g. $P(\text{magic coin}) \sim 0.99$

$$\begin{aligned} P(\text{magic} \mid 3 \text{ heads}) &= P(3 \text{ heads} \mid \text{magic}) P(\text{magic}) / P(3 \text{ heads}) \\ &= 1.0 \times 0.99 / [0.99 \times 1.0 + 0.01 \times 0.125] \approx 99.99\% \end{aligned}$$

- If you do not believe in magic, e.g. $P(\text{magic coin}) \sim 0.01$

$$\begin{aligned} P(\text{magic} \mid 3 \text{ heads}) &= P(3 \text{ heads} \mid \text{magic}) P(\text{magic}) / P(3 \text{ heads}) \\ &= 1.0 \times 0.01 / [0.01 \times 1.0 + 0.99 \times 0.125] \approx 45\% \end{aligned}$$

The result does depend on your degree of belief!

You may find this interpretation is kind of odd, given **the prior** is something you cannot avoid when introducing Bayesian probability!

COMMENT: THE PRIOR

- The degrees of belief in a hypothesis depends on *the experimental results* and *the prior probability before the experiment*. Or, one can say that **Bayesian statistics is subjectivity by definition**.
- Surely for the physicists this is not very appealing; people tried very hard to look for a way to avoid introducing prior into the experiments, but without a real success.
 - Bayesian may comment that it is actually intersubjective, i.e. the real nature of learning and knowing physics.
- **Frequentist approach is generally preferred by a large fraction of physicists** (*probably the majority, but Bayesian statistics is getting more and more popular in many application, also thanks to its easier application in many of the cases*).

INTERMISSION

- The probabilities can be either Frequentist defined or Bayesian defined. Try to identify some of the probabilities used in daily life if they can be defined in Bayesian way or in Frequentist way.
- In the previous two Bayes theorem examples, the resulting probabilities are all depending on the given prior probability (only). You can try to derive the exact relationship between

$P(u|+)$ versus $P(u)$

$P(\text{magic coin}|\text{all 3 heads})$ versus $P(\text{magic coin})$

and see how it works.





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

Time for all lovely distributions!

CONTINUOUS RANDOM VARIABLE

- A random event may be associated a random variable X , which takes different possible numerical values $X_1, X_2, \dots$, corresponding to the different possible outcome.
- Those probabilities $P(X_1), P(X_2), \dots$, form a **probability distribution**.
- When an experiment consists of N repeated observations of the same random variable X , it can be considered as the single observation of a random vector $X = \{X_1, X_2, \dots, X_N\}$.
- Instead of probability for discrete cases, now we can generalize probabilities of events to probability distributions of random variable, using the tools like **probability density functions**.

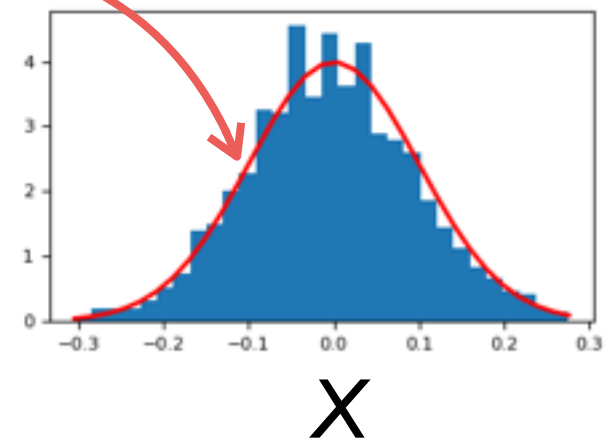
PROBABILITY DENSITY FUNCTION (PDF)

- Consider a random histogram of X , collecting the data with a pistol shooting location for example.
- The probability distribution of finding particles at X is denoted by $P(X)$, which is still discrete.
- However it is more convenient to describe this using a continuous function $f(X)$ by introducing infinite small steps:

$$f(X) = \lim_{\Delta X \rightarrow 0} \frac{P(X)}{\Delta X}$$

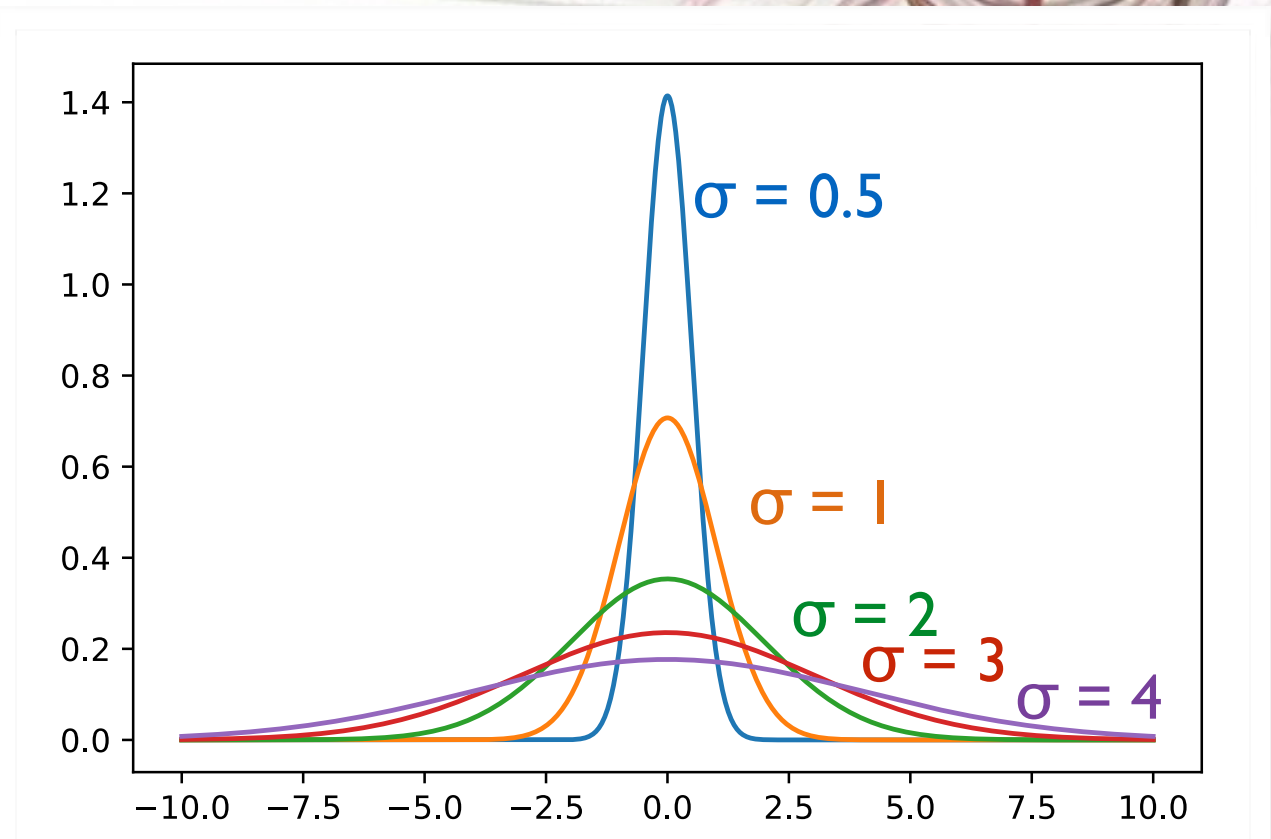
- The **probability density function $f(X)$** represents a probability density per unit array length of X . The normalization condition must be held:

$$\int_X f(X) dx = 1$$



DRAWING GAUSSIANS

- Here are a brief example of drawing Gaussian PDFs with various width parameter.
- All of these PDFs should be **normalized** to have area = 1.



```
gaus = lambda x,mu,sigma: np.exp(-0.5*(x-mu)**2/sigma**2)/sigma/2**0.5

fig = plt.figure(figsize=(6,4), dpi=80)
x = np.linspace(-10.,10.,1000)
for w in [0.5,1.,2.,3.,4.]:
    y = gaus(x, 0., w)
    plt.plot(x,y)
plt.show()
```

↑↑ You can also take `scipy.stats.norm!`

I304-example-01.py (partial)

PROPERTIES OF DISTRIBUTIONS

- Several useful quantities which characterize probability distributions. The PDF $f(X)$ is used as a weighting function to obtain the corresponding quantities.

- The **expectation E** of a function $g(X)$ is given by

$$E(g) = \langle g(X) \rangle = \int_{\Omega} g(X) f(X) dx$$

where Ω is the entire space.

- The **mean μ** is simply the expected value of X :

$$\mu = E(X) = \langle X \rangle = \int_{\Omega} X f(X) dx$$

- The expectation of the function $(X - \mu)^2$ is the **variance V** :

$$V = \sigma^2 = E((X - \mu)^2) = \int_{\Omega} (X - \mu)^2 f(X) dx = \int_{\Omega} X^2 f(X) dx - \mu^2$$

EXPECTATION OF POISSON

- The following example code is to calculate the expectation of random variable n with a Poisson PDF (or PMF, probability mass function) is μ :

$$P(n) = \frac{\mu^n e^{-\mu}}{n!}$$

```
from scipy.stats import poisson
```

```
mu = 4.5
```

```
E = 0.
```

```
x = np.linspace(0.,15.,16)
```

```
y = poisson.pmf(x,mu)
```

```
for n,p in zip(x,y):
```

```
    E += p*n
```

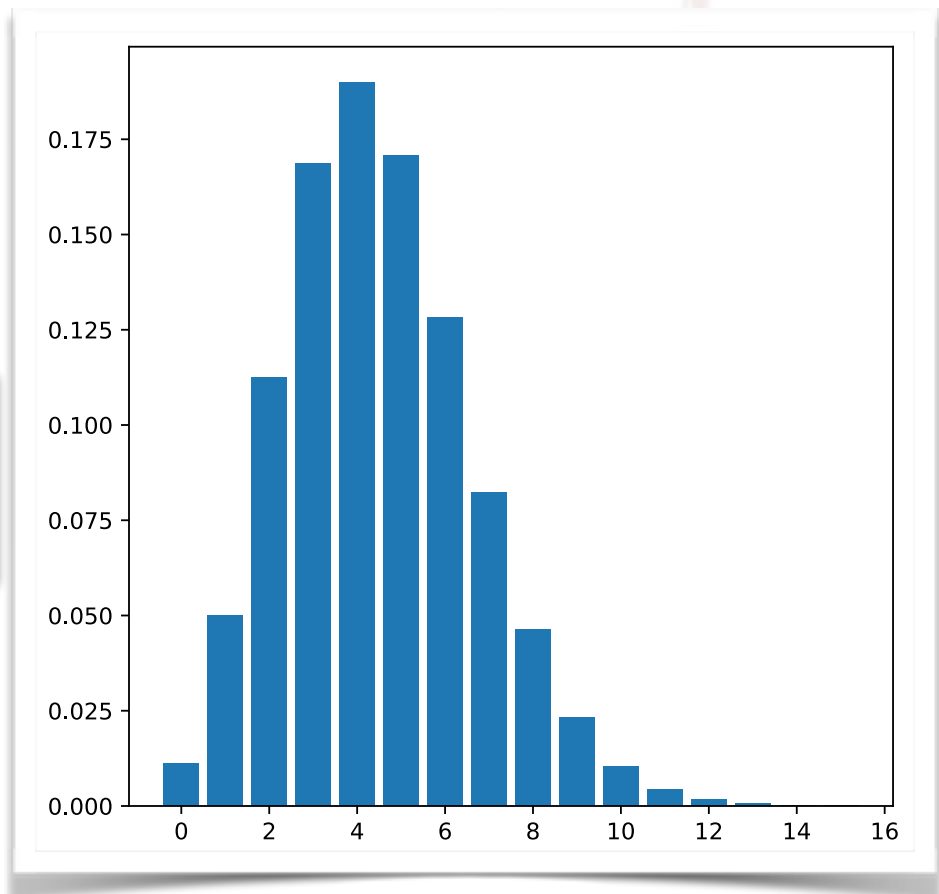
```
print("Expectation:", E)
```

```
fig = plt.figure(figsize=(6,6), dpi=80)
```

```
plt.bar(x,y)
```

```
plt.show()
```

Expectation:
4.49966852764



I304-example-02.py (partial)

COVARIANCE AND CORRELATION

- **Covariance** and **correlation** are two further useful numerical characteristics. Consider a joint density $f(X, Y)$ of two variables, the covariance is the expectation of $(X - \mu_X)(Y - \mu_Y)$:

$$\text{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y)$$

- Another one is the correlation coefficient, which is defined by

$$\text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- When there are more than 2 variables, the covariance (and correlation) can be still defined for each 2D joint distribution for X_i and X_j . The matrix with elements $\text{cov}(X_i, X_j)$ is called the **covariance matrix** (or *variance/error matrix*). The diagonal elements are just the variances:

$$\text{cov}(X_i, X_i) = E(X_i^2) - E(X_i)^2 = \sigma_{X_i}^2$$

COMMONLY USED DISTRIBUTIONS: BINOMIAL

- Consider a distribution of “# of successes” with N trials, while each trial has a probability of success p , which should follow the binomial distribution:

$$P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

- Average: $E = \langle n \rangle = Np$
- Variance: $V = \sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = Np(1-p)$
- Frequently used for efficiency estimation with a limited size of sample, in this case the efficiency $\epsilon = \langle n \rangle / N = p$, the uncertainty is given by

$$\sigma_{\epsilon} = \sqrt{\frac{\epsilon(1-\epsilon)}{N}}$$

Remark:

$\sigma_{\epsilon} \rightarrow 0$ when $\epsilon \rightarrow 0$ or 1

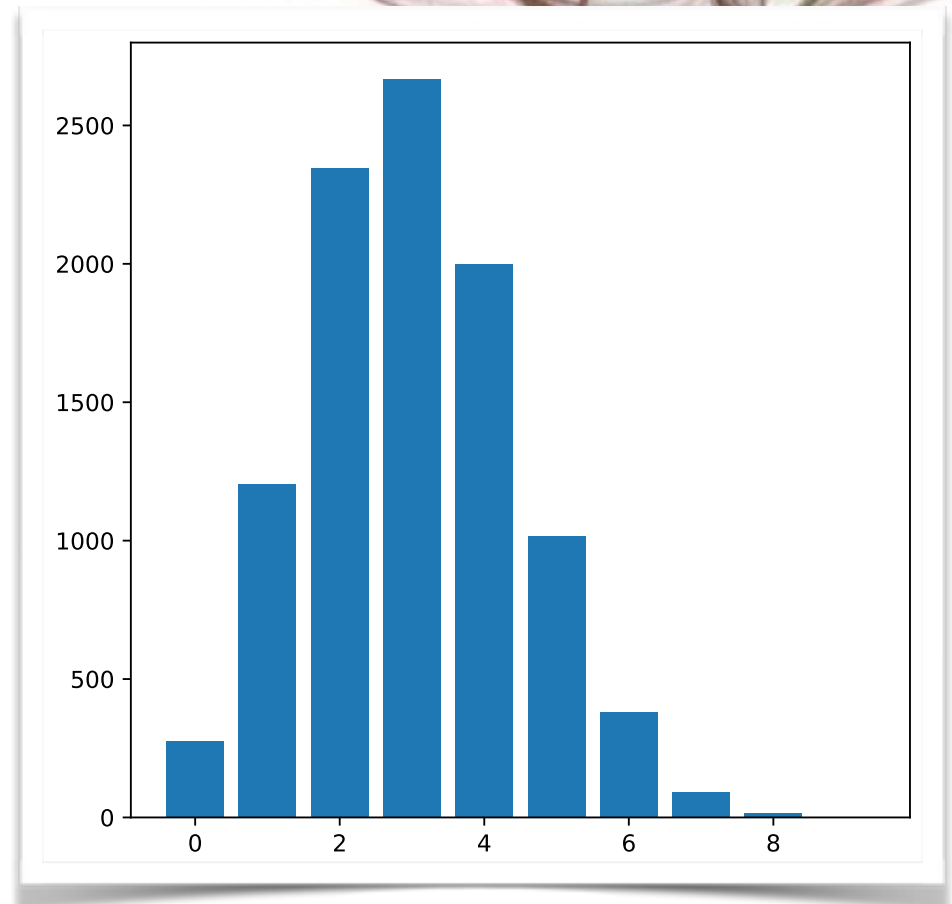
RANDOM EXTRACTION

- Suppose you have a bag of marbles, there are 3 red ones and 7 white ones. Let's define a "success", which is the extraction of a red marble out of this bag:
 - Red: $p = 3/10$
 - White: $1 - p = 7/10$
- Note the "success" can be finding an event passing your selection criteria.
- Suppose you can only do the extraction for a fixed N trials, then the # of successful trials n should follow the binomial distribution given in the previous slides!

→ When you are doing such an extraction **continuously**, the n success becomes **Poisson distribution**.

GENERATING BINOMIAL FROM THE PRINCIPLE

- Let's use random number to mimic the random extraction and see if our resulting distribution matches to the binomial.
- Set $p = 0.3$, $N = 10$. The resulting # of success n should distributed like this:



```
p, N = 0.3, 10
n = [np.sum(np.random.rand(N)<p) for i in range(10000)]

fig = plt.figure(figsize=(6,6), dpi=80)
plt.hist(n, bins=10, range=(-0.5,9.5), rwidth=0.8)
plt.show()
```

I304-example-03.py (partial)

COMMONLY USED DISTRIBUTIONS:

POISSON

- The Poisson distribution gives the probability of finding exactly n events in a given **length of time** (and/or space), if the events occur independently at **a constant rate**.
- It is a special case of binomial distribution with $p \rightarrow 0$, $N \rightarrow \infty$, $\mu = Np$ as the finite constant; as $\mu \rightarrow \infty$, the Poisson distribution converges to the Normal distribution (Gaussian).
- Properties:
 - variable: positive integer n
 - parameter: positive real number μ
 - probability function: $P(n) = \frac{\mu^n e^{-\mu}}{n!}$
 - expected value: $E(n) = \mu$
 - variance: $V(n) = \mu$



Siméon Denis Poisson

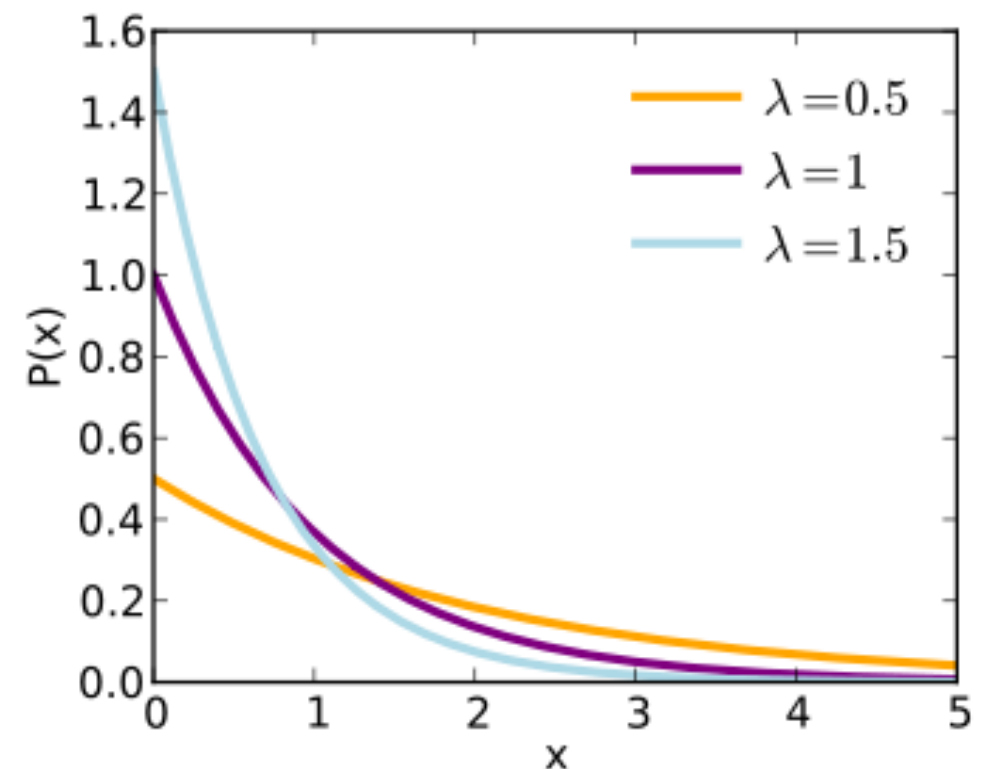
COMMONLY USED DISTRIBUTIONS: POISSON (II)

- Poisson distributions apply to various phenomena of **discrete properties** (*those that may happen 0, 1, 2, 3, ... times during a given period of time or in a given area*) whenever the probability of the phenomenon happening is constant in time or space.
- For example:
 - number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry (*quote: L. J. Bortkiewicz*).
 - number of yeast cells used when brewing Guinness beer (*quote: W. S. Gosset*).
- The time **interval between two successive events is actually exponentially distributed**, and this is true for any Poissonian process!



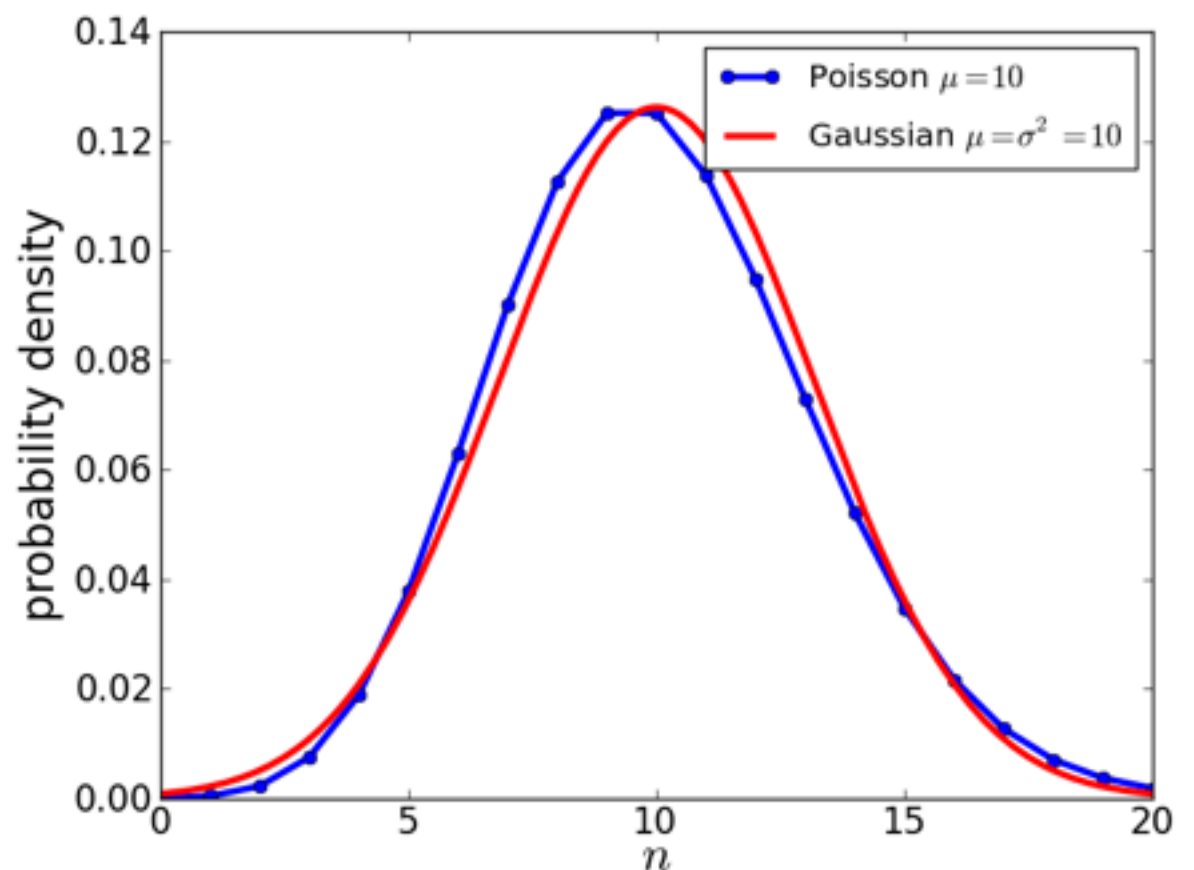
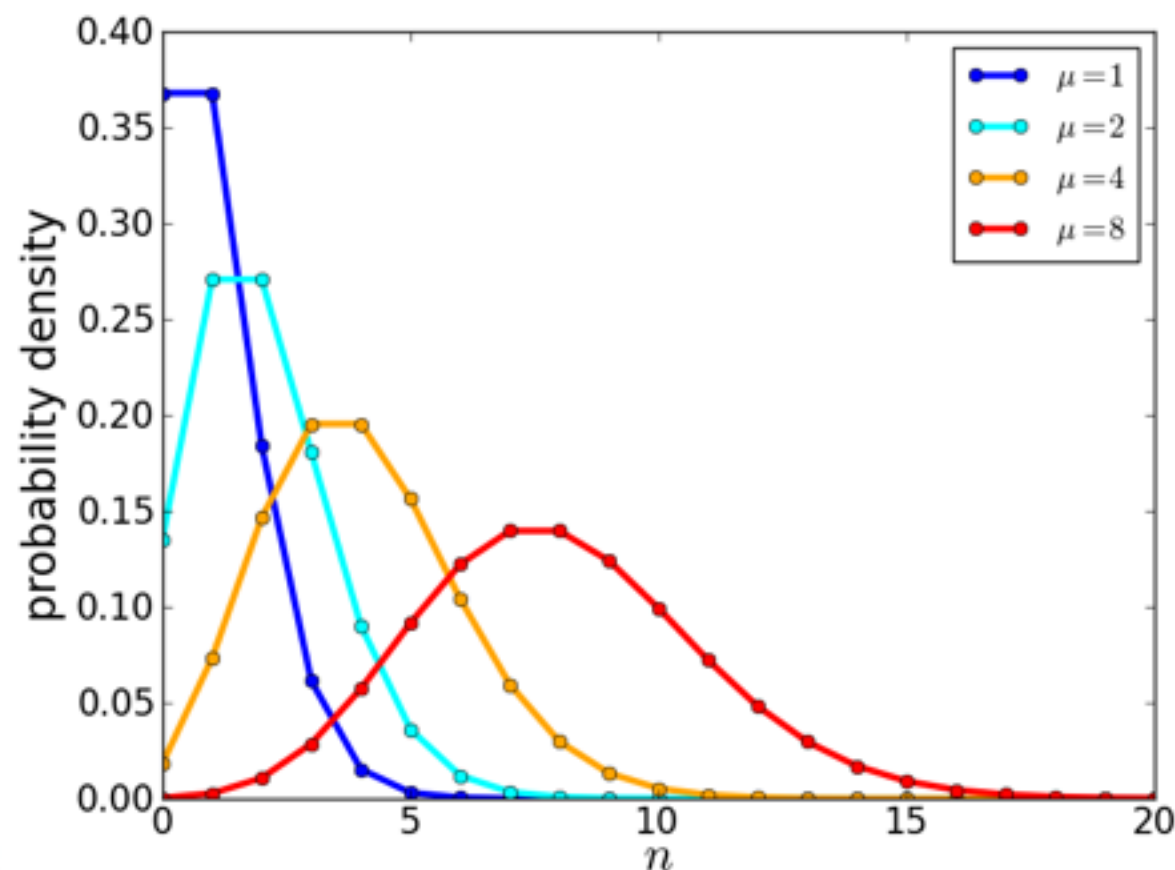
COMMONLY USED DISTRIBUTIONS: EXPONENTIAL

- Consider events occurring randomly in time, with an average of λ events per unit time.
- The Poisson distribution describe the probability of N events occurring in a time interval t ; then **the probability of no events in time t follows** the exponential distribution $\exp(-\lambda t)$.
- Properties:
 - variable: real number x
 - parameter: real numbers λ
 - probability function: $P(n) = \frac{\mu^n e^{-\mu}}{n!}$
 - expected value: $E(x) = 1/\lambda$
 - variance: $V(x) = 1/\lambda^2$



FROM POISSON TO GAUSSIAN

- When the expected value μ of the Poisson distribution increases, it converges to the **Normal distribution (Gaussian)**.
- Even the value of μ is only ~ 10 , the distribution is already rather close to a Gaussian with the same variance ($V=\sigma^2=\mu$).



COMMONLY USED DISTRIBUTIONS: NORMAL / GAUSSIAN

- Gaussian is probably the most important / well-known / useful probability distribution.

- Properties:

- variable: real number x
- parameter: real numbers μ, σ
- probability function:

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

- expected value: $E(x) = \mu$
- variance: $V(x) = \sigma^2$

- A Gaussian distribution with $\mu=0$ and $\sigma=1$ is the standard Normal density function.



Carl Friedrich Gauss

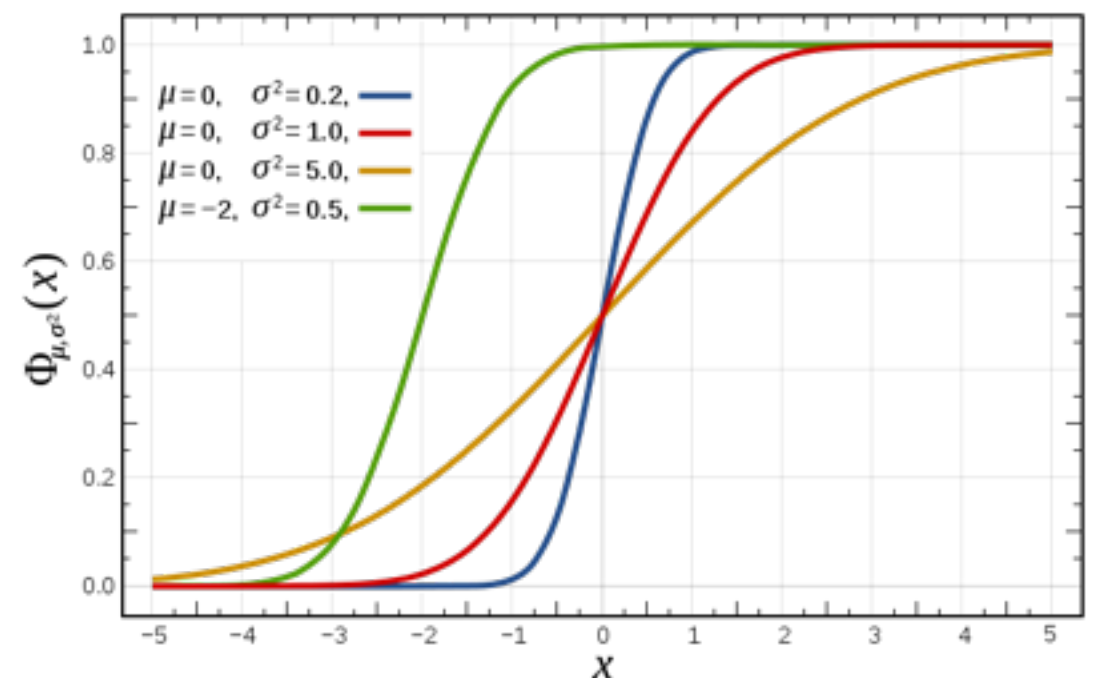
NORMAL / GAUSSIAN DISTRIBUTION (II)

- The **cumulative distribution** of the standard normal distribution can be related to the error function, **erf(x)**

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$


$$\Phi(x) = \int_{-\infty}^x G(x'; \mu = 0, \sigma = 1) dx' = \frac{1}{2} \left[1 + \text{erf} \left(\frac{x}{\sqrt{2}} \right) \right]$$

- The error function is what you can easily call within your program, if you want to calculate the integration of a Gaussian!

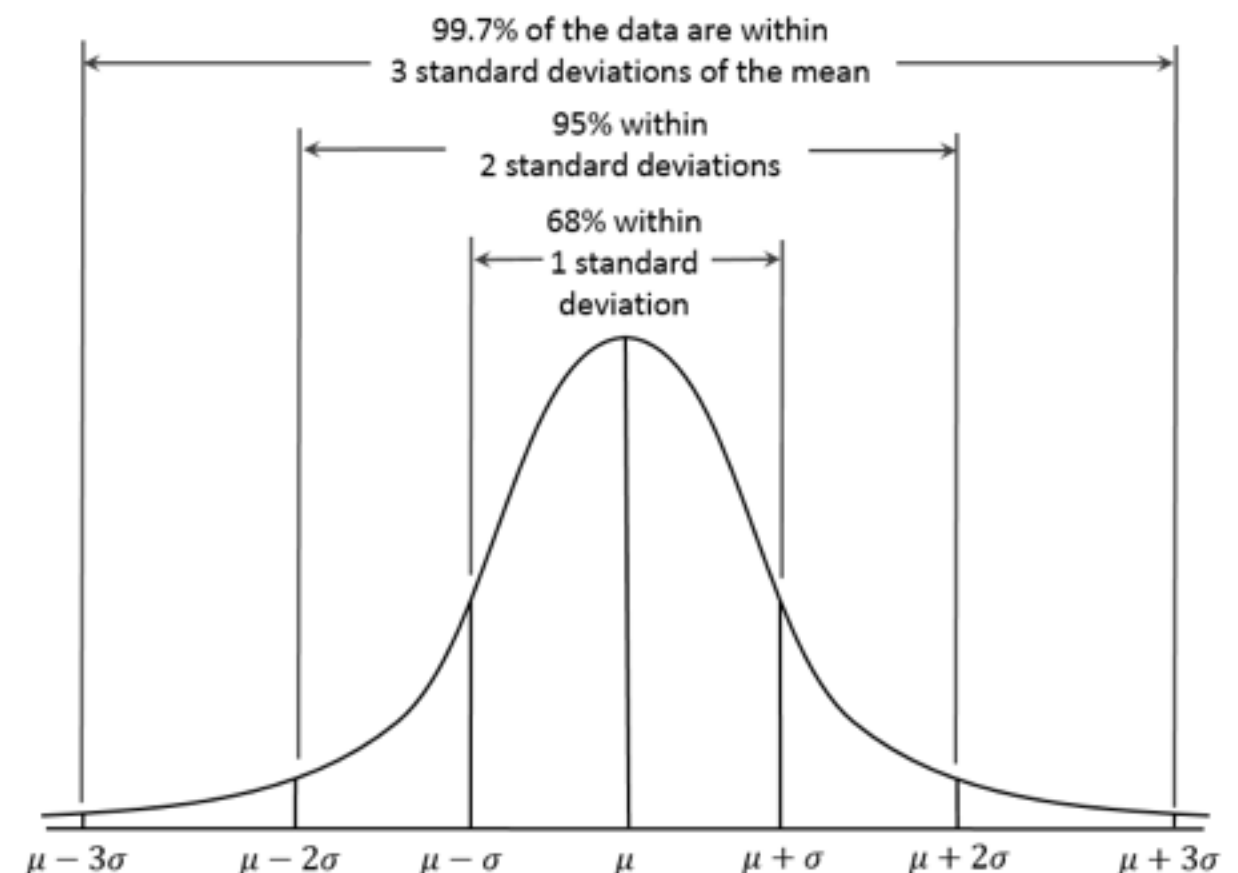


NORMAL / GAUSSIAN DISTRIBUTION (III)

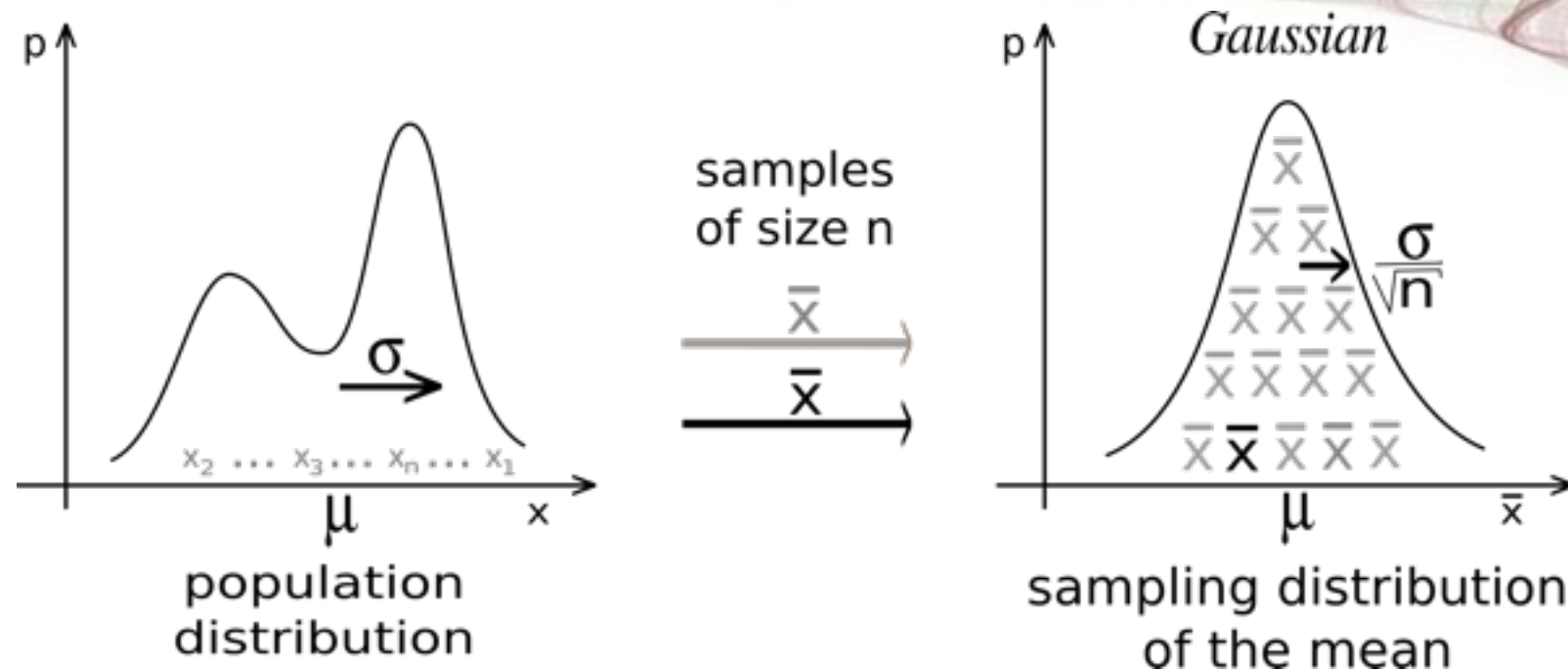
- On the other hand, the error function can be easily used to derive the **coverage probability** for a given standard deviation, e.g. **68.3% of a normal distribution is just within $\pm 1\sigma$ region, etc.**

$$p(n) = \Phi(n) - \Phi(-n) = \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right)$$

n	p(n)	1-p(n)
1σ	0.682 689	0.317 310
2σ	0.954 499	0.045 500
3σ	0.997 300	0.002 699
4σ	0.999 936	0.000 063



CENTRAL LIMIT THEOREM

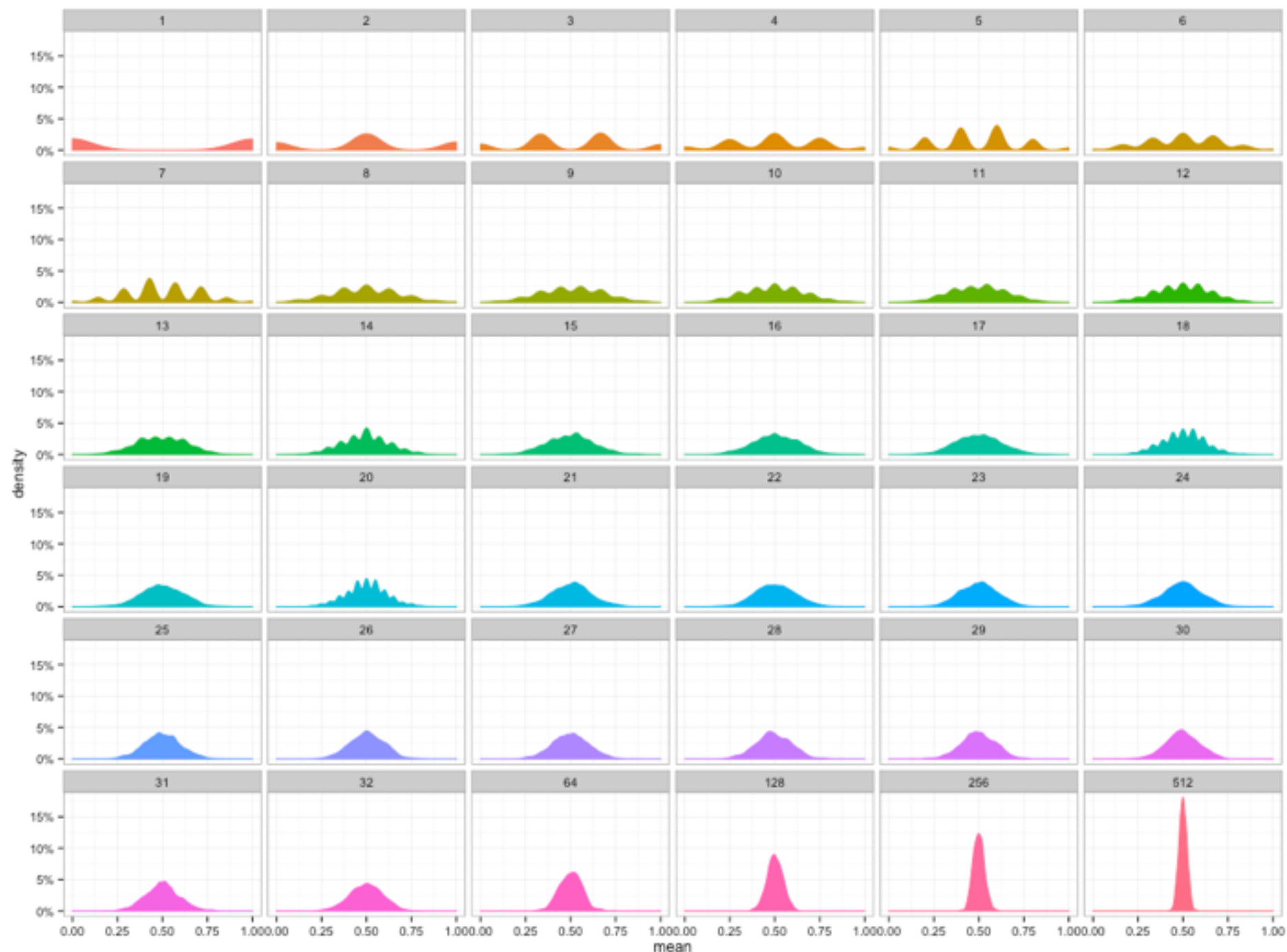


- If we have a sequence of independent variable X_i , each from a distribution with mean μ_i and variance σ_i^2 .
- The sum $S = \sum X_i$ will have a mean $\sum \mu_i$ and a variance $\sum \sigma_i^2$.
- This holds for **ANY distributions** with **finite individual means and variances** exist. The Central Limit theorem states, in the limit of large $N \rightarrow \infty$,

$$\frac{S - \sum_i^N \mu_i}{\sqrt{\sum_i^N \sigma_i^2}} \rightarrow \text{Gaussian}(x; \mu = 0, \sigma = 1)$$

CENTRAL LIMIT THEOREM

(II)

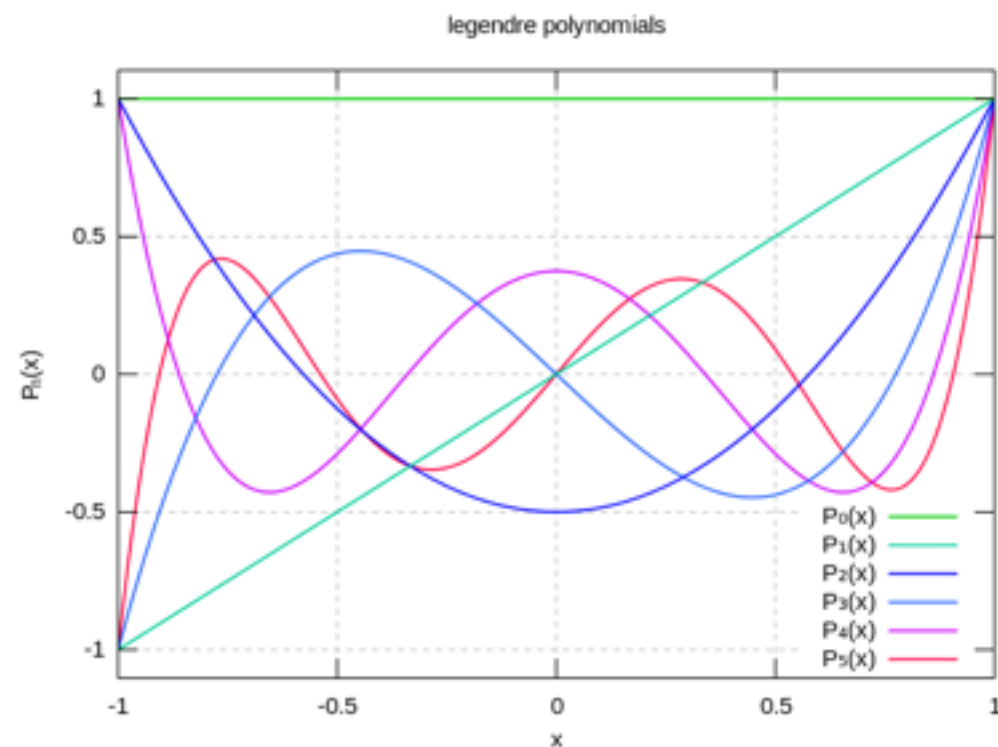


A simulation
with binomial
distributions up
to $N=512$

COMMONLY USED DISTRIBUTIONS: POLYNOMIALS

- Polynomials are probably the simplest way to model any unknown distributions. Although different definitions of polynomials are mathematically equivalent, but different polynomials indeed have different behavior.
- In particular, some of the polynomials (e.g. Legendre or Chebyshev) are orthogonal, they usually have a better behavior when expanding the order of polynomials.
- Simple polynomials:
 - **Power series:** $a_0 + a_1x + a_2x^2 + a_3x^3 + \dots = \sum_{k=0}^N a_k x^k$
 - **Laurent polynomial:** same as above but k can be negative.

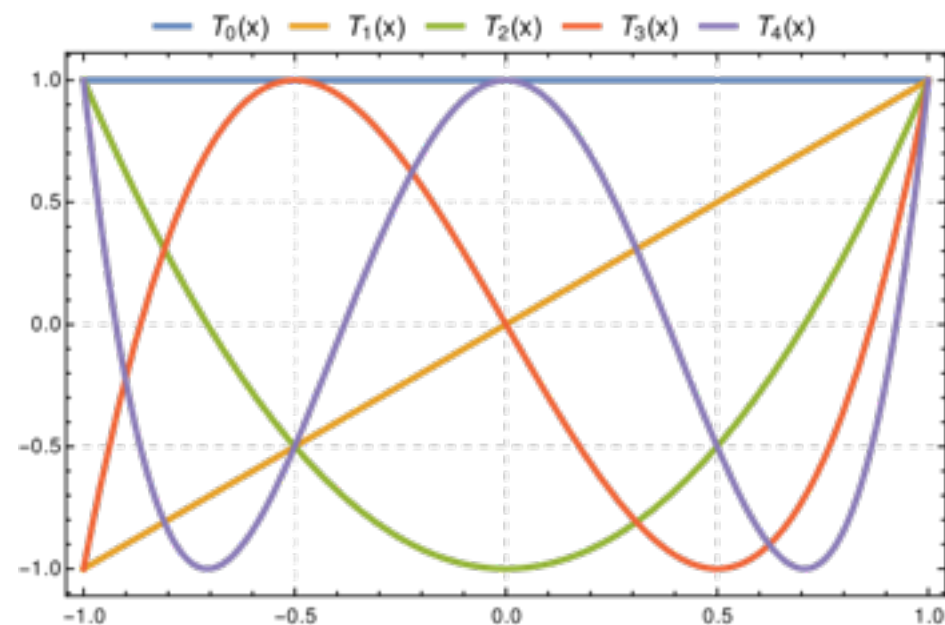
POLYNOMIALS (II)



- **Legendre polynomials:** as general solutions to Legendre's Equation, and are azimuthally symmetric.

$$P_0(x) = 1, \quad P_1(x) = x$$

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$



- **Chebyshev polynomials:** as a sequence of orthogonal polynomials and can be defined recursively.

$$T_0(x) = 1, \quad T_1(x) = x$$

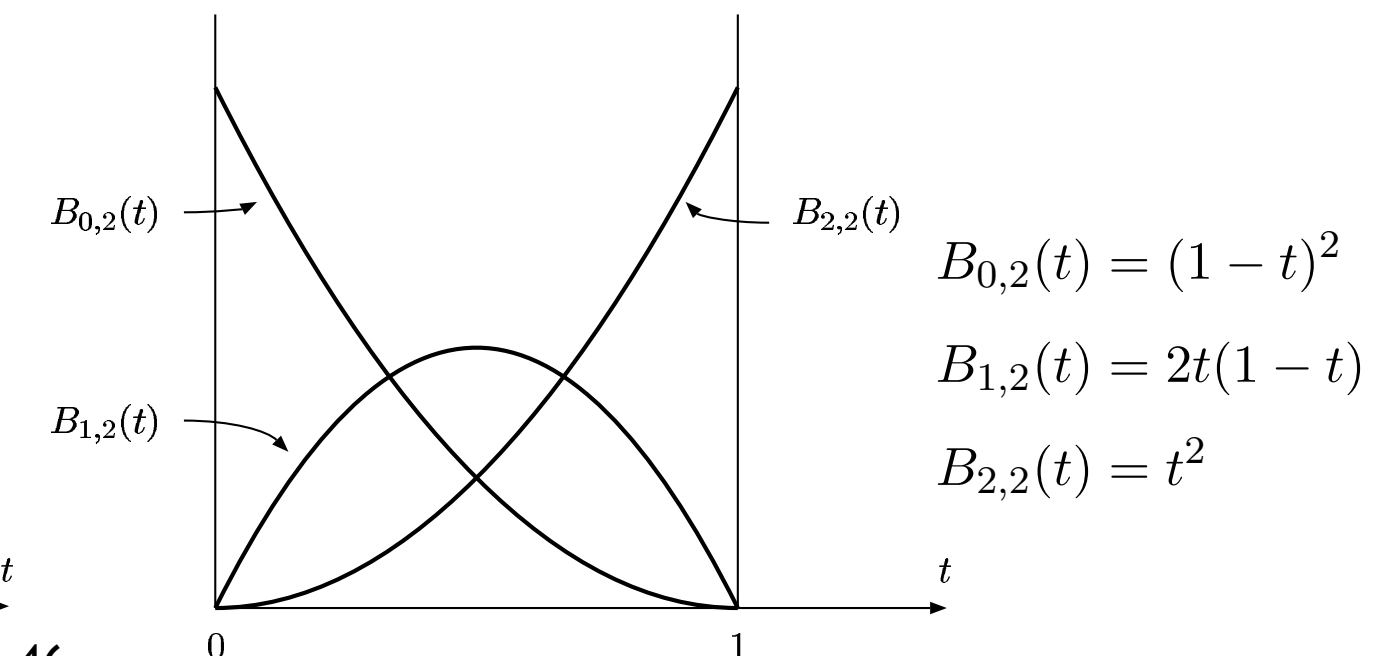
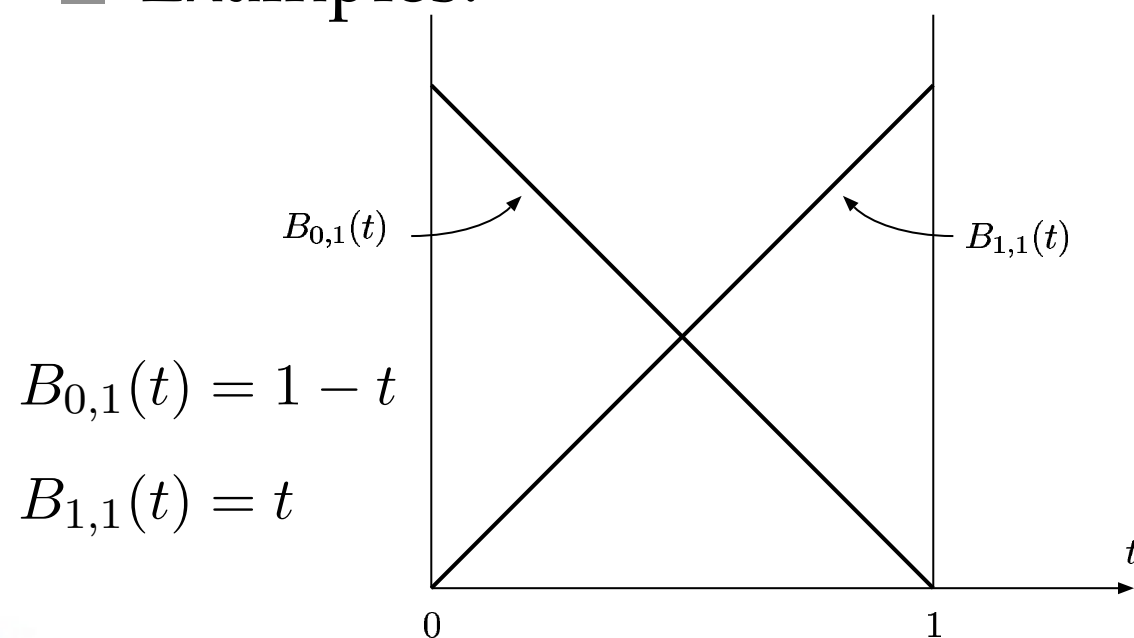
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

POLYNOMIALS (III)

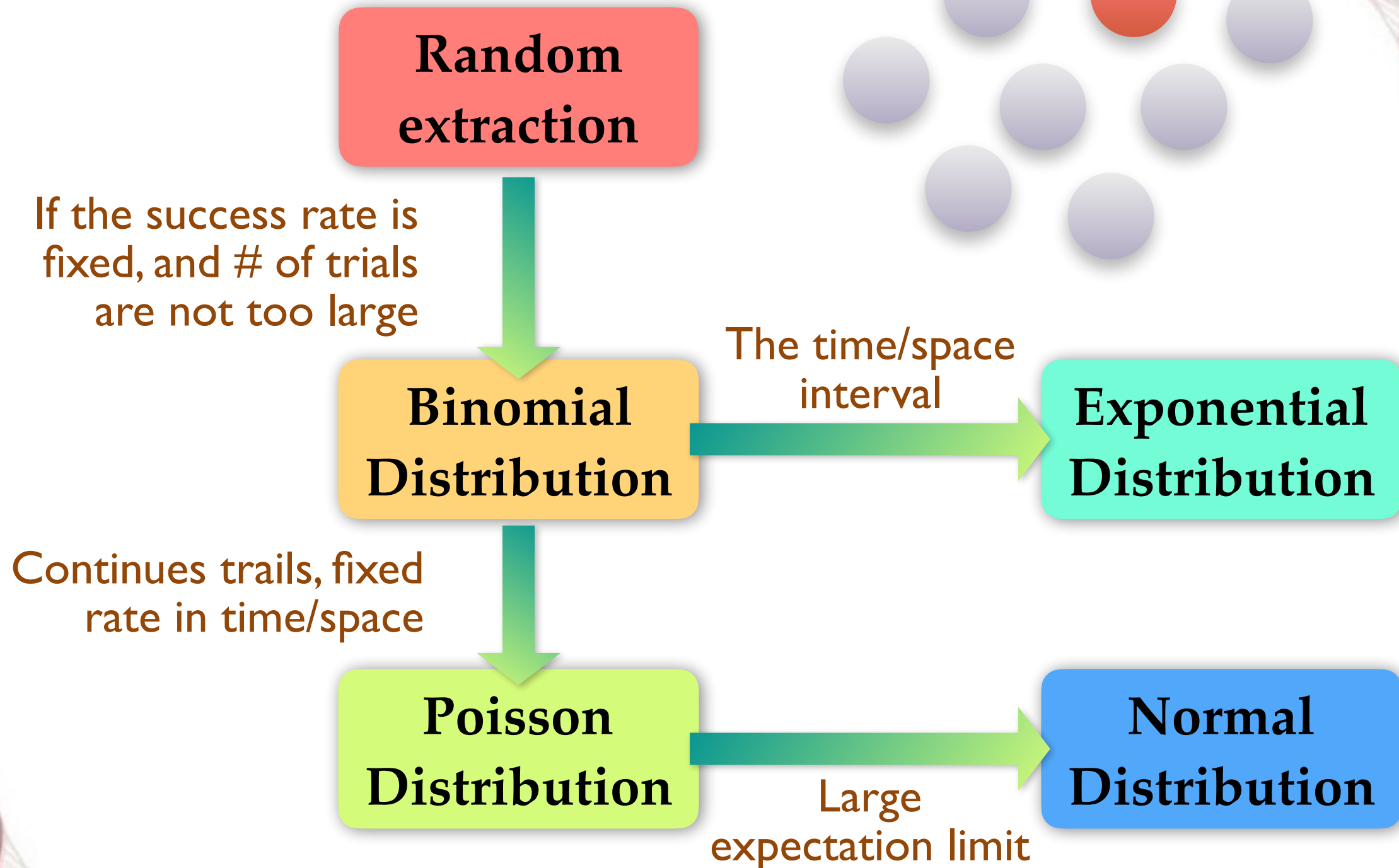
- **Probabilities should be always “positive defined”**, but this is not the case for usual power-series based polynomials. The function can easily go to negative and break the evaluation of probability.
- **Bernstein polynomials** are constructed with sets of non-negative bases and are generally convenient for PDF modeling.
- Bernstein polynomials of degree n are defined by

$$B_{i,n}(t) = \frac{n!}{i!(n-i)!} t^i (1-t)^{(n-i)} \quad (0 \leq t \leq 1)$$

- Examples:



THE RELATIONSHIPS



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

Let's come back to the Bayes theorem and the Bayesian estimation now.

BAYES THEOREM FOR CONTINUOUS VARIABLE

- Consider N independent observation of a continuous variable X_i , and for a continuous hypothesis θ (for example, a physics parameter like particle mass). The PDF for i th variable is $f_i(X_i | \theta)$. The joint density function is

$$p(X|\theta) = \prod_{i=1}^N f_i(X_i|\theta)$$

- Question: having made N observations from the distributions $f_i(X_i | \theta)$, what can one say about the value of θ ?
- Answer: **classically θ has a fixed true value**. So in principle when fits (for example, maximum likelihood fits, will be discussed in the next lecture) applied, the value of θ can be estimated. But **this is not performed with Bayes theorem**.

BAYES THEOREM FOR CONTINUOUS VARIABLE (II)

- However with Bayesian methods introduced, the distributions of θ (using PDF of θ) can be taken to represent the degree of belief in different possible value of θ .
- We can obtain the form of Bayes theorem used in Bayesian parameter estimation for a particular set of data, X^0 :

$$p(\theta|X^0) = \frac{p(X^0|\theta)p(\theta)}{\int p(X^0|\theta)p(\theta)d\theta}$$

where

- $p(\theta|X^0)$ is posterior probability density for θ .
- $p(X^0|\theta)$ is the likelihood function.
- $p(\theta)$ is the prior probability density for θ . Again this is the major problem in the evaluation!
- The integration in the denominator is just a normalization factor.

PRIORS AND POSTERIOR

- The **prior PDF** represents your personal, subjective, degree of belief about parameter θ before you do any experiments.
 - If you already have some experimental knowledge about θ (e.g. from a previous experiment), the posterior PDF from the previous experiment can be introduced as the prior for the new one.
 - But this implies that, somewhere in the beginning, there must be a prior which contained no experimental evidence!
- **The very first prior** can be thought of as a kind of phase space, or density of possible states of nature. But there is no law of nature that tells us what this density is!
- On the other hand, the **posterior density** already represents all our knowledge about θ , so there is no need to process this PDF any further. But since we want a point estimate here, further operations does require.

BAYESIAN INFERENCE

- The posterior probability is proportional to the product of likelihood function times the prior probability for the unknown parameters θ :

$$p(\theta|X^0) \propto \prod_{i=1}^N f_i(X_i|\theta) \cdot p(\theta)$$

- Based on the **posterior probability** one can evaluate then the average and variance of θ , as well as the point with **highest posterior density (HPD)**!
 - Note the value which gives the highest posterior density and the average don't coincide in general!
- By looking for the highest posterior density point (maximizing the posterior probability), it is just the **maximum likelihood estimator with a flat prior $p(\theta)$** :

$$L(\theta|X^0) \propto \prod_{i=1}^N f_i(X_i|\theta)$$

To be discussed in the next lecture!

BAYESIAN INFERENCE EXAMPLE

- Consider a counting experiment and have been performed for once. **The result is 3**. Assuming the result should follow the **Poisson distribution**. What should be the estimated value of μ ?
- Let's calculate the **posterior probability** with Bayes theorem, assuming a prior $p(\mu)$ and the likelihood function $p(n|\mu)$ is Poisson:

$$p(n|\mu) = \frac{\mu^n e^{-\mu}}{n!} \Rightarrow p(\mu|n) = \frac{\frac{\mu^n e^{-\mu}}{n!} \cdot p(\mu)}{\int_0^\infty \frac{\mu^n e^{-\mu}}{n!} p(\mu) d\mu}$$

- If the prior is uniform, the normalization calculation is basically straightforward:

$$\int_0^\infty \frac{\mu^n e^{-\mu}}{n!} p(\mu) d\mu = 1 \Rightarrow p(\mu|n) = \frac{\mu^n e^{-\mu}}{n!}$$

Wait, is this just the Poisson distribution?

BAYESIAN INFERENCE

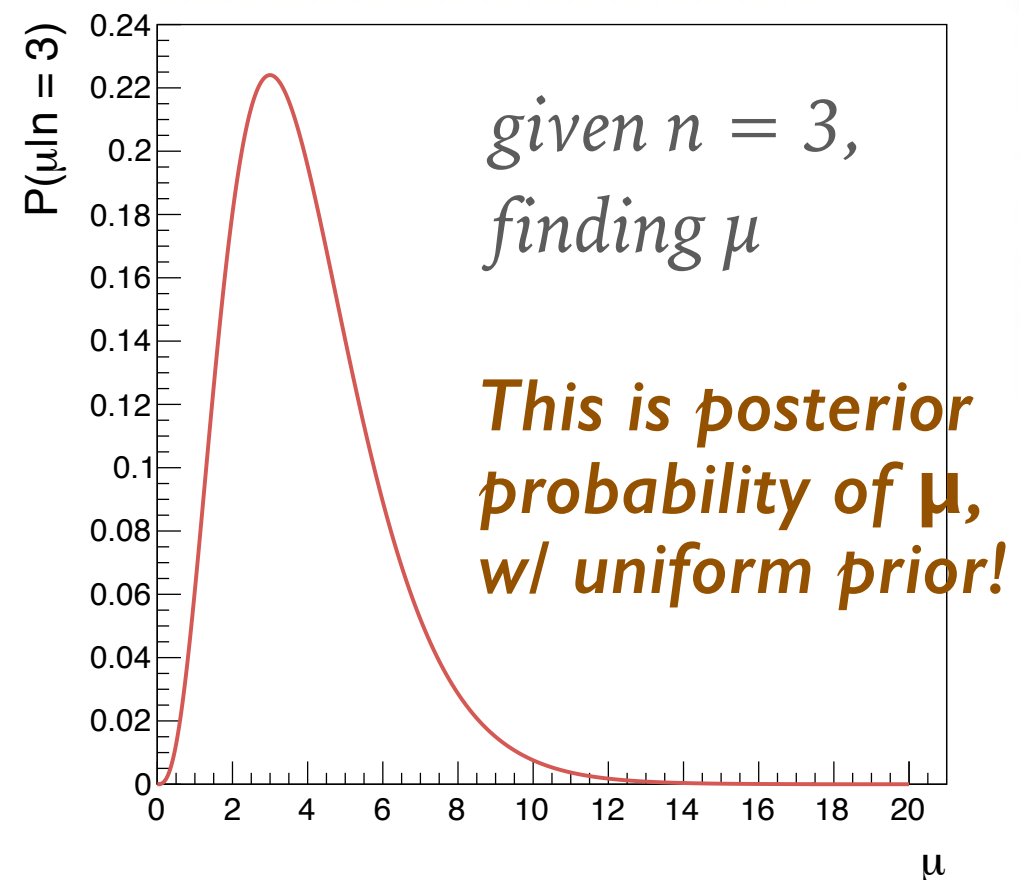
EXAMPLE (II)

■ Poisson PDF versus posterior probability?

$$p(n|\mu) = \frac{\mu^n e^{-\mu}}{n!}$$



$$p(\mu|n) = \frac{\mu^n e^{-\mu}}{n!}$$



Here the value $\mu=3$ has the **highest posterior density**, by maximizing $p(\mu|n)$. **But the mean is not 3 but 4!**

BAYESIAN INFERENCE

EXAMPLE (III)

- If we repeat the experiment again and obtain the second result: **n = 5**, what would be the **posterior probability**?
- Similar the expression can be given by

$$p(\mu|n = 5) = \frac{\frac{\mu^5 e^{-\mu}}{5!} \cdot p(\mu)}{\int_0^\infty \frac{\mu^5 e^{-\mu}}{5!} p(\mu) d\mu}$$

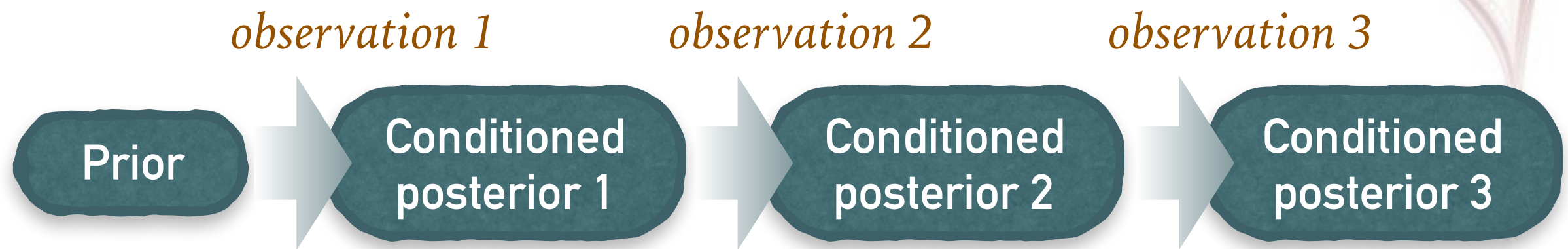
- But now the prior **p(μ)** is not “flat” anymore. Since we already have the first experiment, we can use the posterior probability of the first experiment and the prior of the second experiment.
- Regardless of the normalization term, the resulting posterior should be

$$p(\mu|n = 5) \propto \frac{\mu^5 e^{-\mu}}{5!} \cdot \frac{\mu^3 e^{-\mu}}{3!}$$

Here the value **μ=4** has the **highest posterior density**, by maximizing **p(μ|n)**.

REPEATED USE OF BAYES THEOREM

- Bayes theorem can be applied sequentially for repeated data observations: **posterior** $\Leftrightarrow$ **learning from experiments**.



$$P_0 = \text{Prior} \quad P_1 \propto P_0 \times L_1 \quad P_2 \propto P_0 \times L_1 \times L_2 \quad P_3 \propto P_0 \times L_1 \times L_2 \times L_3$$

...accumulating more and more observations = **multiply probabilities**

- The observation modifies the prior knowledge of the unknown parameters as if L is a probability distribution function.
- Note applying Bayes theorem directly from prior to multiple observables leads to the same result:

$$P_{1+2+3} = P_0 \times L_{1+2+3} = P_0 \times L_1 \times L_2 \times L_3 = P_3$$

COMMENT: BAYESIAN INFERENCE

- Bayesian point estimation is a coherent method which provides a reasonable way to estimate parameters. But it involves two arbitrary choices (*issues*):
 - Which prior PDF to use, and how sensitive is the result to the choice? (*e.g. taking “flat”, some theory function, or anything else*)?
 - How to connect the posterior probability to the point estimate (*e.g. taking the HPD, mean, or anything else*)?
 - **Very difficult to work on multiple dimensional cases.**
- Well, one can increase the observations, **the prior probability is significantly modified by data** — then the final posterior probability will depend much less from the initial prior probability.
 - But under such a condition, using frequentist or Bayesian approaches does not make much difference.

TOWARD THE NEXT LECTURE

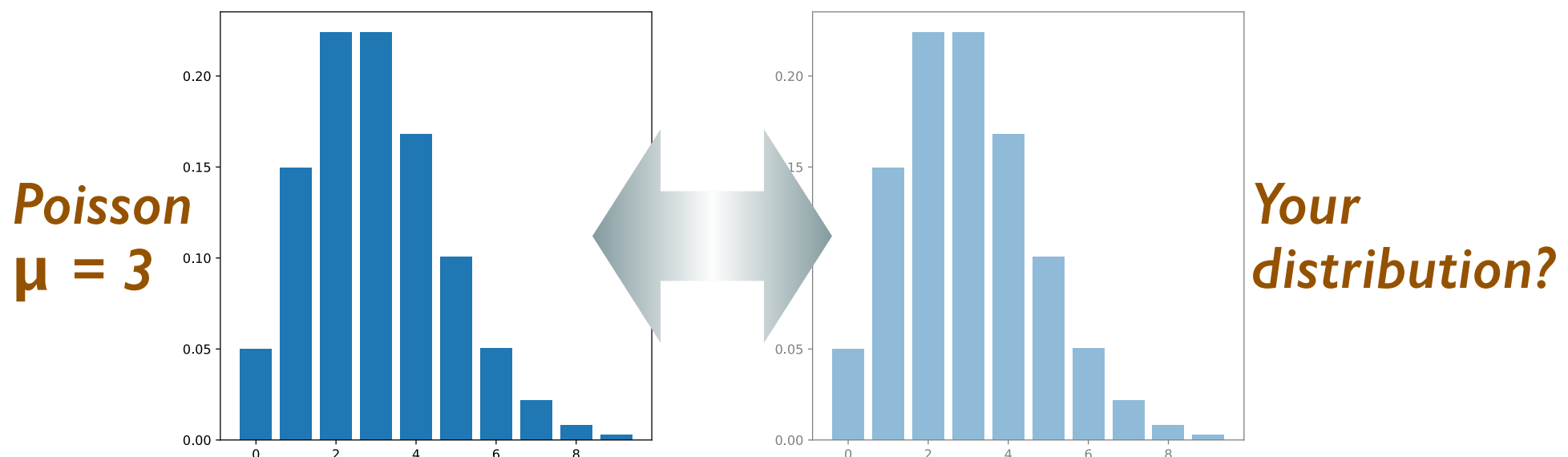
- In this lecture we have discussed the basic concept of Bayesian probability and Frequentist probability, and some commonly used probability distributions.
- Before ending we just discussed about how to use the Bayes theorem to estimate the parameter of a given model. We have shown the principle but no real application yet (no example code!)
- This is because in many of the cases we are facing a more complex problem of many experiment data and complex models.
- In the next lecture we will introduction how to perform the parameter estimation with many data points: **the least square estimator** and **the (extend) maximum likelihood estimator**.

HANDS-ON SESSION

■ Practice 01:

In the `l304-example-03.py` code we have generated a binomial distribution from the “principle”. Now we can use it to approach a Poisson distribution by setting a very large N and very small p , e.g. $N = 10000$ and $p = 0.0003$.

- Compare your resulting distribution and the Poisson distribution of $\mu = Np = 3$ directly (you can take it from `l304-example-02.py`), do you observe a good agreement?



HANDS-ON SESSION

■ Practice 02:

Again, following the discussion before, in a Poisson process the time / space between two defined success would distribute exponentially.

- Suppose you are playing a phone / tablet game which has a character lottery. **Assuming the chance to get your target character is 5%**, perform this lottery with the random numbers for many times and record the number of failures (*not getting your target 5% character*) between two success (*getting your character*).
- Plot the number of failures and show that it is actually **exponentially** distributed!