Berkeley
MSSE
MOLECULAR SCIENCE & SOFTWARE ENGINEERING

Lecture 7:

Random Variables and Distributions

Markus Hohle

University California, Berkeley

Numerical Methods for Computational Science

Course Map

Week 1: Introduction to Scientific Computing and Python Libraries

Week 2: Linear Algebra Fundamentals

Week 3: Vector Calculus

Week 4: Numerical Differentiation and Integration

Week 5: Solving Nonlinear Equations

Week 6: Probability Theory Basics

Week 7: Random Variables and Distributions

Week 8: Statistics for Data Science

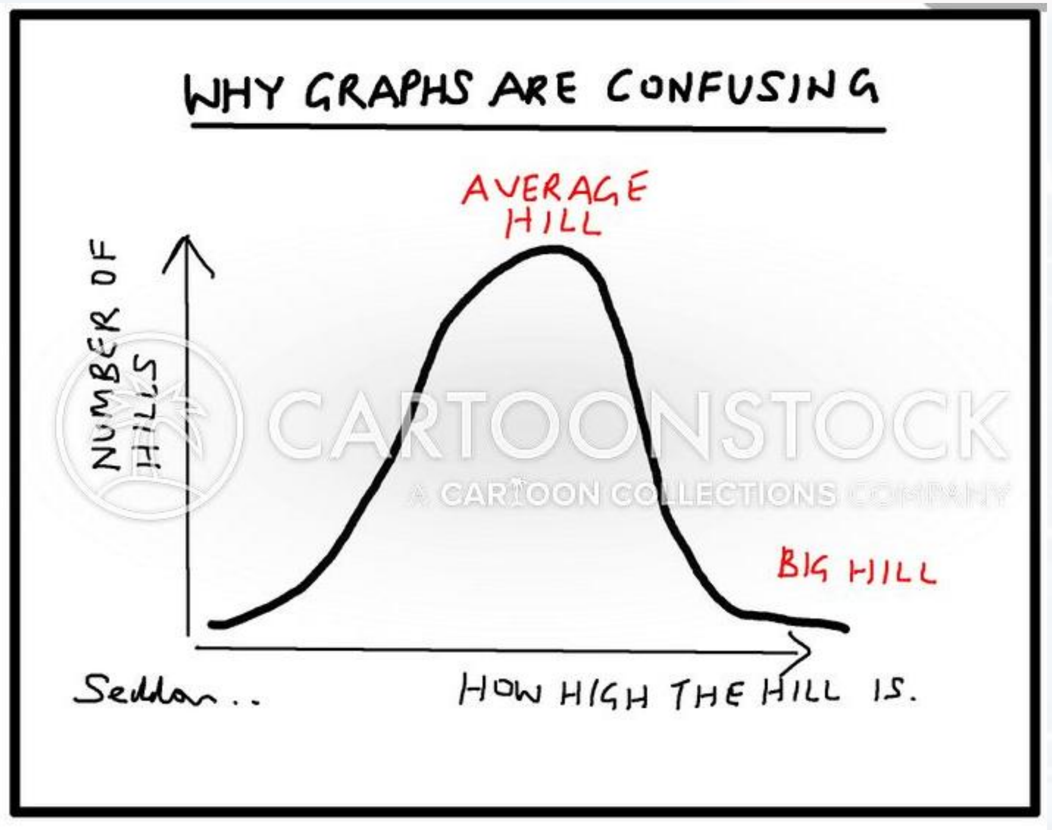
Week 9: Eigenvalues and Eigenvectors

Week 10: Simulation and Monte Carlo Method

Week 11: Data Fitting and Regression

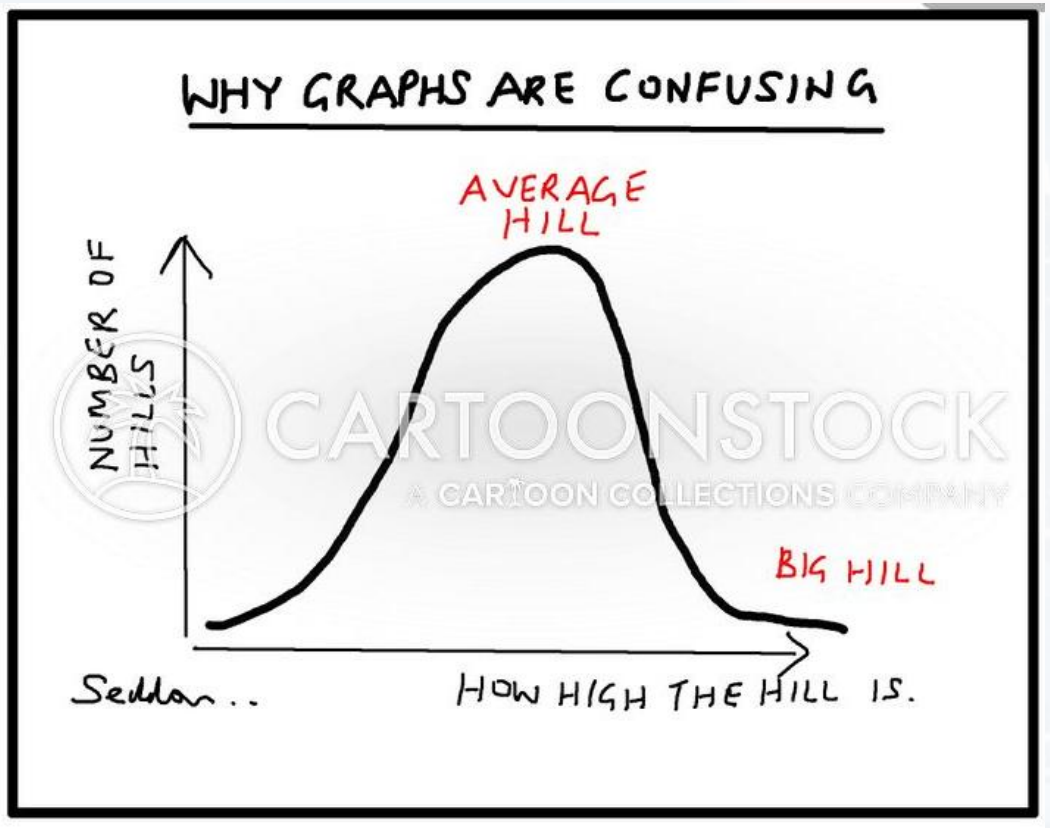
Week 12: Optimization Techniques

Week 13: Machine Learning Fundamentals



<u>Outline</u>

- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Central Limit Theorem



Outline

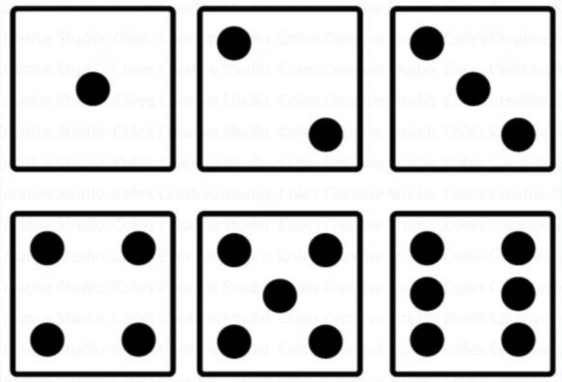
- **Uniform Distribution**
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Central Limit Theorem

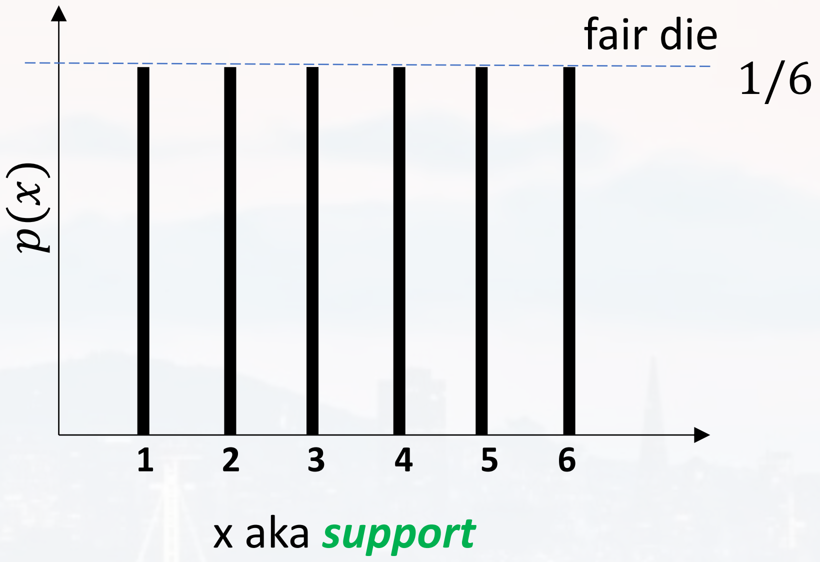
Random Variables and Distributions:

Uniform Distribution

distributions

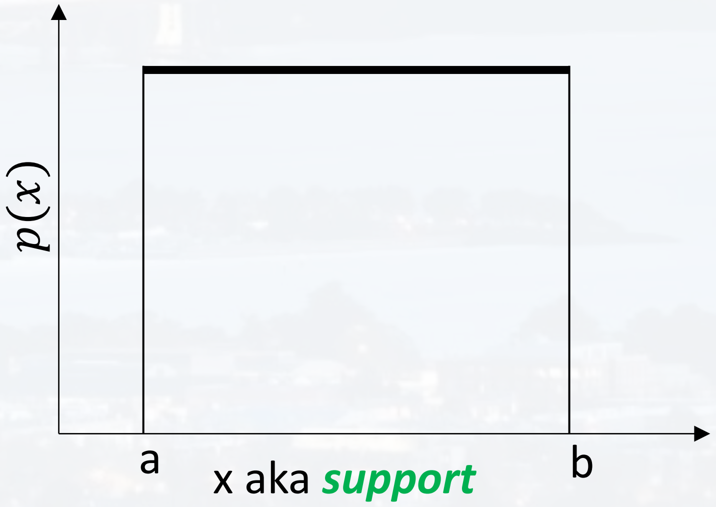
discrete (= countable)





continuous

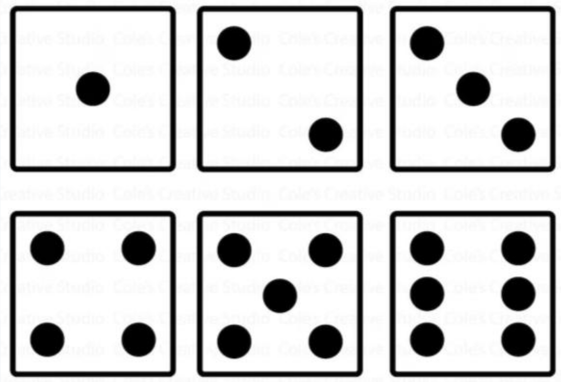
[a ≤ x ≤ b]

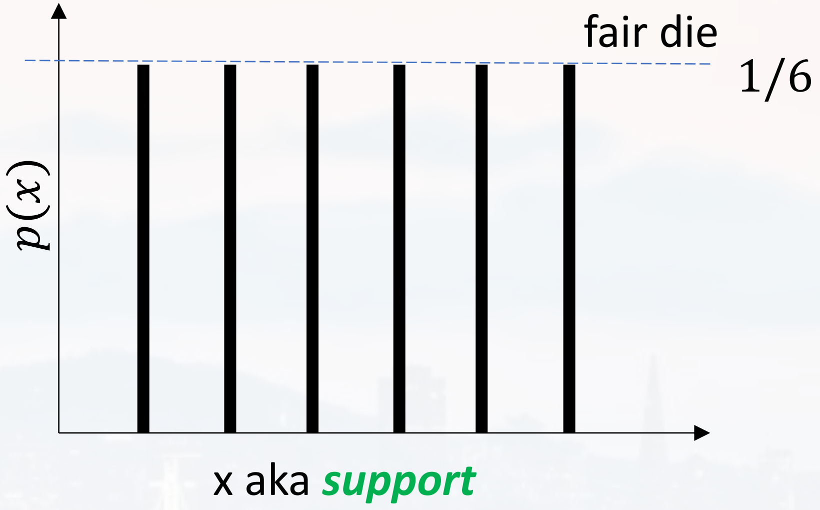


Random Variables and Distributions: Uniform Distribution

distributions

discrete (= countable)





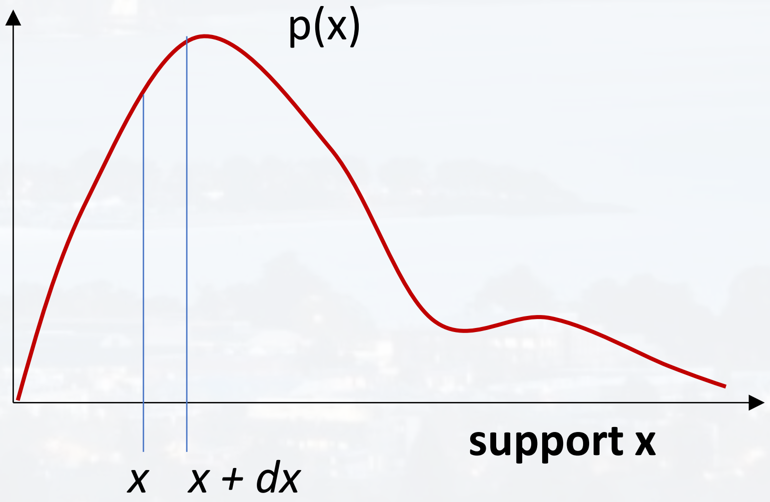
continuous

[a ≤ x ≤ b]

p(x) doesn't make sense

→ *p(x) dx*

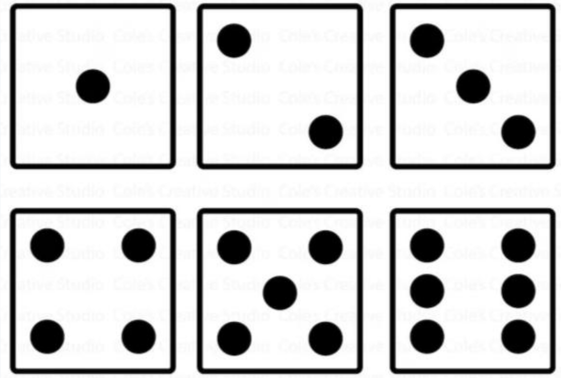
probability **d**ensity **f**unction

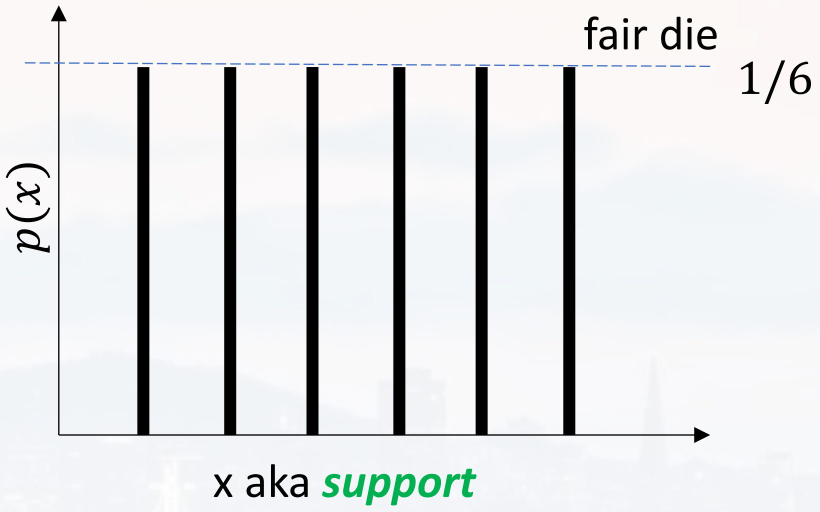


Random Variables and Distributions: Uniform Distribution

<u>distributions</u>

discrete (= countable)





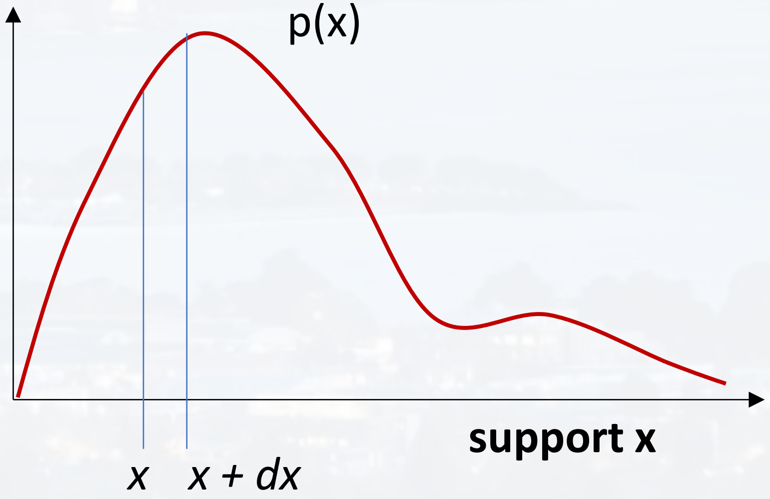
continuous

p(x) doesn't make sense
→ *p(x) dx*

probability **d**ensity **f**unction

dx defines the probability!

*What is the probability to find a person who is **EXACTLY** 6ft tall?*
*Depends on how **accurate** (dx) you measure!*



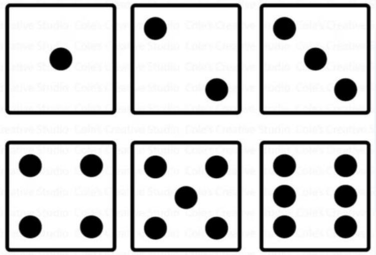
	the mean μ	the variance σ^2
	(barycenter)	(natural scatter)
discrete (= countable)	$\mu = E(x) = \sum_i x_i \, p(x_i)$	$\sigma^2 = var(x) = \sum_i (x_i - \mu)^2 \, p(x_i)$
continuous	$\mu = E(x) = \int x \, p(x) \, dx$	$\sigma^2 = var(x) = \int (x - \mu)^2 \, p(x) \, dx$

Random Variables and Distributions: Uniform Distribution

<u>uniform distribution: discrete vs continuous</u>

the mean μ

discrete (= countable)



$$\mu = \sum_i x_i\, p(x_i) = \sum_{i=1}^{6} i\,\frac{1}{6} = \mathbf{3.5}$$

continuous

[a ≤ x ≤ b]

$$\mu = \int x\, p(x)\, dx$$

$\boldsymbol{p(x) = const}$
(uniform)

$$= const \int_a^b x\, dx = const\,\frac{1}{2}(b^2 - a^2)$$

2nd axiom

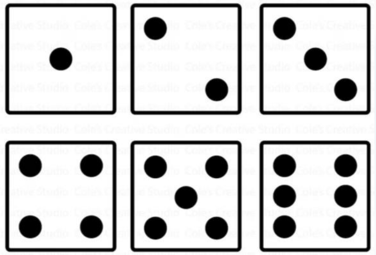
$$\int_a^b p(x)\, dx = 1$$

$$const \int_a^b dx = 1 \qquad \rightarrow const = \frac{1}{b-a}$$

uniform distribution: discrete vs continuous

the mean μ

discrete (= countable)



$$\mu = \sum_i x_i\, p(x_i) = \sum_{i=1}^{6} i\, \frac{1}{6} = 3.5$$

continuous

[a ≤ x ≤ b]

$$\mu = \int x\, p(x)\, dx = const\, \frac{1}{2}(b^2 - a^2)$$

$$const = \frac{1}{b-a}$$

$$\boxed{\mu = \frac{1}{2}\frac{b^2 - a^2}{b-a}}$$

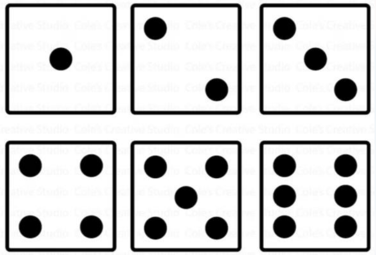
= 3.5 for a = 1 and b = 6

Random Variables and Distributions: Uniform Distribution

<u>uniform distribution: discrete vs continuous</u>

the variance σ^2

discrete (= countable)



$$\sigma^2 = \sum_i (x_i - \mu)^2 \, p(x_i) = \frac{1}{6}\sum_{i=1}^{6}(i - 3.5)^2 \approx \mathbf{2.9}$$

continuous

[a ≤ x ≤ b]

$$\sigma^2 = \int (x - \mu)^2 \, p(x) \, dx = \frac{1}{b-a}\int_a^b (x - 3.5)^2 dx$$

$$\boxed{\sigma^2 = \frac{1}{12}\,(b-a)^2}$$

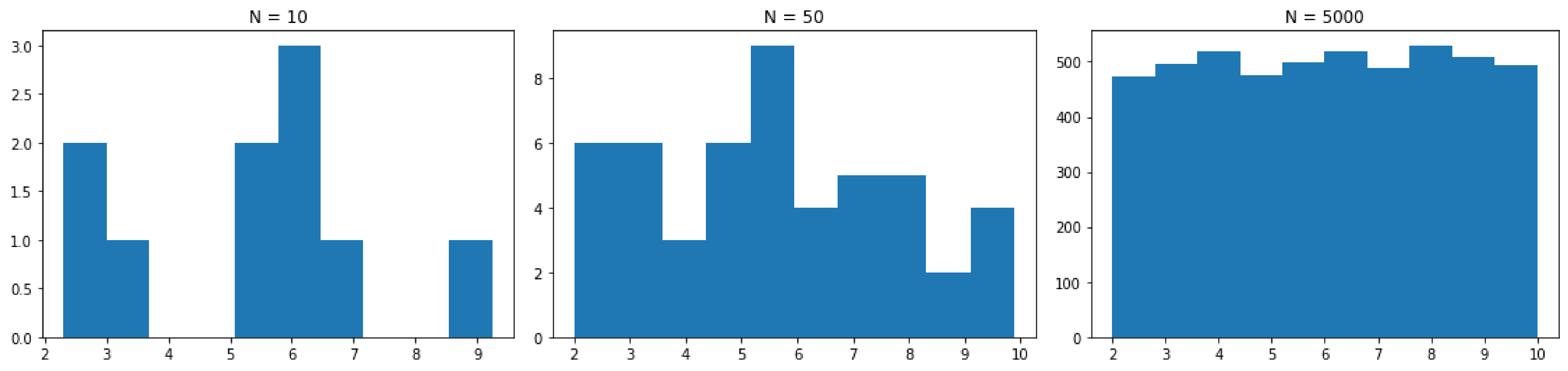
= 25/12 ≈ **2.1** for a = 1 and b = 6

$$p(x) = const$$

plotting the pdf

```
U = np.random.uniform(low, high, shape)
plt.hist(U)
```

continuous support



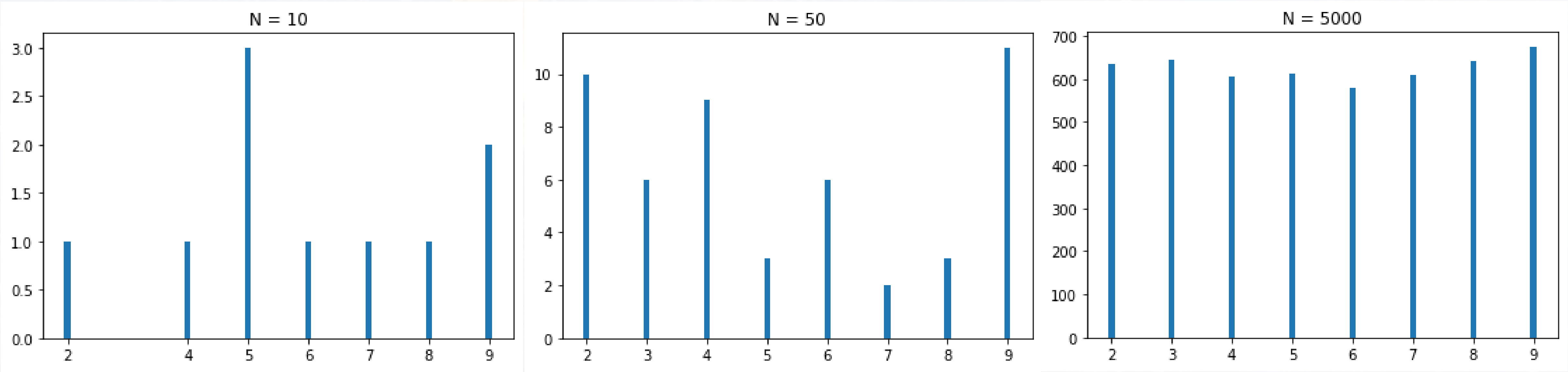
Random Variables and Distributions:

$p(x) = const$

plotting the pdf

```
U = np.random.randint(low, high, shape)

labels, counts = np.unique(U, return_counts = True)
plt.bar(labels, counts, align = 'center', width = 0.1)
plt.gca().set_xticks(labels)
plt.title('N = ' + str(N))
```

discrete support



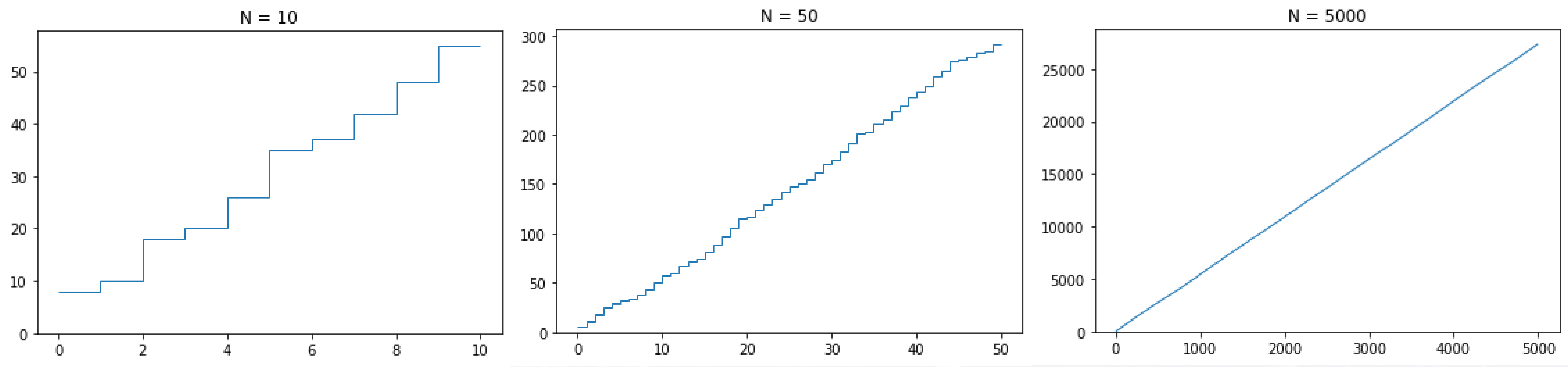
Cumulative **d**ensity **f**unction

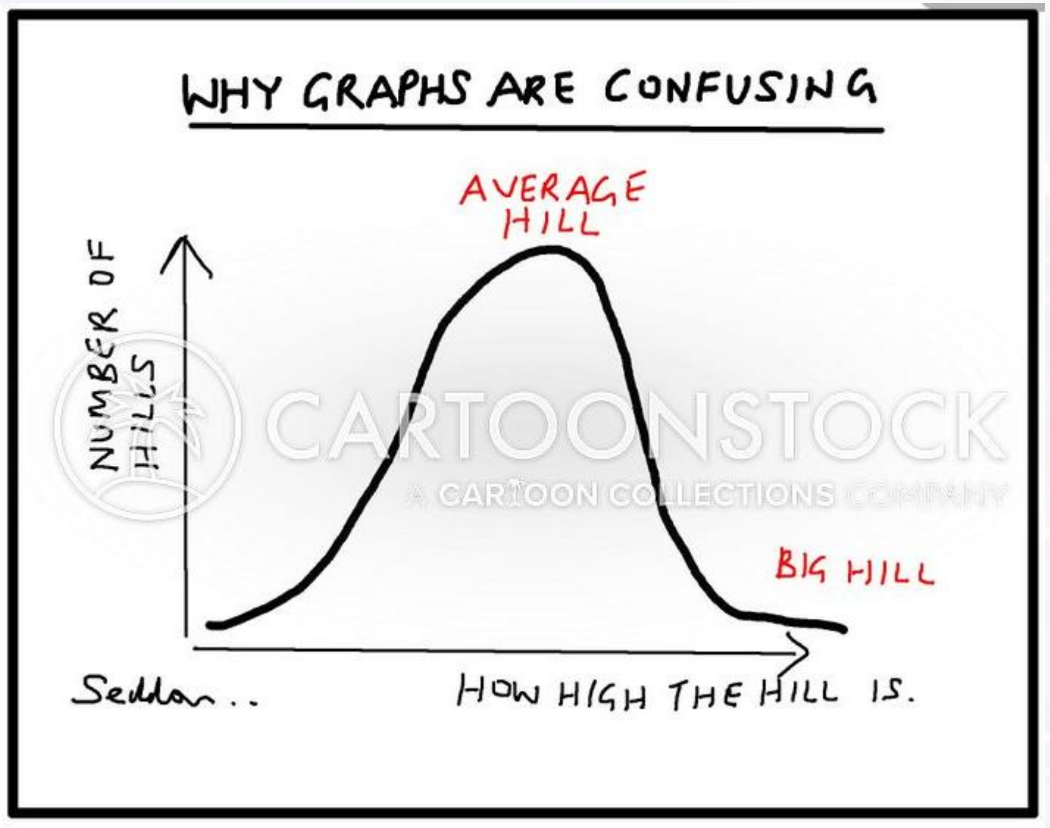
$$P(x) = \int_a^x p(x)\,dx$$

```
U = np.random.randint(low, high, shape)
```

discrete support

```
C = np.cumsum(U)
plt.stairs(C, baseline = None)
plt.title('N = ' + str(N))
```



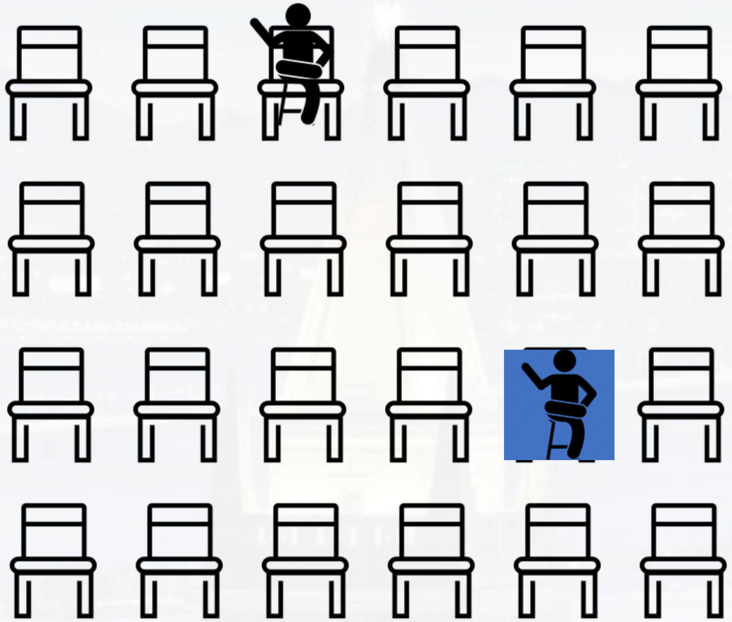


Outline

- Uniform Distribution
- **Binomial Distribution**
- Poisson Distribution
- Normal Distribution
- Central Limit Theorem

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

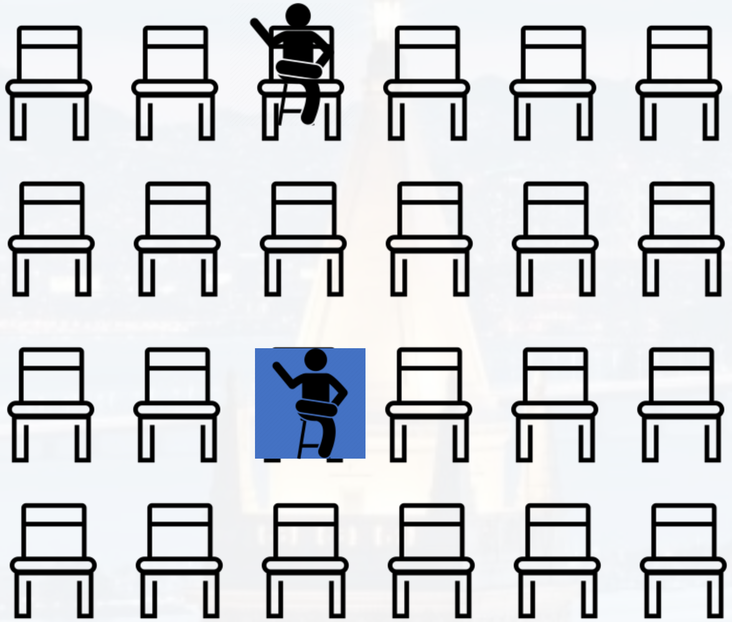


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

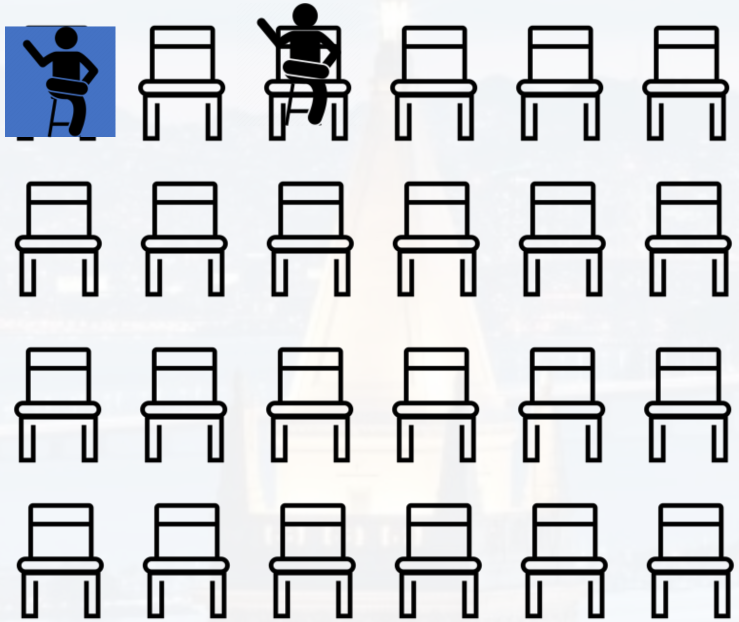


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

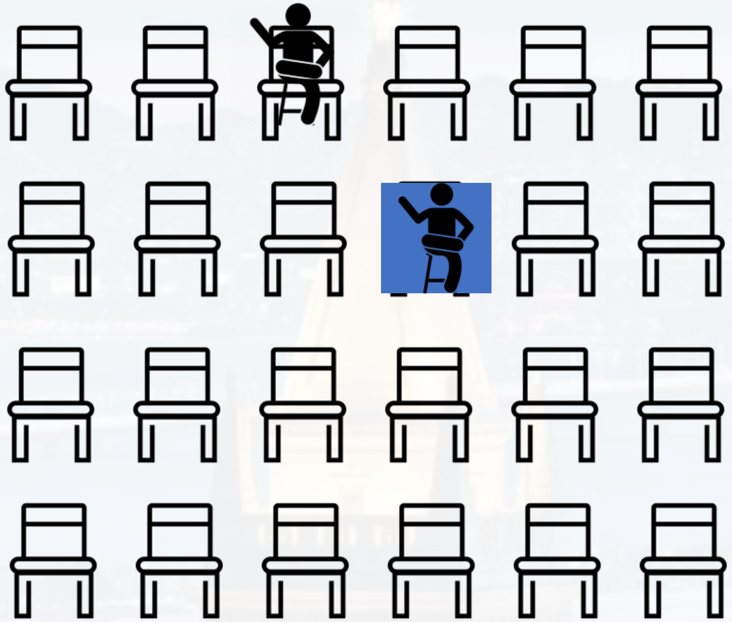


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

<u>seating arrangements in a classroom</u>

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

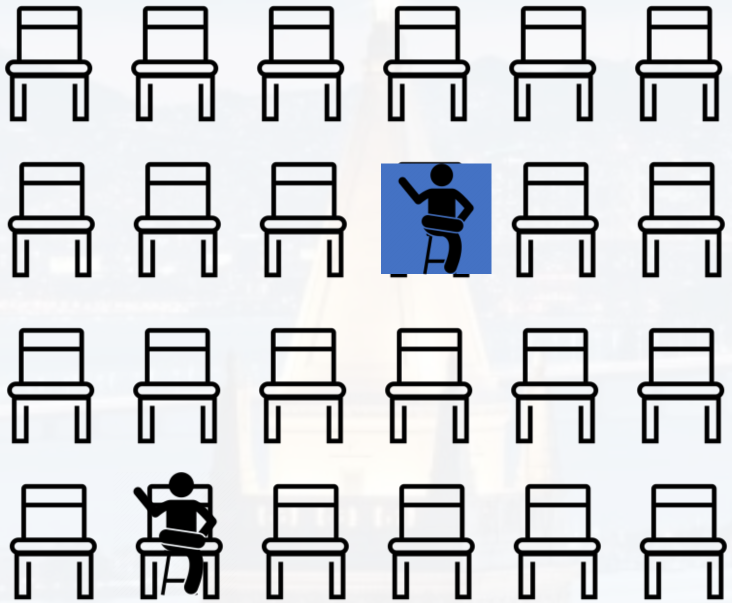


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

<u>seating arrangements in a classroom</u>

- How many arrangements *Ω* for *k* students and *n* seats?
- How many arrangements *Ω* for *k* occupied seats among *n* seats?

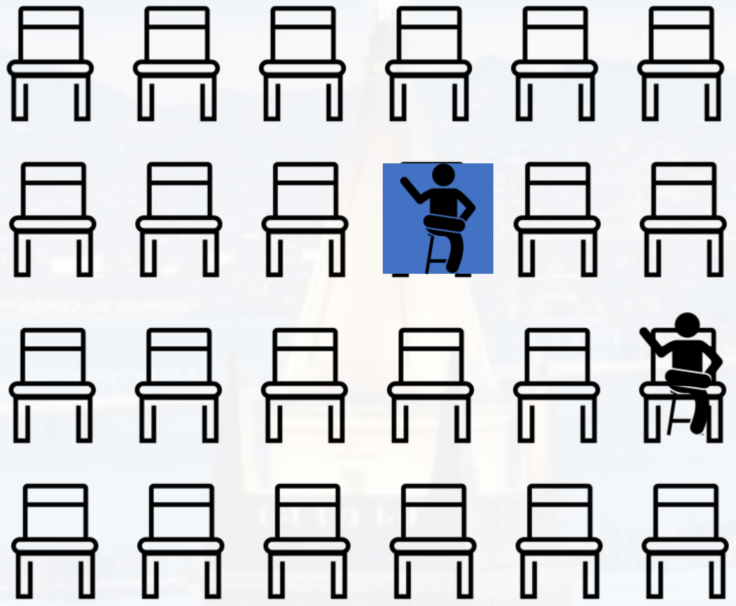


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

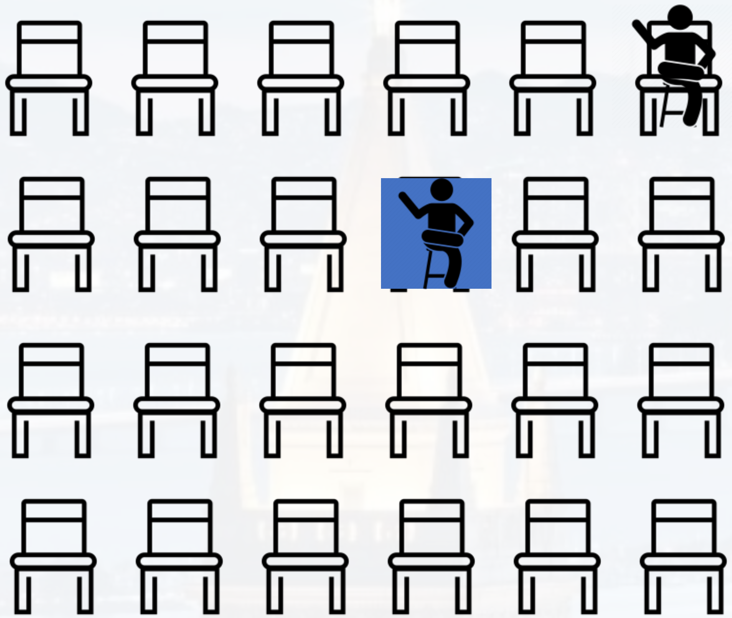


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?

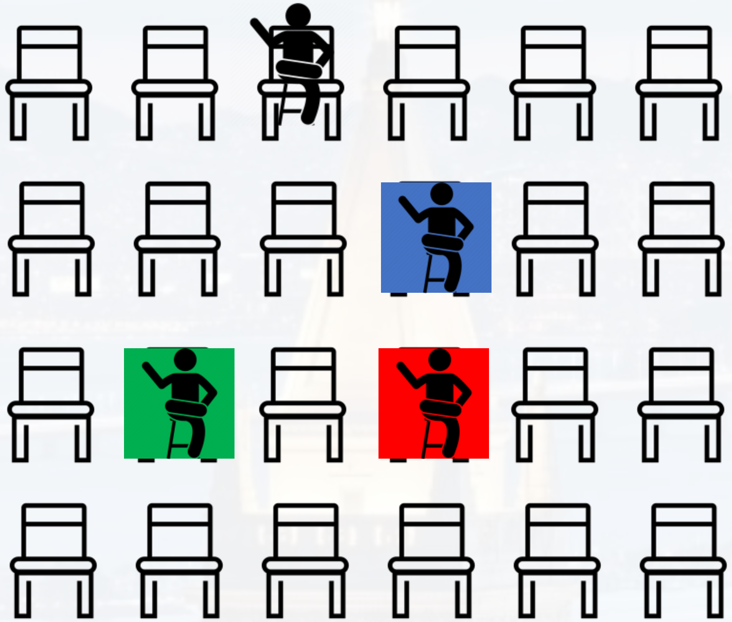


student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

seating arrangements in a classroom

- How many arrangements *Ω* for *k* students and *n* seats?
- How many arrangements *Ω* for *k* occupied seats among *n* seats?



student 1: $\Omega = n$

student 2: $\Omega(2) = n - 1$

$\Omega = n\,(n-1)$

student 3: $\Omega(3) = n - 2$

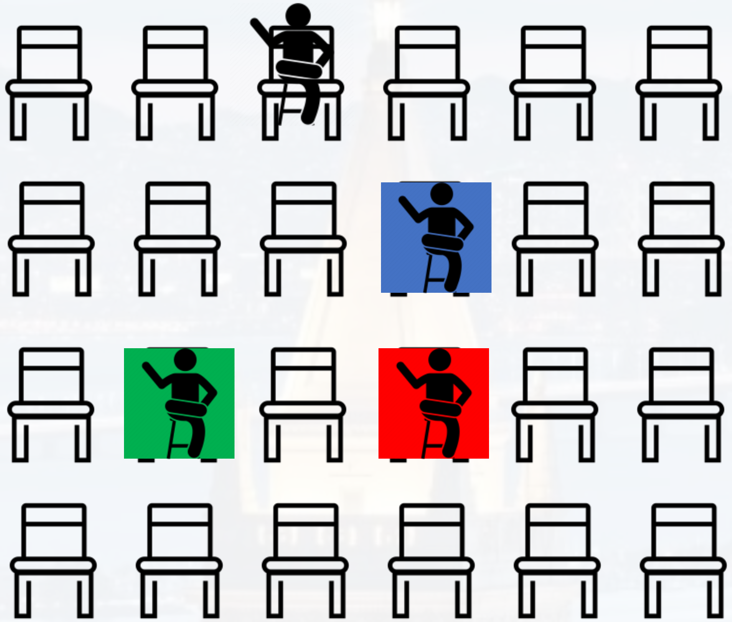
$\Omega = n\,(n-1)(n-2)$

student 4: $\Omega(4) = n - 3$

$\Omega = n\,(n-1)(n-2)(n-3)$

seating arrangements in a classroom

- How many arrangements Ω for k students and n seats?
- How many arrangements Ω for k occupied seats among n seats?



student 1: $\Omega = n$

student 2: $\Omega = n\,(n-1)$

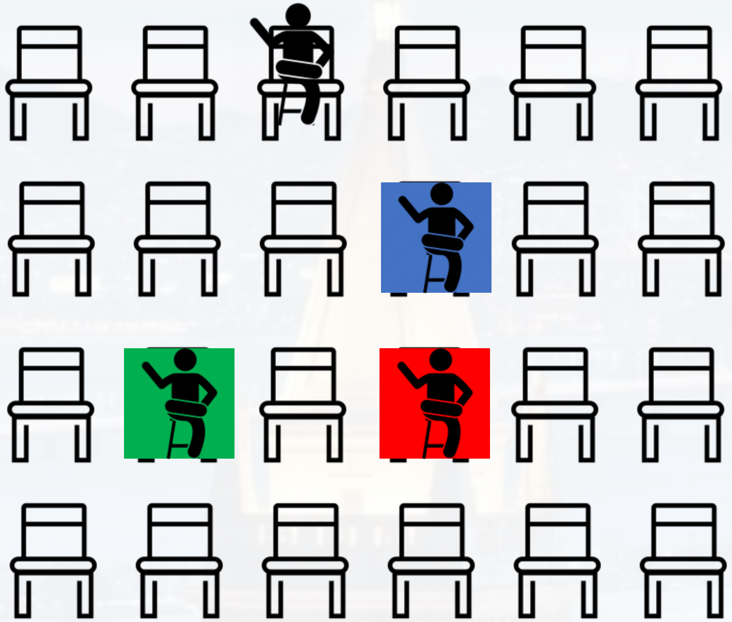
student 3: $\Omega = n\,(n-1)(n-2)$

student 4: $\Omega = n\,(n-1)(n-2)(n-3)$

student k: $\Omega = n\,(n-1)(n-2)(n-3)\ldots(n-\mathrm{k}+1)$

<u>seating arrangements in a classroom</u>

- How many arrangements *Ω* for *k* students and *n* seats?
- How many arrangements *Ω* for *k* occupied seats among *n* seats?



student *k*:
$$\Omega = n\,(n-1)(n-2)(n-3)\ldots(n-\mathrm{k}+1)$$

<u>for *k* = *n*:</u>

$$\Omega = k\,(k-1)(k-2)\ldots 1 = k!$$

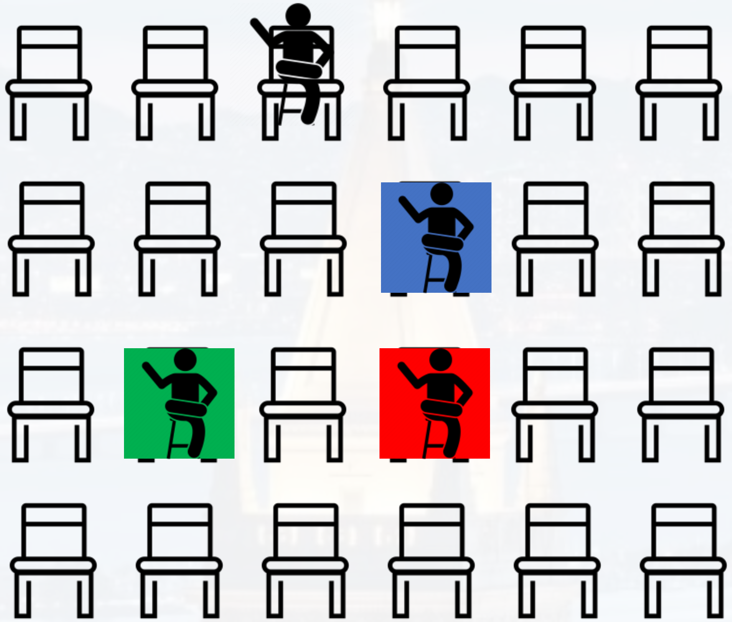
k factorial

5! = 5 * 4 * 3 * 2 *1 = 120

note: 0! = 1

<u>seating arrangements in a classroom</u>

- How many arrangements *Ω* for *k* students and *n* seats?
- How many arrangements *Ω* for *k* occupied seats among *n* seats?



student *k*:
$$\Omega = n\,(n-1)(n-2)(n-3)\ldots(n-k+1)$$

interrupting at *(n-k)!*

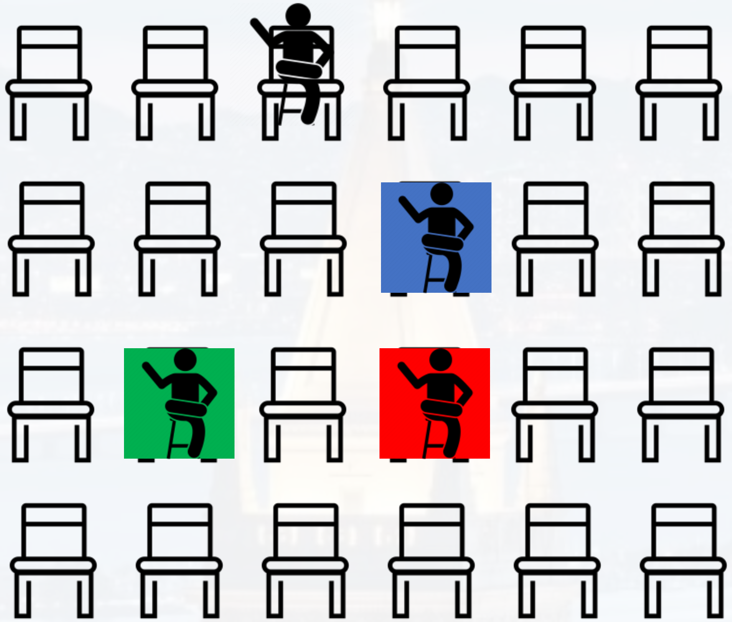
$$\Omega = \frac{n!}{(n-k)!} = \frac{n(n-1)\ldots(n-k+1)(n-k)(n-k-1)\ldots 1}{(n-k)(n-k-1)\ldots 1}$$

$$\boxed{\Omega = \frac{n!}{(n-k)!}}$$

<u>seating arrangements in a classroom</u>

- How many arrangements *Ω* for *k* students and *n* seats?

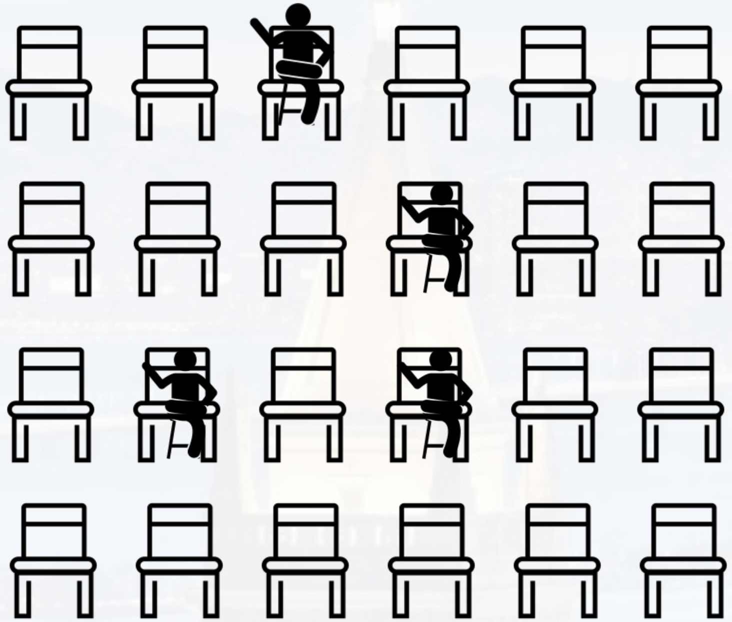
- How many arrangements *Ω* for *k* occupied seats among *n* seats?



$$\boxed{\Omega = \frac{n!}{(n-k)!}}$$

<u>seating arrangements in a classroom</u>

- How many arrangements *Ω* for *k* students and *n* seats?
- **How many arrangements *Ω* for *k* occupied seats among *n* seats?**



$$\boxed{\Omega = \frac{n!}{(n-k)!}}$$

k students on *k* seats

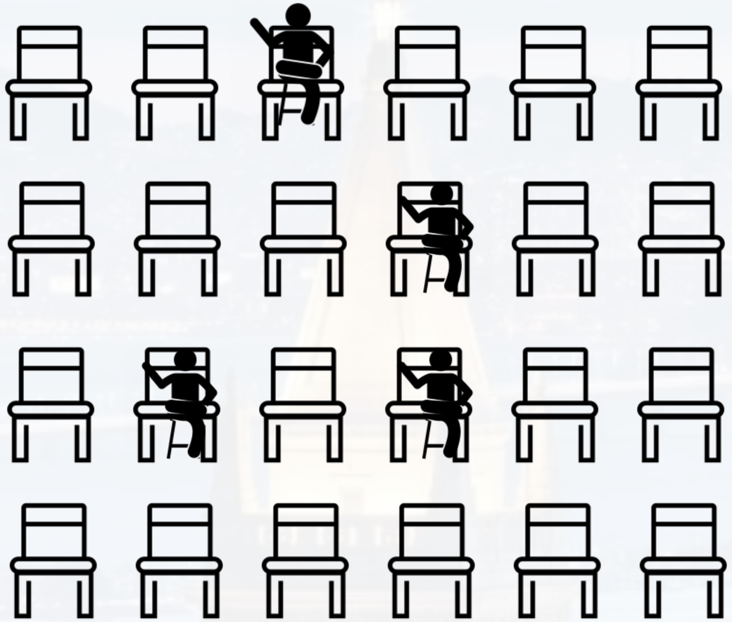
→ *k!* arrangements, which are ***indistinguishable***

$$\boxed{\Omega = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}}$$

n choose k

<u>seating arrangements in a classroom</u>

- How many arrangements Ω for k students and n seats?
- **How many arrangements Ω for k occupied seats among n seats?**



k are ***indistinguishable***

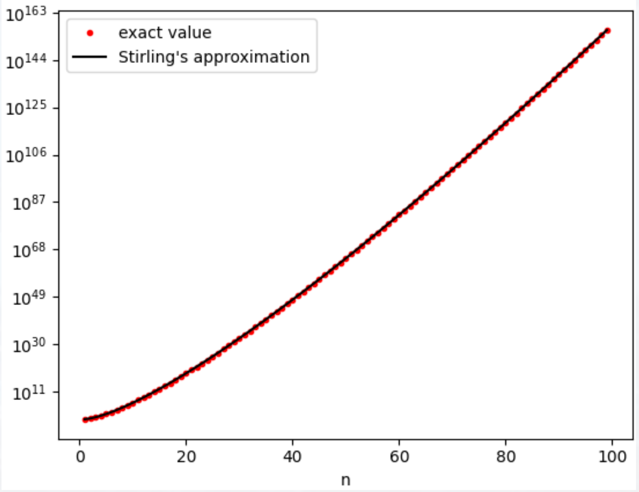
see `nchoosek.ipynb`

$$\boxed{\Omega = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}}$$

n choose k

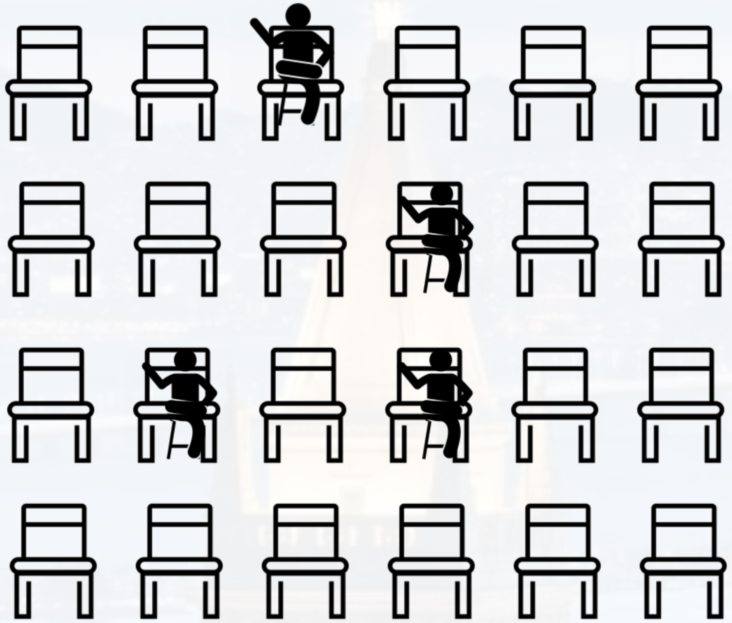
Stirling's approximation

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^{n}$$



seating arrangements in a classroom

- How many arrangements *Ω* for *k* students and *n* seats?
- **How many arrangements *Ω* for *k* occupied seats among *n* seats?**



k are ***indistinguishable***

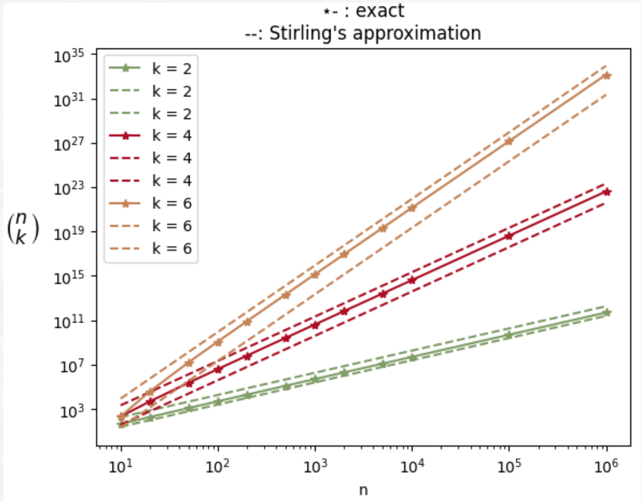
see `nchoosek.ipynb`

$$\boxed{\Omega = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}}$$

n choose k

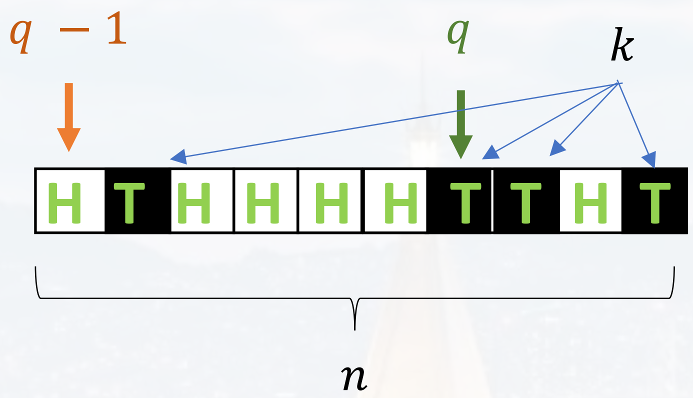
Stirling's approximation

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

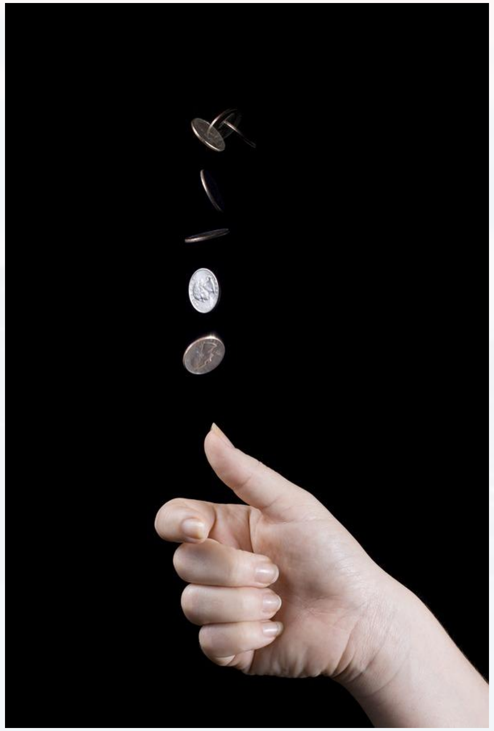


bino = “two”

probability of having **a** sequence of **k tails** and **n - k heads**



fair coin? q = 0.5 ???



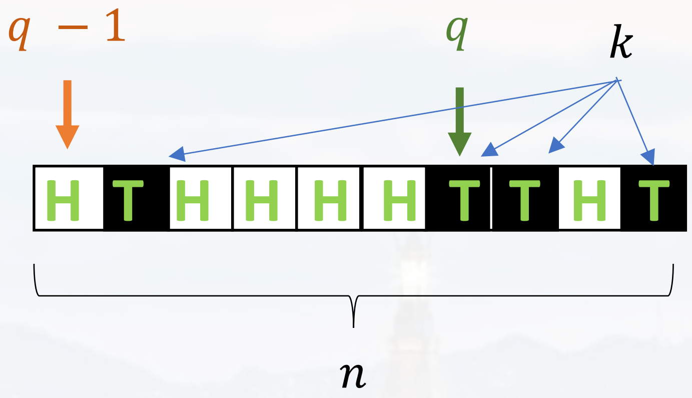
$$p_{tot} = (q-1)q(q-1)(q-1)(q-1)(q-1)qq(q-1)q$$

$$p_{tot} = q^k\,(1-q)^{n-k}$$

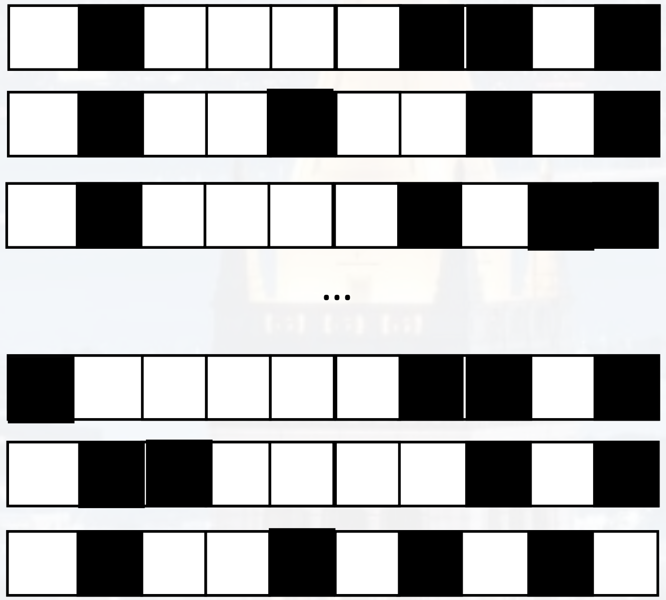
Probability of having any sequence of **k tails** and **n - k heads**?

n choose k
The Binomial Distribution

Probability of having **any** sequence of **k tails** and **n - k heads**?



$$p_{tot} = q^k \, (1-q)^{n-k}$$



$$p_{tot} = q^k \, (1-q)^{n-k}$$

$$p_{tot} = q^k \, (1-q)^{n-k}$$

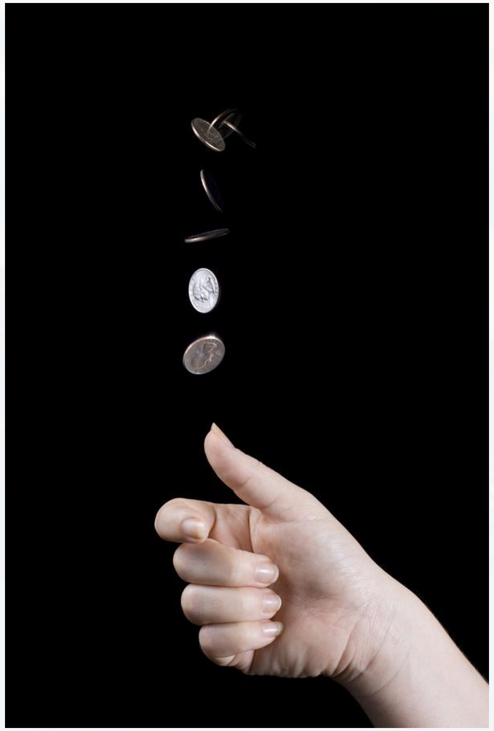
$$p_{tot} = q^k \, (1-q)^{n-k}$$

...

$$p_{tot} = q^k \, (1-q)^{n-k}$$

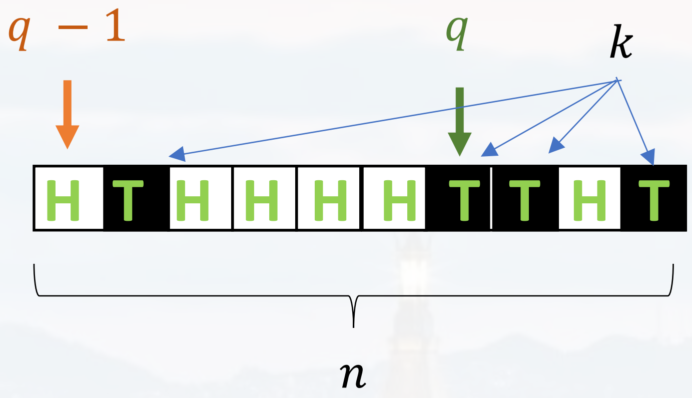
$$p_{tot} = q^k \, (1-q)^{n-k}$$

$$p_{tot} = q^k \, (1-q)^{n-k}$$

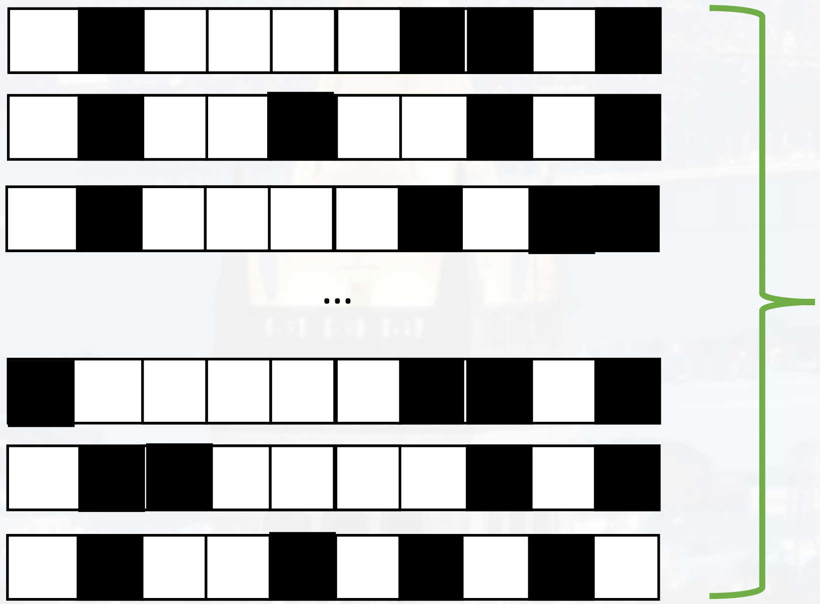


Random Variables and Distributions:

Probability of having **any** sequence of **k tails** and **n - k heads**?



$$p_{tot} = q^k\,(1-q)^{n-k}$$

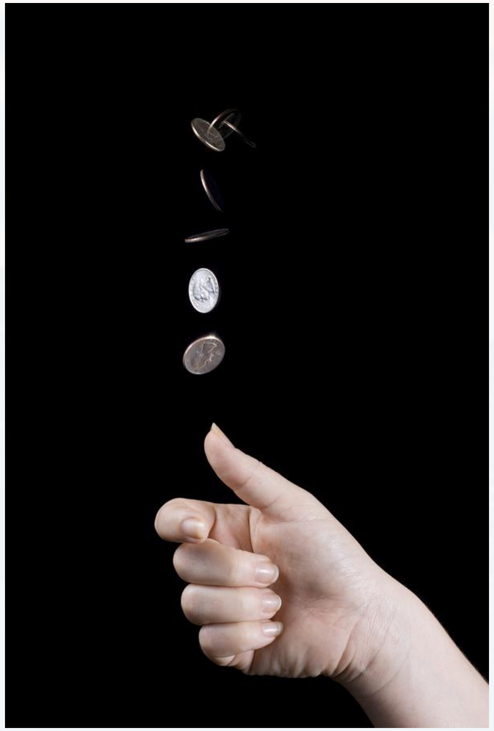


k are ***indistinguishable***

$$\Omega = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}$$

$$\boxed{P(k|q,n) = \binom{n}{k}\,q^k\,(1-q)^{n-k}}$$

binomial distribution

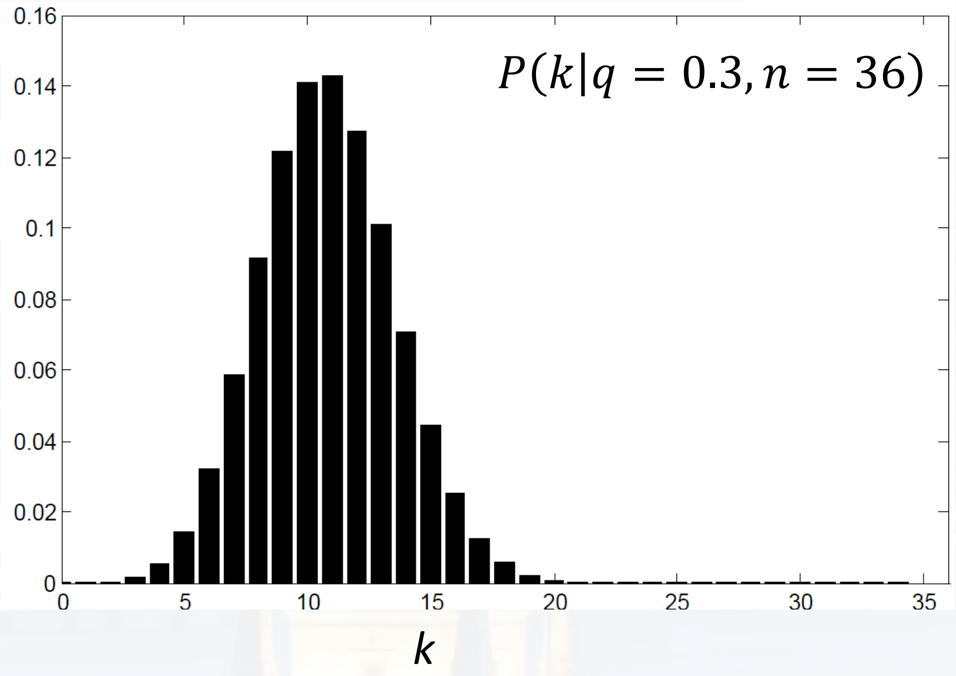


$$P(k|q,n) = \binom{n}{k} q^k (1-q)^{n-k}$$

binomial distribution

$$\mu = \sum_i x_i \, p(x_i)$$

$$\mu = \int x \, p(x) \, dx$$



$$\mu = \sum_{k=0}^{n} k \binom{n}{k} q^k (1-q)^{n-k} = \boldsymbol{qn}$$

$$var(k) = \sum_{k=0}^{n} (k - qn)^2 \binom{n}{k} q^k (1-q)^{n-k} = \boldsymbol{qn(1-q)}$$

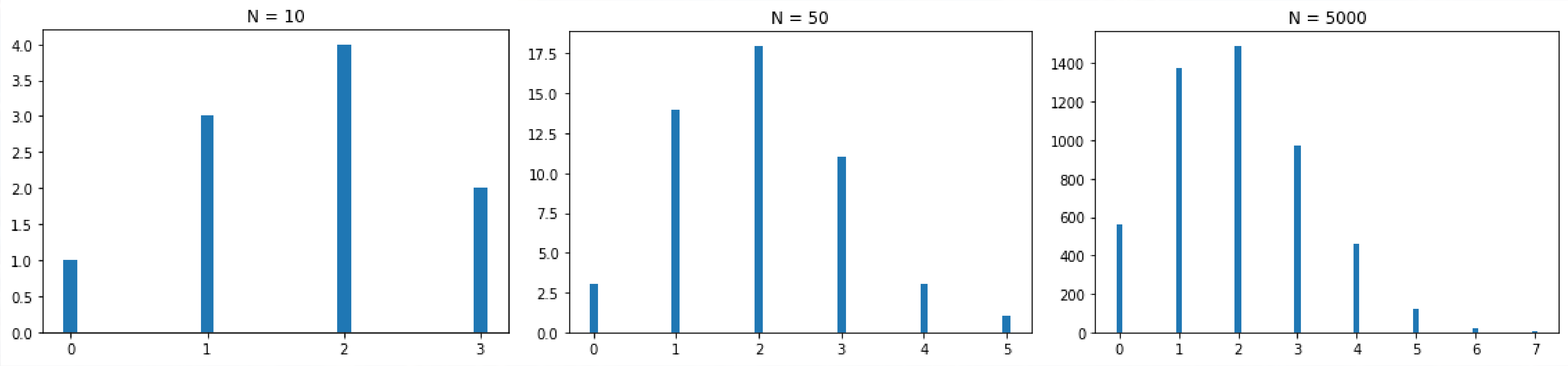
n choose k

The Binomial Distribution

$$\boxed{P(k|q,n) = \binom{n}{k} q^k (1-q)^{n-k}}$$

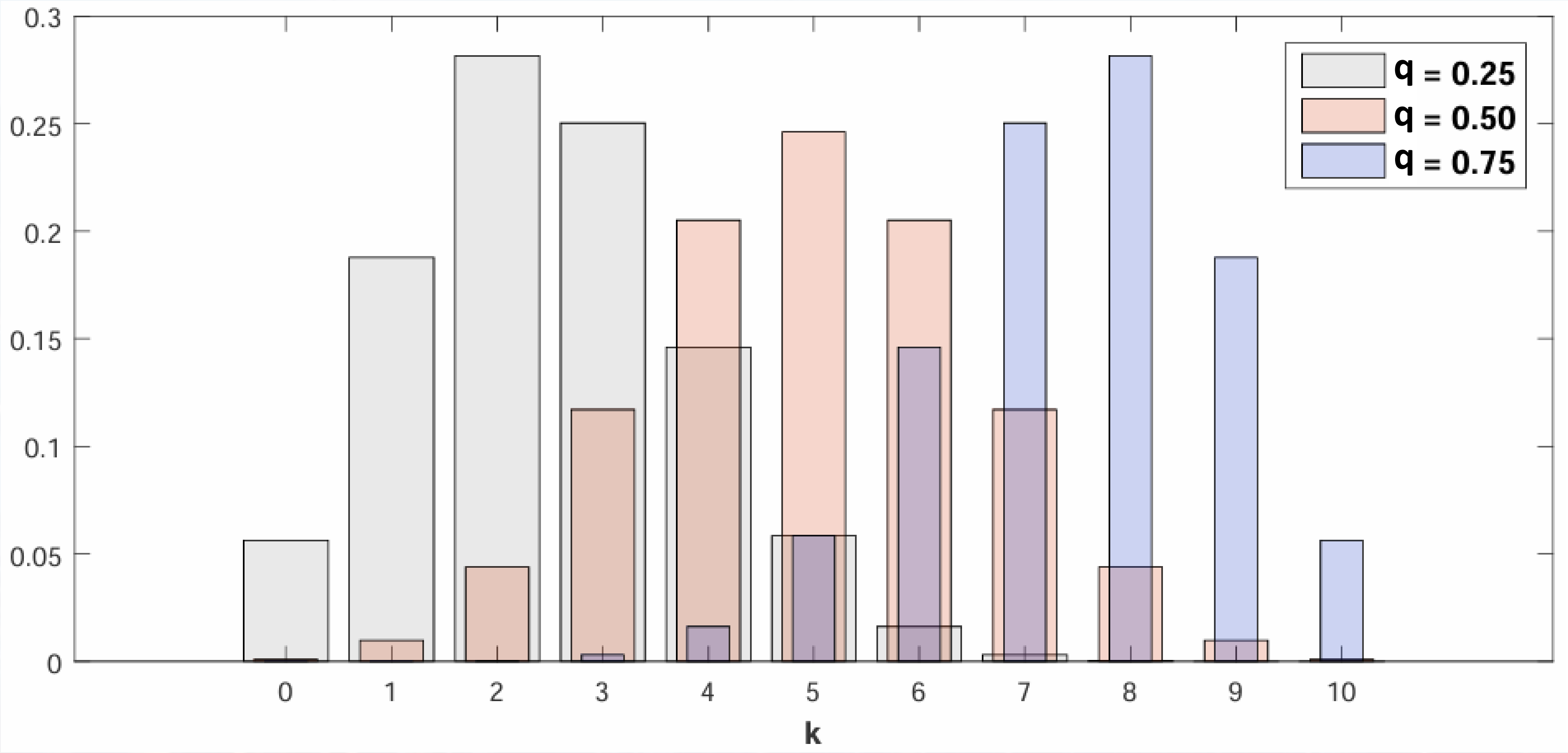
binomial distribution

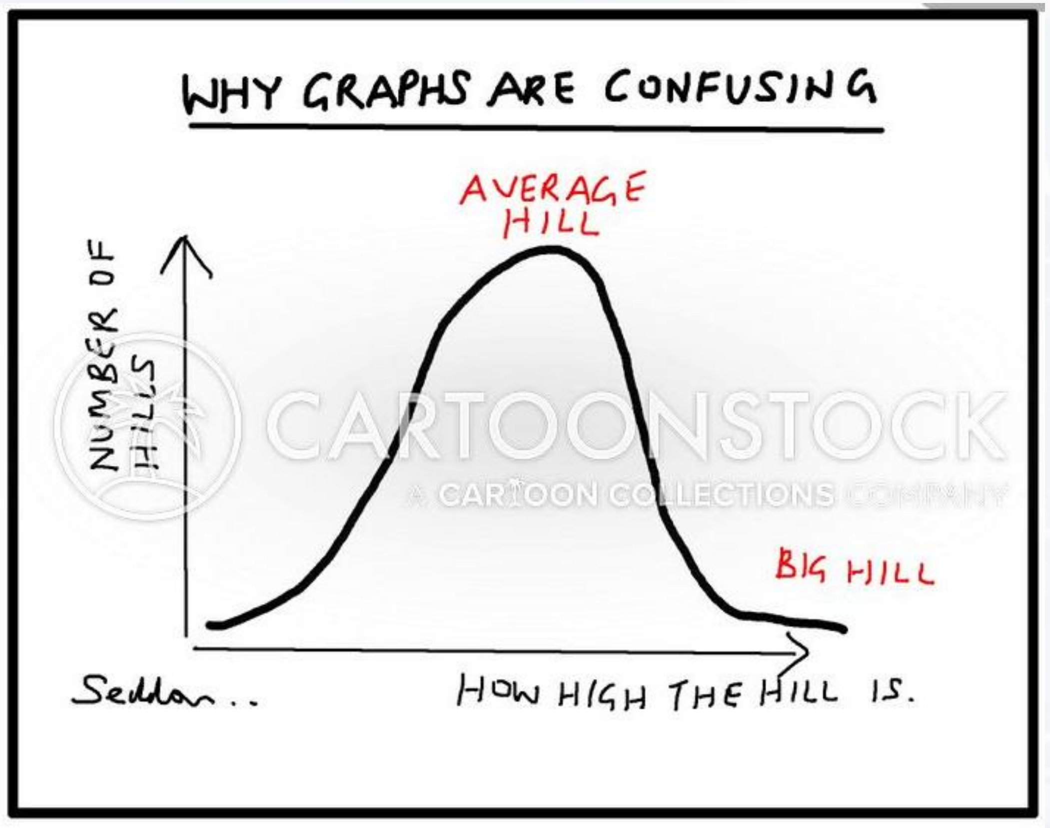
```
q = 0.2
n = 10
K = np.random.binomial(n, q, N)

labels, counts = np.unique(K, return_counts = True)
plt.bar(labels, counts, align = 'center', width = 0.1)
plt.gca().set_xticks(labels)
plt.title('N = ' + str(N))
```



$$\boxed{P(k|q,n) = \binom{n}{k}\, q^k\,(1-q)^{n-k}}$$

binomial distribution





Outline

- Uniform Distribution
- Binomial Distribution
- **Poisson Distribution**
- Normal Distribution
- Central Limit Theorem

$$\boxed{P(k|q,n) = \binom{n}{k}\, q^k\, (1-q)^{n-k}}$$

binomial distribution

rare events → q << 1

Taylor expansion for $(1-q)^{n-k}$ around q = 0

$$(1-q)^{n-k} = 1 - nq + \frac{(nq)^2}{2} - \frac{(nq)^3}{6} + \ldots = e^{-nq}$$

→ n → ∞

Stirling's approximation for n!

$$\frac{n!}{(n-k)!} \approx \sqrt{\frac{n}{n-k}}\,\frac{n^n e^{n-k}}{e^n (n-k)^{n-k}} \approx n^k$$

$$\binom{n}{k}\, q^k\, (1-q)^{n-k} \approx \frac{(nq)^k e^{-nq}}{k!}$$

Random Variables and Distributions:

$$\binom{n}{k} q^k (1-q)^{n-k} \approx \frac{(nq)^k e^{-nq}}{k!}$$

often: nq : = λ

events per time interval: λ = c Δt

H	T	H	H	H	H	T	T	H	T

time

$\underbrace{\qquad\qquad\qquad}_{\Delta t}$

rate c = 4 tails per Δt

$$\boxed{P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}}$$

Poisson distribution

$\mu \quad = qn \quad \rightarrow$ $\boldsymbol{qn = \lambda}$

$var(k) \quad = qn(1-q) \rightarrow$ $\boldsymbol{qn = \lambda}$

Random Variables and Distributions:

$$\boxed{P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}}$$

Poisson distribution

$$\mu = \lambda$$
$$var(k) = \lambda$$

- **rare** events
- events are mutually **independent**
- events have **no duration**

examples:

- radioactive decay
- single photon detection
- lightning
- mutation of a gene
- receiving WhatsApp messages/SMS

rare: not that **a** atom decays,
→ that **this** atom decays within Δt

$$P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}$$

Poisson distribution

$$\mu = \lambda$$

$$var(k) = \lambda$$

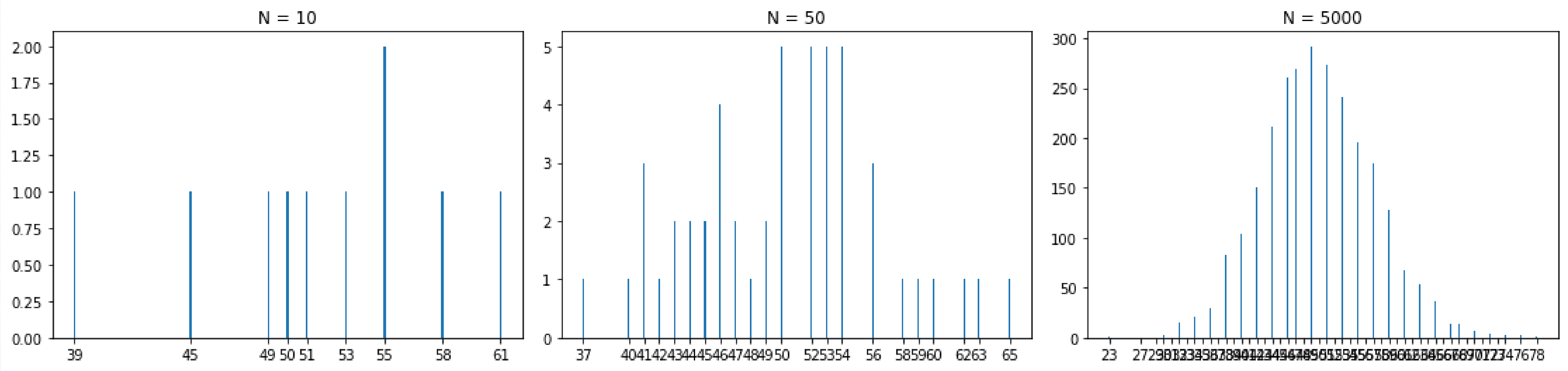
```
c    = 5
delt = 10
lam  = c * delt

K    = np.random.poisson(lam, N)

labels, counts = np.unique(K, return_counts = True)
plt.bar(labels, counts, align = 'center', width = 0.1)
plt.gca().set_xticks(labels)
plt.title('N = ' + str(N))
```

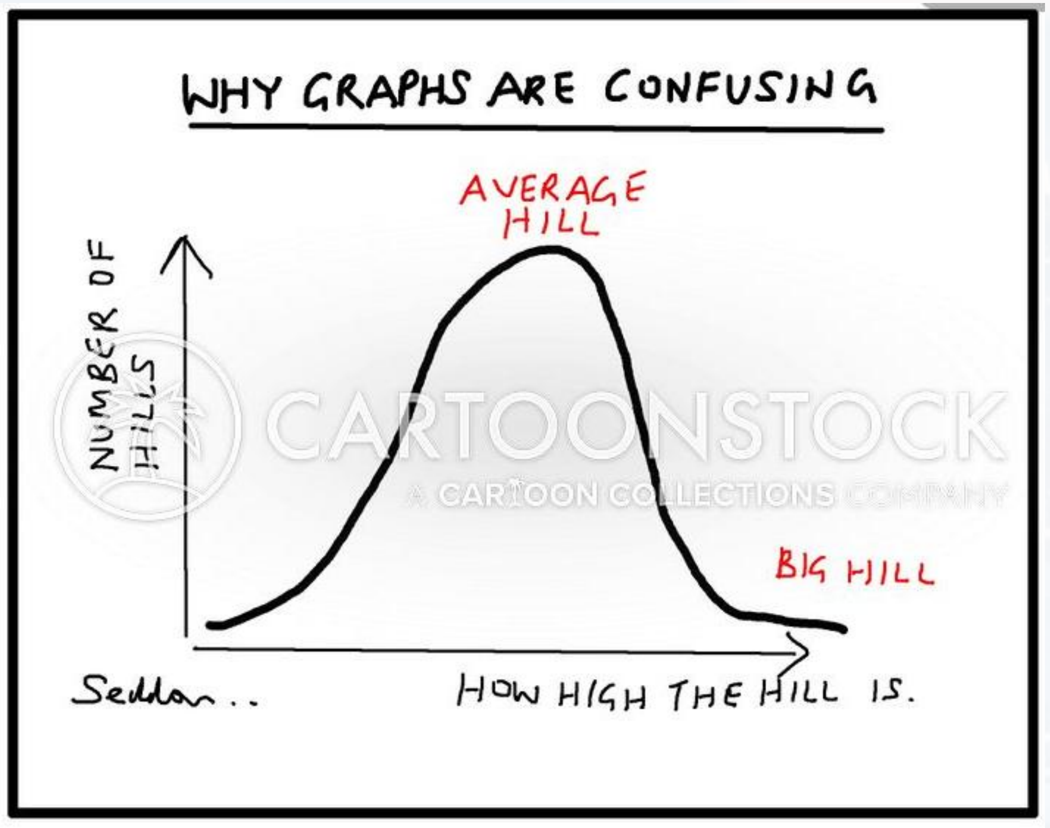
$$\boxed{P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}}$$

Poisson distribution

$$\mu = \lambda$$
$$var(k) = \lambda$$

```
c    = 5
delt = 10
lam  = c * delt
K    = np.random.poisson(lam, N)
labels, counts = np.unique(K, return_counts = True)
plt.bar(labels, counts, align = 'center', width = 0.1)
plt.gca().set_xticks(labels)
plt.title('N = ' + str(N))
```





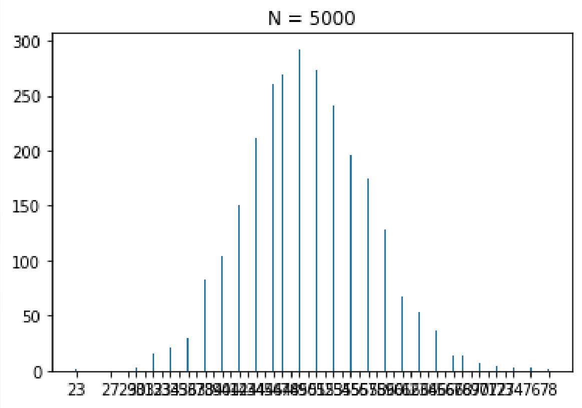
Outline

- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- **Normal Distribution**
- Central Limit Theorem

Random Variables and Distributions: Normal Distribution

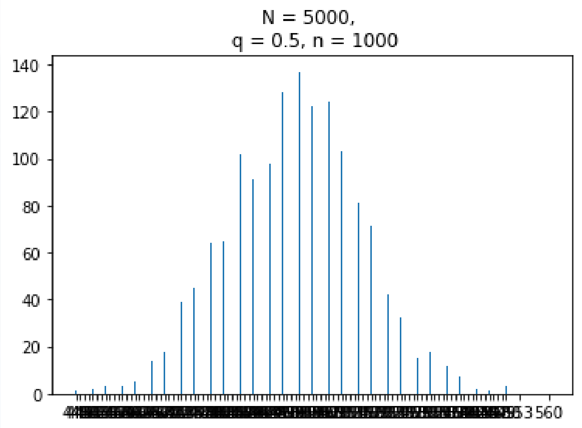
$$P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}$$

Poisson distribution



$$P(k|q,n) = \binom{n}{k}\, q^k\, (1-q)^{n-k}$$

binomial distribution



Stirling's approximation for even larger n

$$P(k|n,p) \approx \frac{1}{\sqrt{2\pi\, nq(1-q)}} \exp\left[-\frac{(k-nq)^2}{2nq(1-q)}\right]$$

Stirling's approximation for even larger n

$$P(k|n,p) \approx \frac{1}{\sqrt{2\pi\, nq(1-q)}} \exp\left[-\frac{(k-nq)^2}{2nq(1-q)}\right]$$

using $\sigma^2 = var(k) = qn(1-q)$

$\mu = qn$

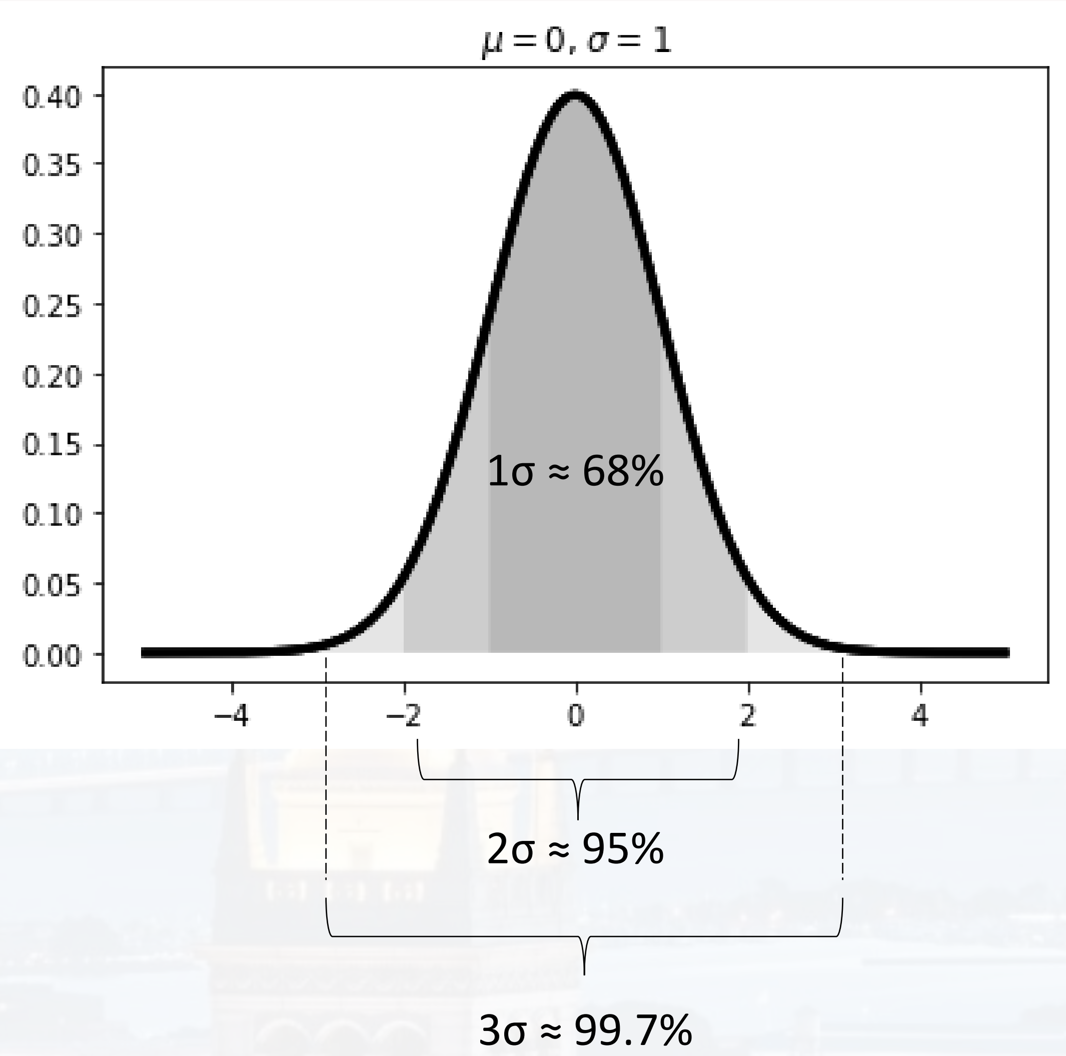
and $k := x$

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

Normal/Gauss distribution

Note, that the **Poisson** and the **Binomial distribution** are *discrete*, whereas the **Normal distribution** is *continuous*!

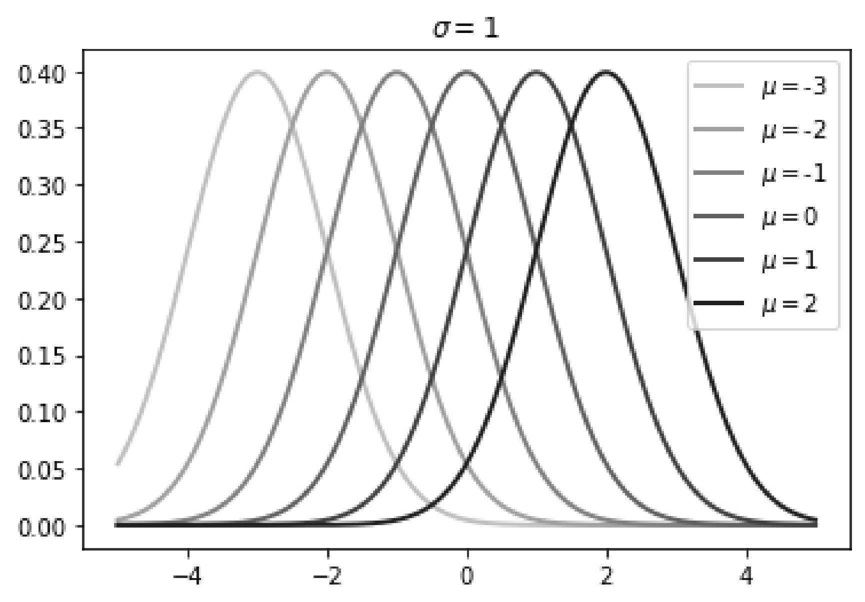
Random Variables and Distributions: Normal Distribution

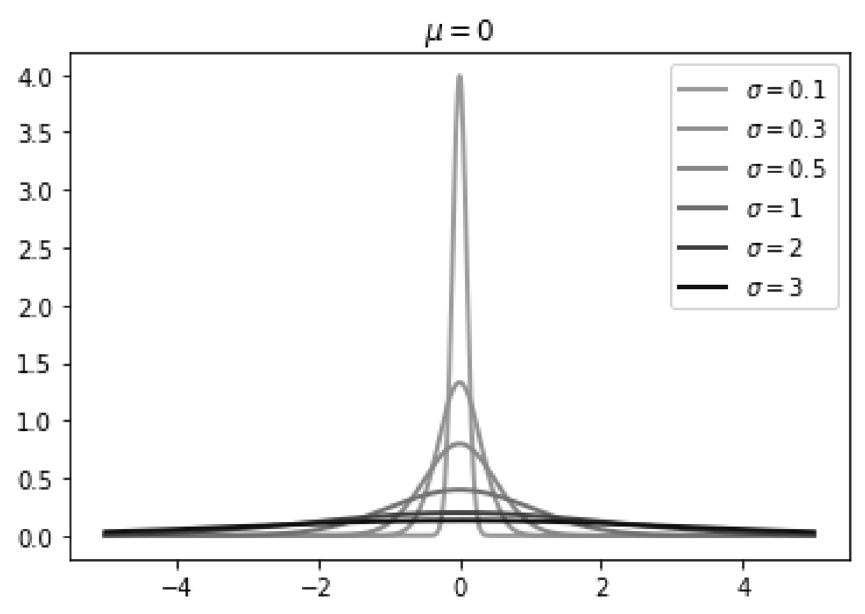


$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\,exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right]$$

Normal/Gauss distribution

Random Variables and Distributions:





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

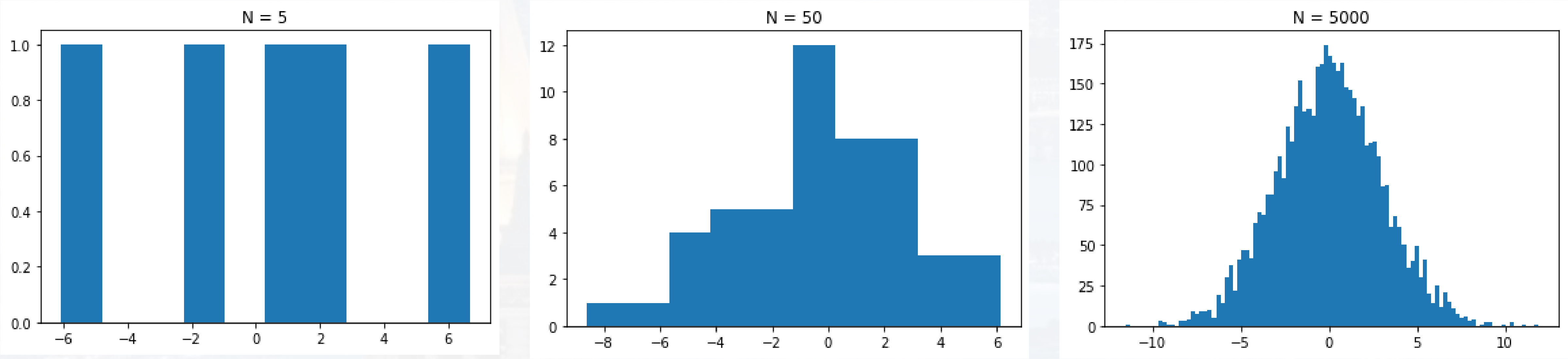
Normal/Gauss distribution

Random Variables and Distributions:

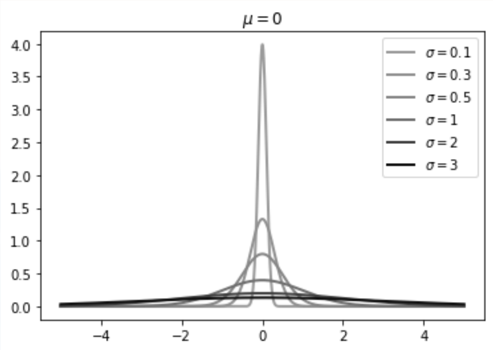

```
mu = 0
s  = 1
P  = np.random.normal(mu, s, N)
plt.hist(P)
plt.title('N = ' + str(N))
```

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

Normal/Gauss distribution



Random Variables and Distributions:



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

Normal/Gauss distribution

examples:

- diffusion processes
- approx. stat. error of data points
- approx. distribution of body height/shoe sizes/ weight, IQ
- approx. blood pressure, blood values
- approx. retirement age

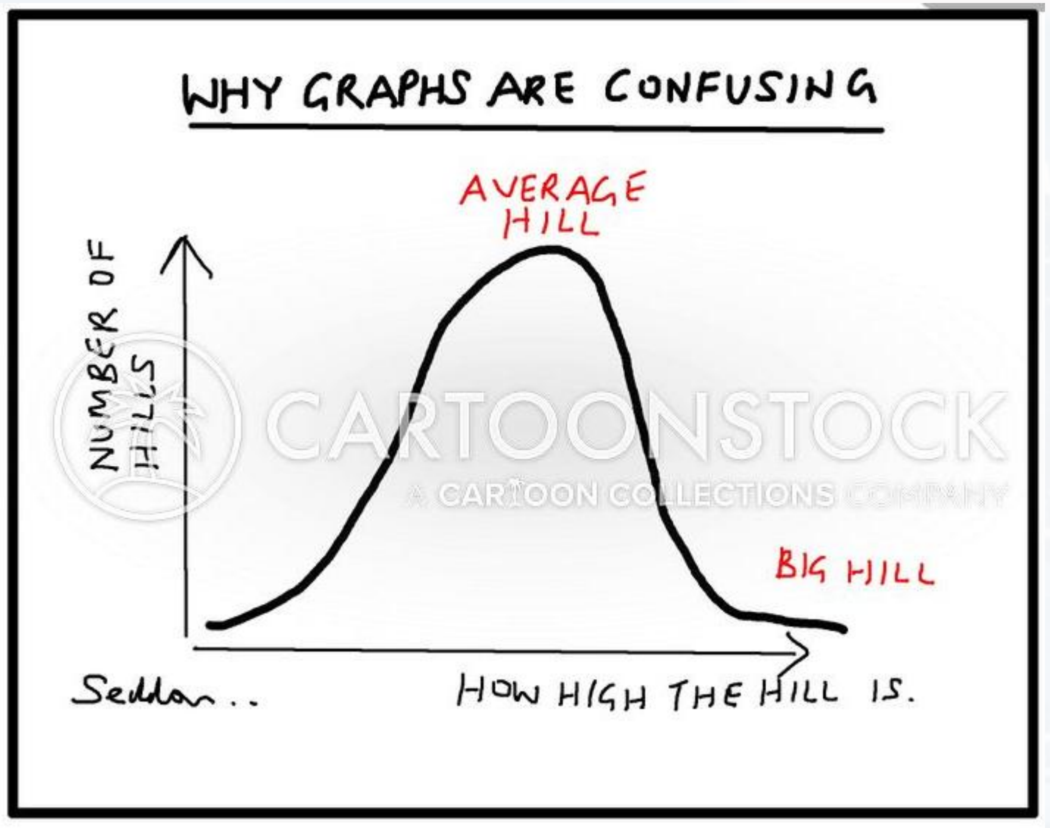
....

applications:

- significance tests
- t-test
- ANOVA/MANOVA
- χ^2 - test
- χ^2 itself and students-t distribution

...

Why do so many quantities follow a normal distribution?



<u>Outline</u>

- Uniform Distribution

- Binomial Distribution

- Poisson Distribution

- Normal Distribution

- Central Limit Theorem

Random Variables and Distributions:

binomial distribution

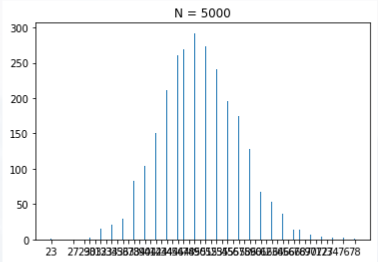
$$P(k|q,n) = \binom{n}{k}\, q^k\, (1-q)^{n-k}$$

q → 0

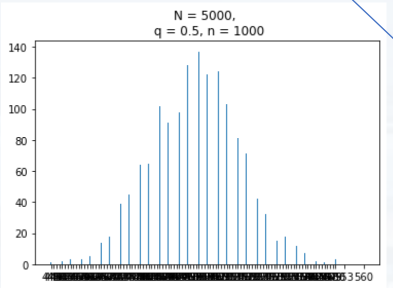
Poisson distribution

$$P(k|c) = \frac{(c\,\Delta t)^k e^{-c\,\Delta t}}{k!}$$

n → ∞



n → ∞



Normal/Gauss distribution

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

The fact that that many datasets can be well approximated by a normal distribution for n → ∞ is called

Central Limit Theorem

Why do so many quantities follow a normal distribution?

At the end… all probability distributions are **Maximum Entropy** Distributions, subject to a **set of constrains**

Distribution name	Probability density / mass function	Maximum Entropy constraint	Support
Uniform (discrete)	$f(k)=\frac{1}{b-a+1}$	None	$\{a,a+1,\dots,b-1,b\}$
Uniform (continuous)	$f(x)=\frac{1}{b-a}$	None	$[a,b]$
Bernoulli	$f(k)=p^k(1-p)^{1-k}$	$\mathbb{E}[\,K\,]=p$	$\{0,1\}$
Geometric	$f(k)=(1-p)^{k-1}\,p$	$\mathbb{E}[\,K\,]=\frac{1}{p}$	$\mathbb{N}\smallsetminus\{0\}=\{1,2,3,\dots\}$
Exponential	$f(x)=\lambda\exp(-\lambda x)$	$\mathbb{E}[\,X\,]=\frac{1}{\lambda}$	$[0,\infty)$
Laplace	$f(x)=\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	$\mathbb{E}[\,\lvert X-\mu\rvert\,]=b$	$(-\infty,\infty)$
Asymmetric Laplace	$f(x)=\frac{\lambda\,\exp\left(-(x-m)\,\lambda\,s\,\kappa^s\right)}{(\kappa+\frac{1}{\kappa})}$ where $s\equiv\operatorname{sgn}(x-m)$	$\mathbb{E}[\,(X-m)\,s\,\kappa^s\,]=\frac{1}{\lambda}$	$(-\infty,\infty)$
Pareto	$f(x)=\frac{\alpha\,x_m^\alpha}{x^{\alpha+1}}$	$\mathbb{E}[\,\ln X\,]=\frac{1}{\alpha}+\ln(x_m)$	$[x_m,\infty)$
Normal	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{E}[\,X\,]=\mu$, $\mathbb{E}[\,X^2\,]=\sigma^2+\mu^2$	$(-\infty,\infty)$

Why do so many quantities follow a normal distribution?

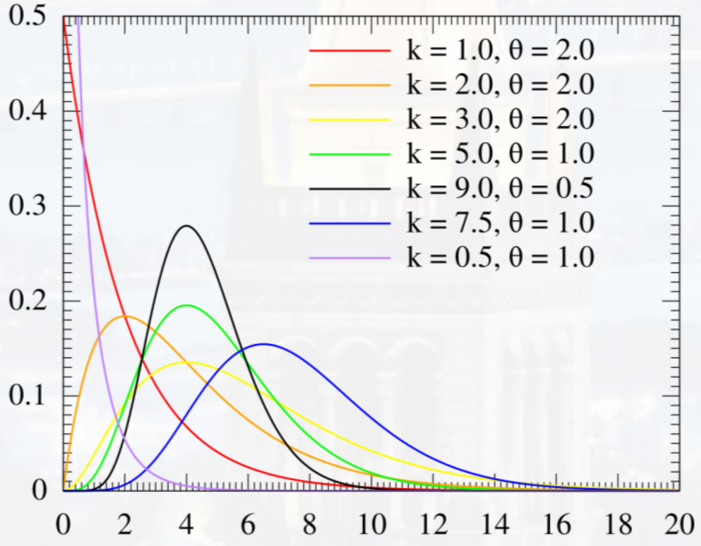
At the end… all probability distributions are **Maximum Entropy** Distributions, subject to a **set of constrains**

<u>examples:</u>

- approx. stat. error of data points
- approx. distribution of body height/shoe sizes/ weight, IQ
- approx. blood pressure, blood values
- approx. retirement age

….

Gamma	$f(x) = \dfrac{x^{k-1}\exp\left(-\frac{x}{\theta}\right)}{\theta^k\,\Gamma(k)}$	$\mathbb{E}[\,X\,] = k\,\theta\,,$ $\mathbb{E}[\,\ln X\,] = \psi(k) + \ln\theta$	$\boxed{[0,\infty)}$



note:

often, the exact model is actually the **gamma distribution**, but a normal distribution is **simpler** and usually a **good approximation**

Thank you very much for your attention!