Statistical Data Analysis 1

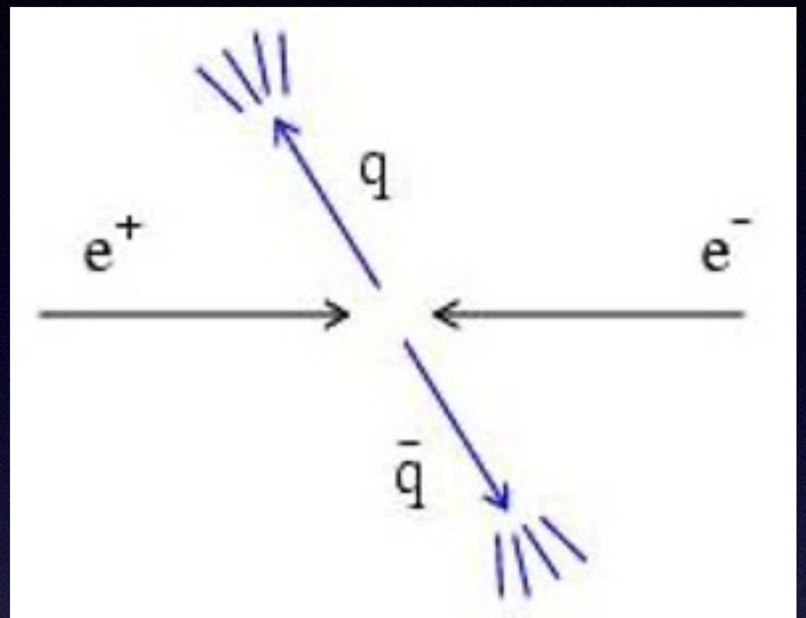
Tae Jeong Kim
(Hanyang University)

For 32nd Vietnam School of Physics:
Particle Physics and Dark Matter
July 17, 2026

Disclaimer

- Materials for this slide are mainly from the book of “Statistical Data Analysis” by Glen Cowan

Data Analysis in Particle Physics



Observe events of a certain type

- Estimate parameters
- Quantify the uncertainty of the parameter esimates (most important)
  - we will have completely different conclusion depending on the uncertainty
- Test hypothesis (Dark matter, SuperSymmetry, Any new physics…)

Dealing with uncertainty

- Statistical uncertainty
  - Random measurement errors
  - We can not repeat the experiment infinitely
- Systematic uncertainty
  - Theory is not deterministic (Quantum Mechanics)
  - We don’t have a perfect experiment (limited cost)

We can quantify the uncertainty using probability

What is probability?

- Relative frequence

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is A}}{n}$$

- Subjective probability

$$P(A) = \text{degree of belief that A is true}$$

- Mostly the frequency interpretation is useful
- Subective probability can provide more natural tretment of non-repeatable phenomena
  - Probability that Big Bang happened
  - Probability that Dark Matter candidate exists

Bayes' theorem

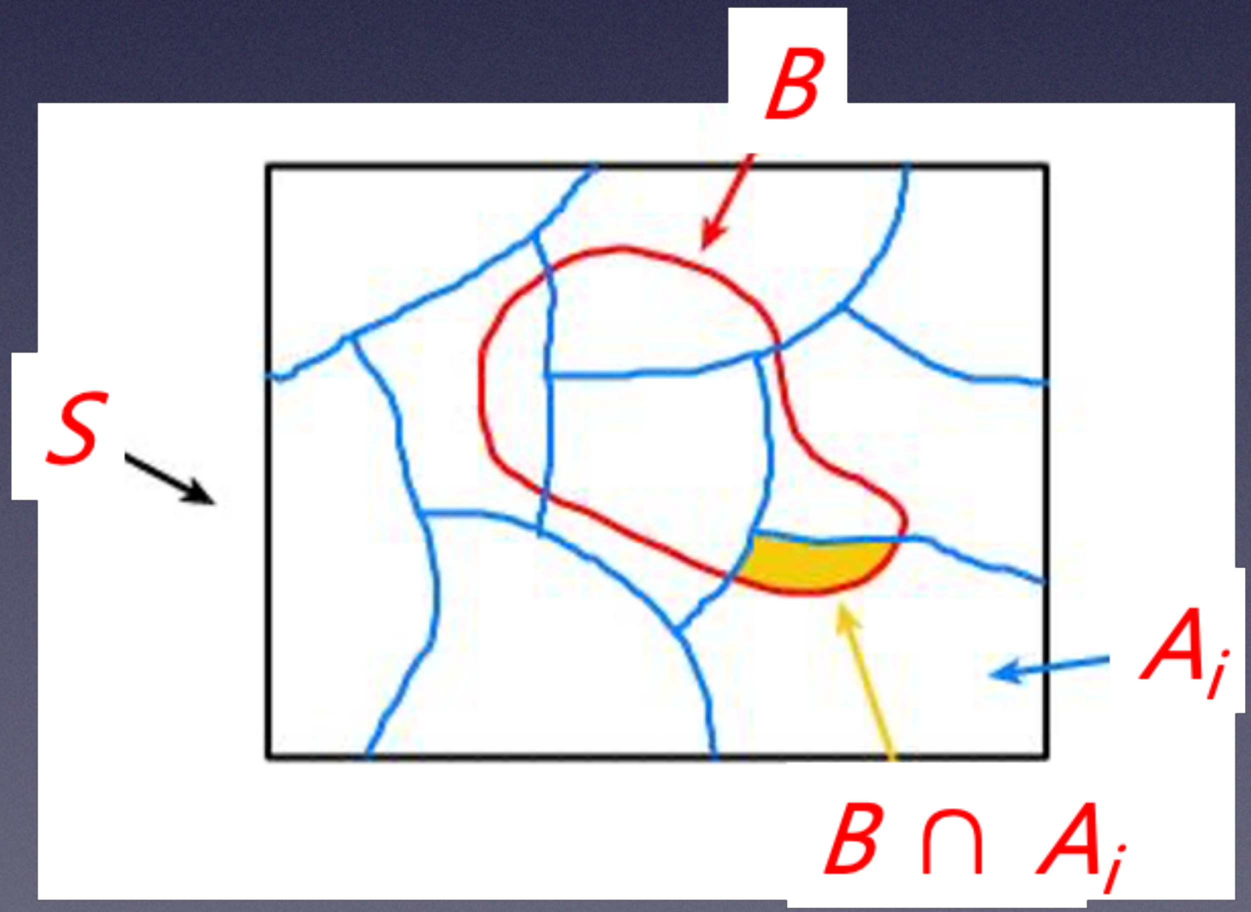
- From the definition of conditional probability, we have

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

- But $P(A \cap B) = P(B \cap A)$, so

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

Using the law of total probability →

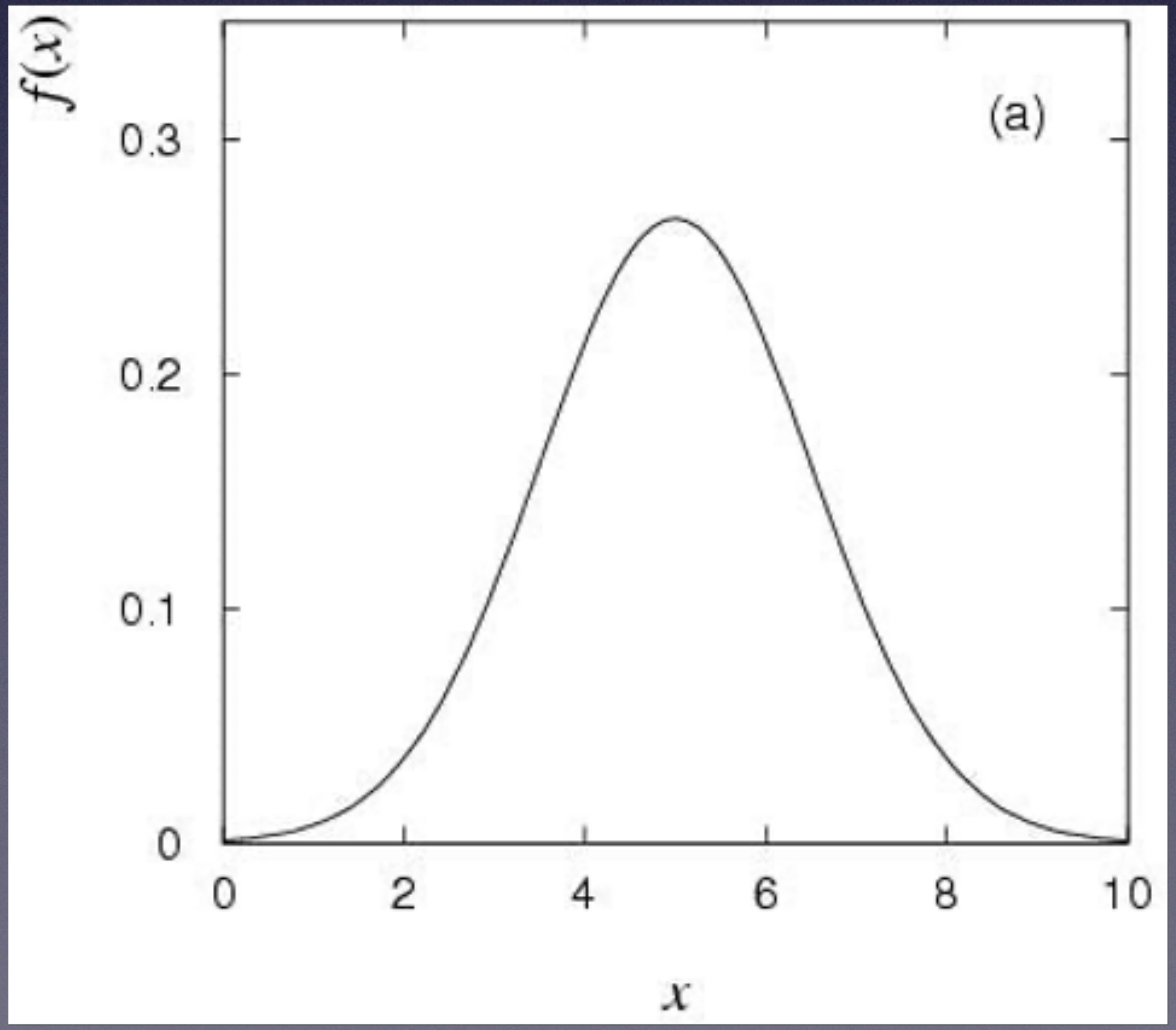


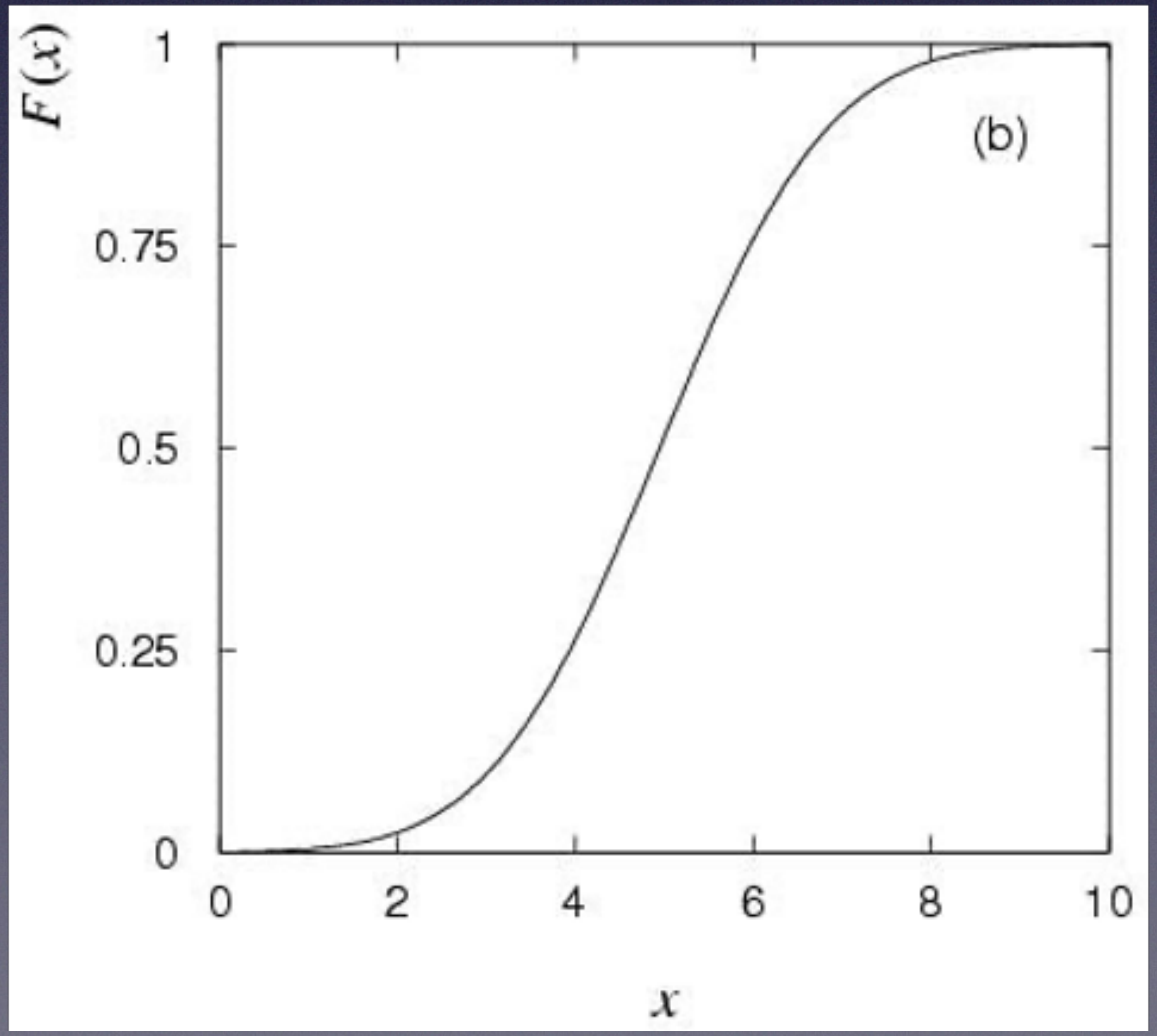
$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}}$$

Cumulative distribution function

- Probability to have outcome less than or equal to x is

$$\int_{-\infty}^{x} f(x')dx' \equiv F(x)$$





Bayesian Statistics

- In Bayesian statistics, use subjective probability for hypotheses:

Probability of the data assuming hypothesis H (likelihood)

Prior probability before seeing the data

$$P(H|obs) = \frac{P(obs|H)P(H)}{\int P(obs|H)P(H)dH}$$

Posterior probability after seeing the data

Sum over all possible hypotheses

- No unique prescription for priors (subjective) → probability given data will change depending on the priors

Example of Bayes' theorem

- Suppose the probability to have a disease D is

$$\begin{aligned} P(\mathrm{D}) &= 0.001 \\ P(\text{no D}) &= 0.999 \end{aligned}$$

(Prior probabilities)

- Consider a test for disease: result + or -

$$\begin{aligned} P(+|\mathrm{D}) &= 0.98 \\ P(-|\mathrm{D}) &= 0.02 \end{aligned}$$

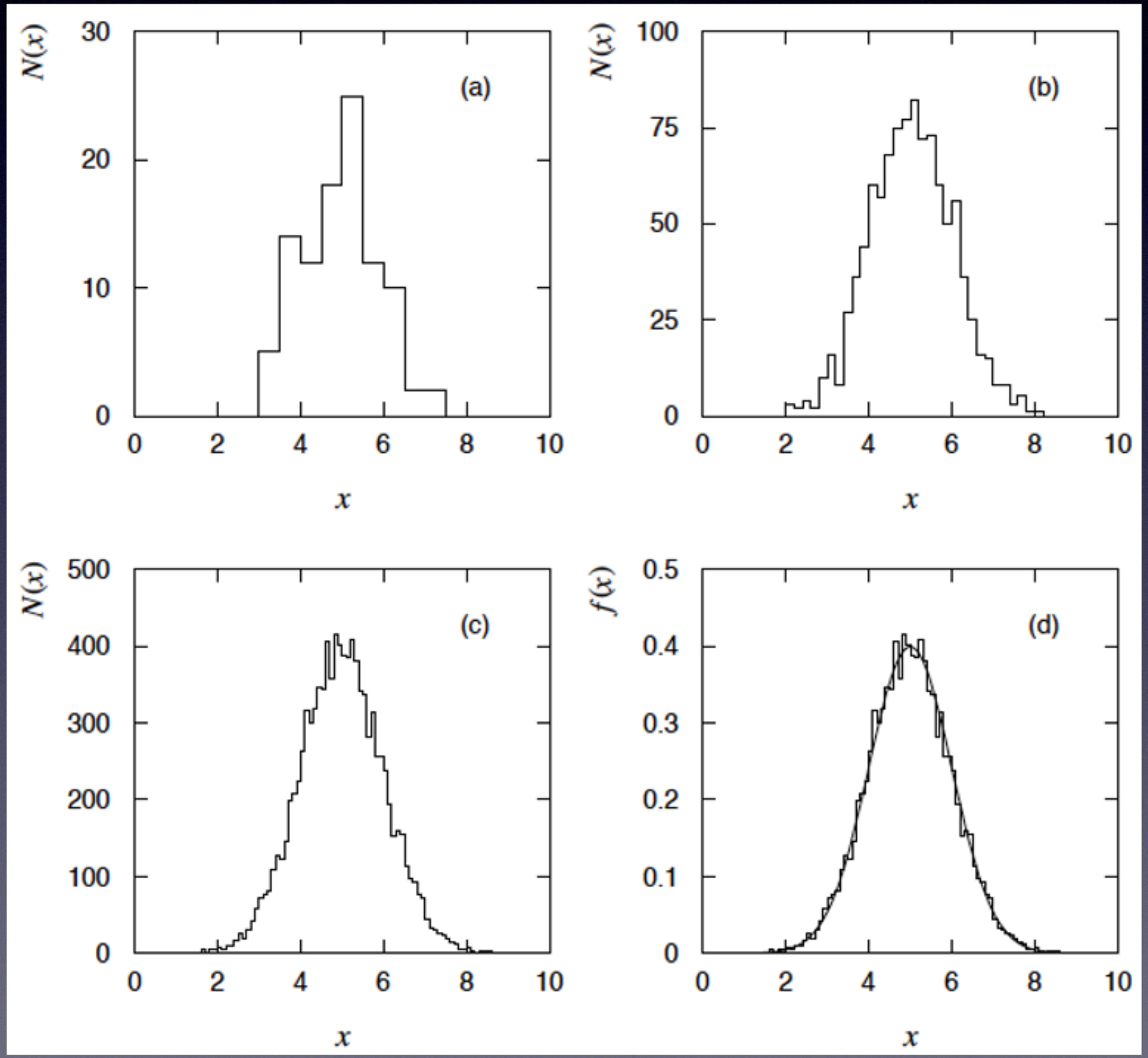
$$\begin{aligned} P(+|\text{no D}) &= 0.03 \\ P(-|\text{no D}) &= 0.97 \end{aligned}$$

- Suppose your result is +. How worried should you be?

$$\begin{aligned} p(\mathrm{D}|+) &= \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\text{no D})P(\text{no D})} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\ &= 0.032 \end{aligned}$$

- You are probably OK!
  - My degree of belief that I have the desease is 3.2%
  - Doctor's viewpoint: 3.2% of people like this have the disease

Histograms and PDF



- From Histogram

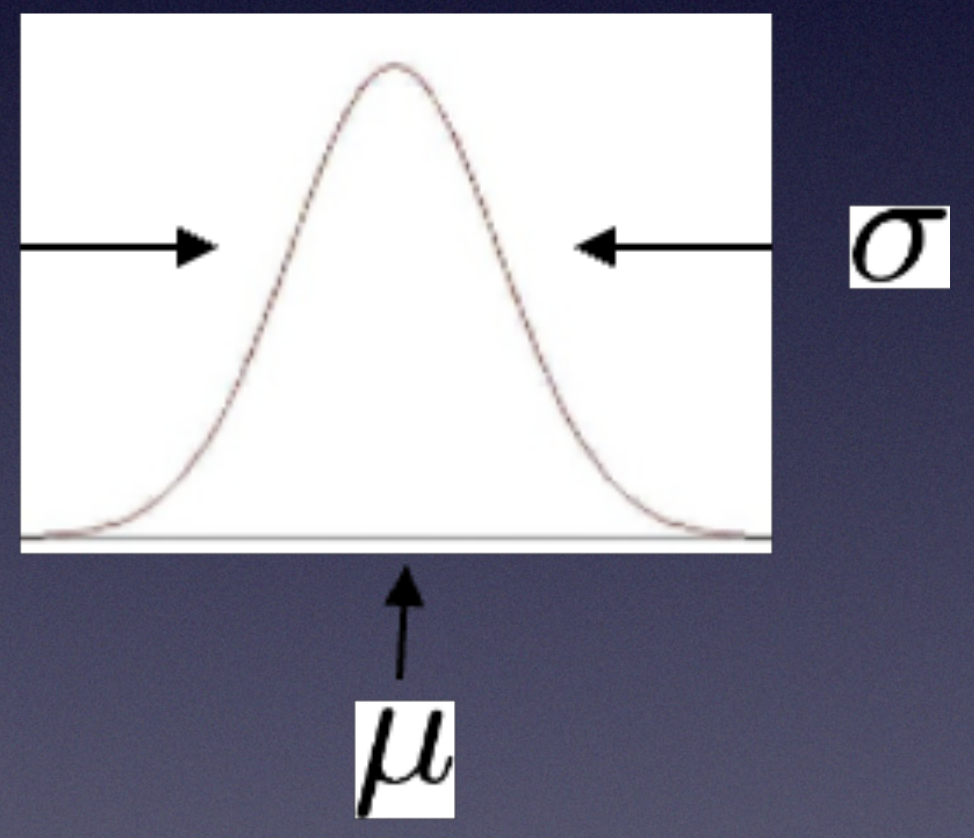
$$f(x) = \frac{N(x)}{n\Delta x},\ \Delta x \to 0$$

- To PDF (Probability Distribution Function)

Expectation values

- Consider a continuous random variable x with PDF $f(x)$
- Define expectation values as

$$E[x] = \int x f(x) dx = \mu$$



- Variance $V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2] = \sigma^2$
- Standard deviation σ ~ width of PDF

Covariance and correlation

- Covariance with x and y

$$\mathrm{COV}[x, y] = E(xy) - \mu_x \mu_y = E[(x - \mu_x)(y - \mu_y)]$$

- Correlation coefficient defined as

$$\rho_{xy} = \frac{\mathrm{COV}[x, y]}{\sigma_x \sigma_y}$$

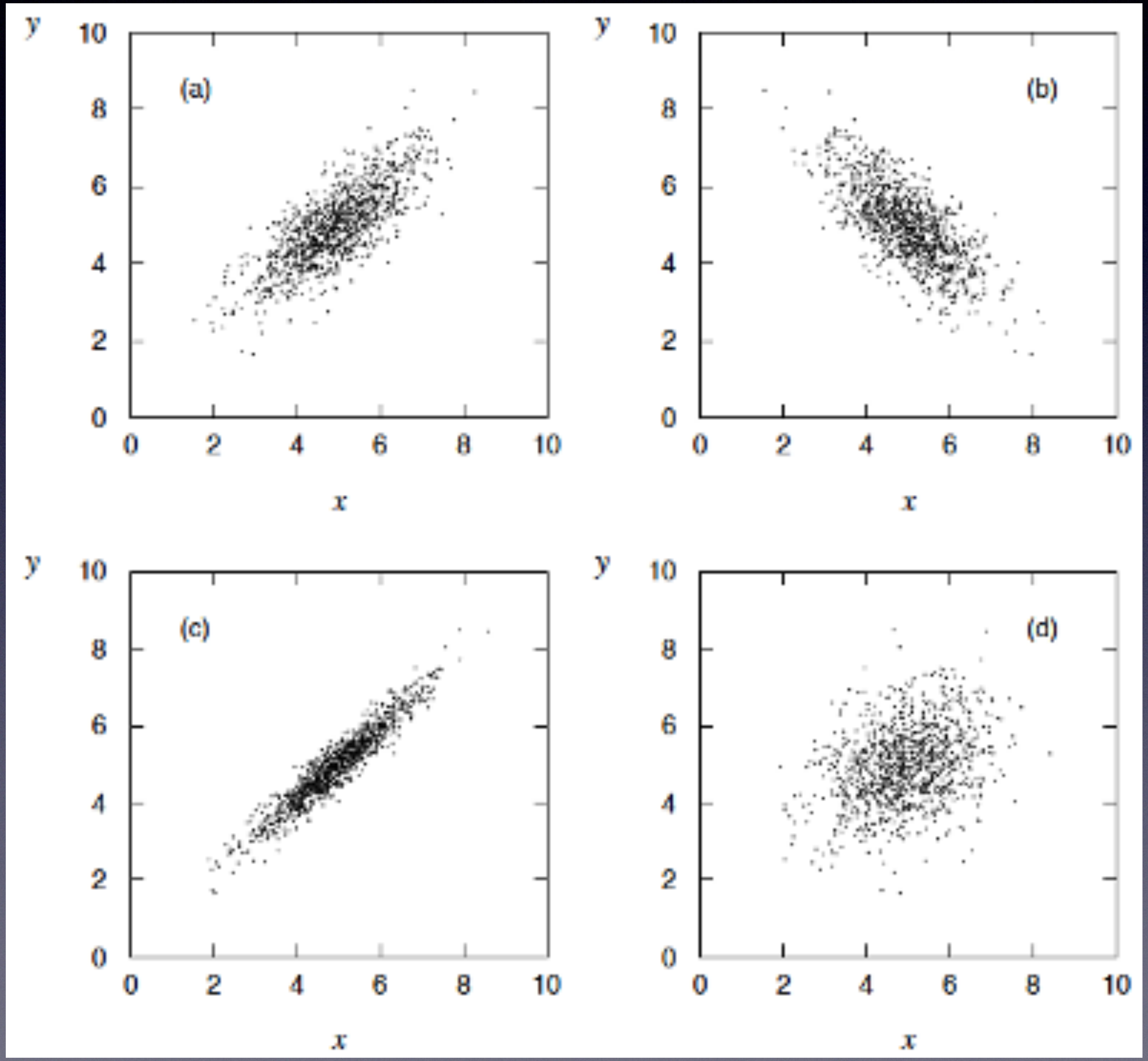
- If x and y are independent

$$\mathrm{COV}[x, y] = 0$$

Correlation

$\rho = 0.75$

$\rho = -0.75$



$\rho = 0.95$

$\rho = 0.25$

Combination of measurements

- Consider two measurements $$a = m_1 \pm \sigma_1$$ $$b = m_2 \pm \sigma_2$$
- Minimize χ^2 to determine the best m $$\chi^2 = \frac{(m - m_1)^2}{\sigma_1^2} + \frac{(m - m_2)^2}{\sigma_2^2}$$

$$\frac{\partial \chi^2}{\partial m} = 0$$

Weight average

$$m = \hat{m} = \frac{\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\omega_1 m_1 + \omega_2 m2}{\omega_1 + \omega_2}$$

The uncertainty on $\hat{m}$ is

$$\hat{m} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

Combination with correlation

$$\chi^2 = \begin{pmatrix} m - m_1 & m - m_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} m - m_1 \\ m - m_2 \end{pmatrix}$$

Solution of the minimization of χ^2

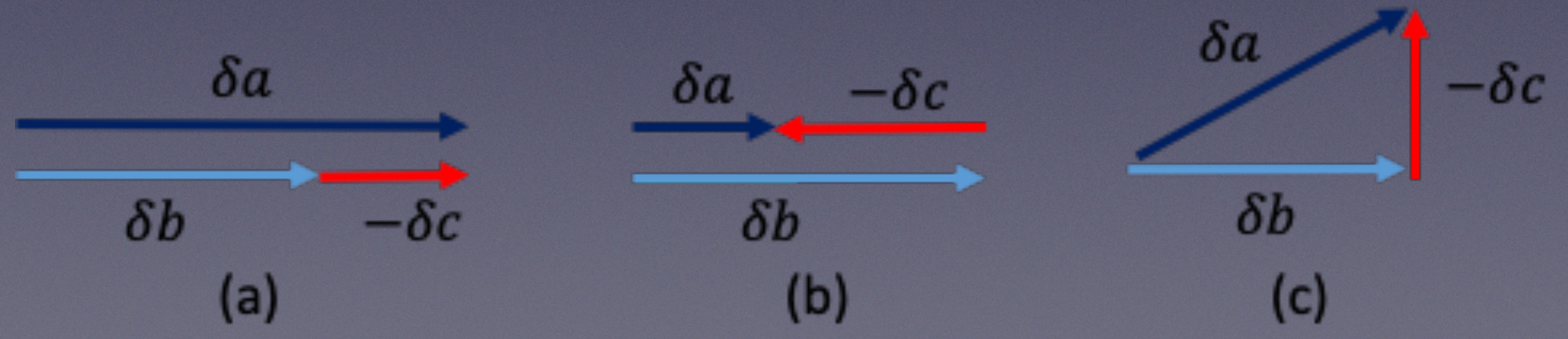
$$\hat{m} = \frac{m_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + m_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

The uncertainty becomes $\sigma_m^2 = \dfrac{\sigma_1^2\sigma_2^2(1-\rho)^2}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$

Error propagation - linear

- example: a = b - c
- What is the uncertainty on a?
- $\sigma_a^2 = \sigma_b^2 + \sigma_c^2 - 2\mathrm{COV}[b, c] \rightarrow$

If the variables are uncorrelated:
add errors quadratically for the sum



Error propagation - nonlinear

- example: a = b/c
- What is the uncertainty on a?
- $$\frac{\sigma_a^2}{a^2} = \frac{\sigma_b^2}{b^2} + \frac{\sigma_c^2}{c^2} - 2\frac{\mathrm{COV}[b,c]}{bc} \rightarrow$$

If the variables are uncorrelated: add relative errors quadratically for the ratio

Distributions

Distribution/PDF	Example use in HEP
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Briet-Wigner	Mass of resonance
Landau	Ionization energy loss

Binomial distribution

- Consider N independent experiment (like tossing a coin)
  - When the outcome is “success” or “failure”
  - Probability of success is p
- Probability of having n times success: $p^n(1-p)^{N-n}$ (order is not important)

there are $\dfrac{N!}{n!(N-n)!}$ permutations to n success in

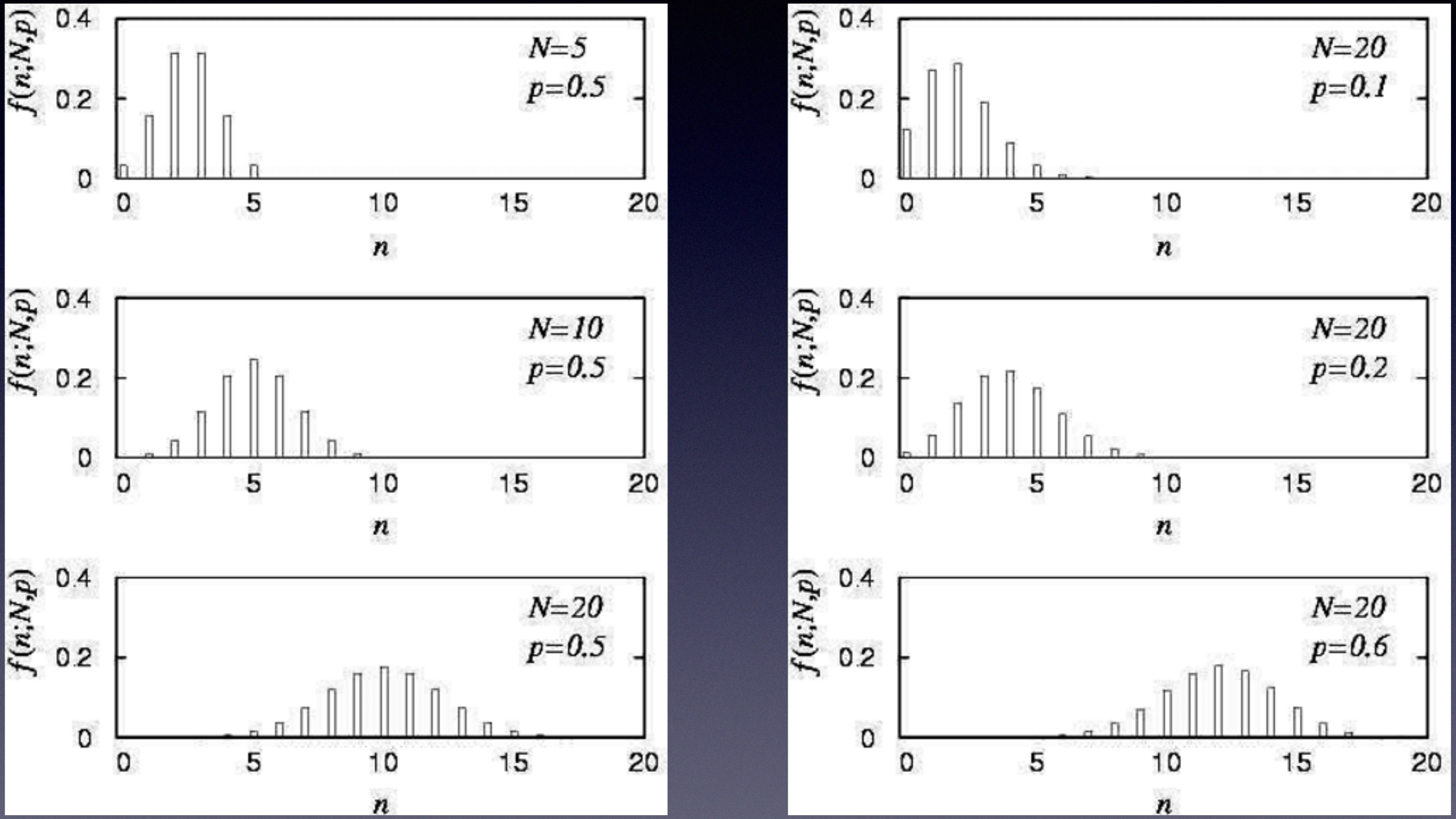
Binomial distribution

- The binomial distribution is

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

- For the expectation value and variance
    - $E[n] = Np$
    - $V[n] = E(n^2] - (E[n])^2 = Np(1-p)$

Binomial distribution



- Example: observe N decays of W boson, what is the number of $W \rightarrow \mu\nu$? this follows the binomial where p is branching ratio

Multinomial distribution

- Like binomial but now m outcomes instead of two, probabilities are

$$\vec{p} = (p_1, \ldots, p_m) \qquad \sum_{i=1}^{m} p_i = 1$$

- For N trials, we want the probability to obtain:

n_1 of outcome 1
n_2 of outcome 2
.
.
.
n_m of outcome m

- This is the multinomial distribution for $\vec{n} = (n_1, \ldots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} p \cdots p_m^{n_m}$$

Multinomial distribution

- For all n_i individually binomial with parameters N, p_i

$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i)$

Covariance $V_{ij} = Np_i(\delta_{ij} - p_j)$, $V_{ij} = - Np_i p_j$ for $i \neq j$

→ negatively correlated

- Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram with m bins, N total entries, all entries independent

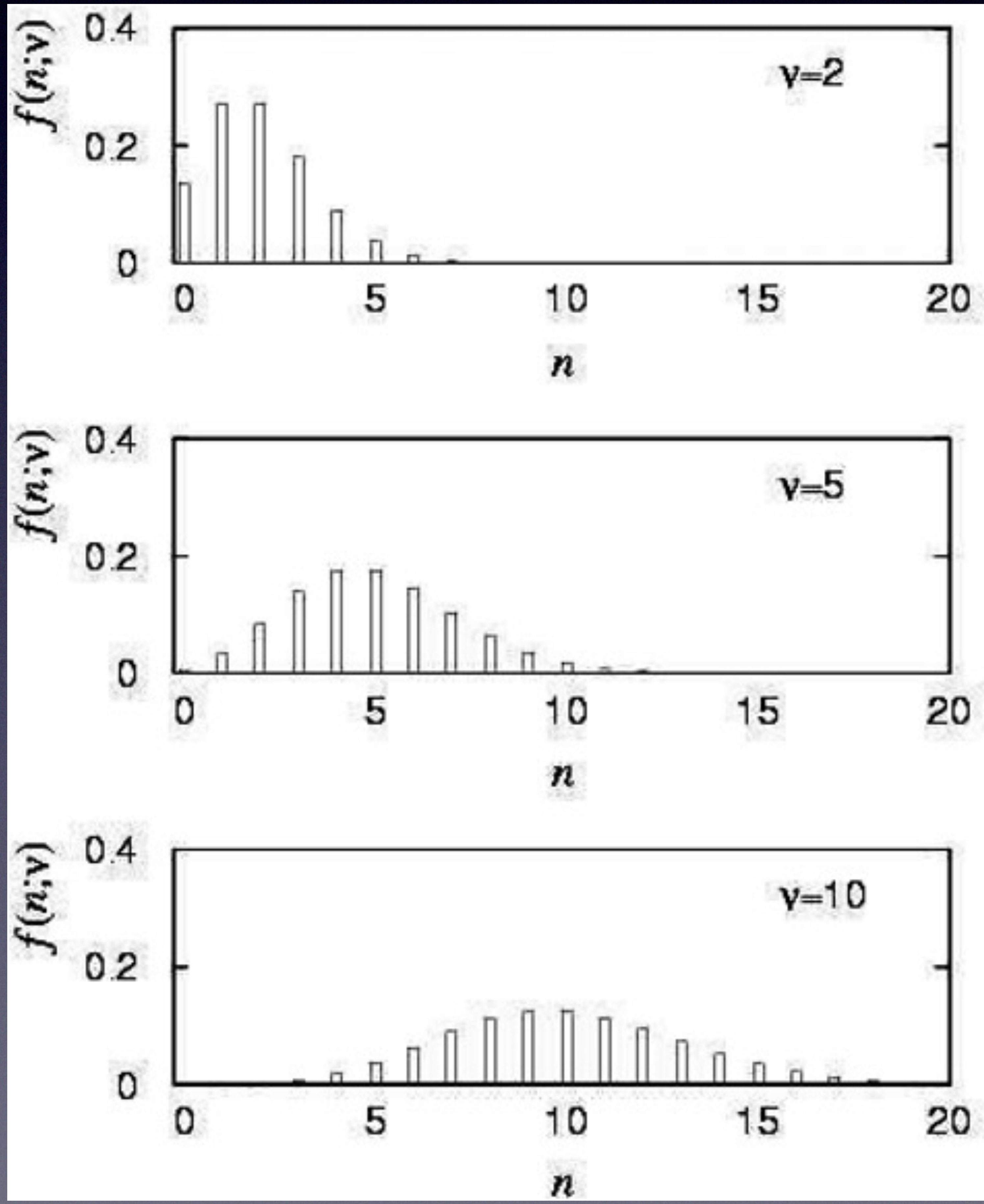
→ One bin goes up, the other bin goes down

Poisson distribution

- probability is small ($p \to 0$) and the number of experiments is large ($N \to \infty$), n follows the Poisson distribution:

$$f(n;\nu) = \frac{\nu^n}{n!} e^{-\nu} \; (n \geq 0)$$

$$E[n] = \nu,\; V[n] = \nu$$



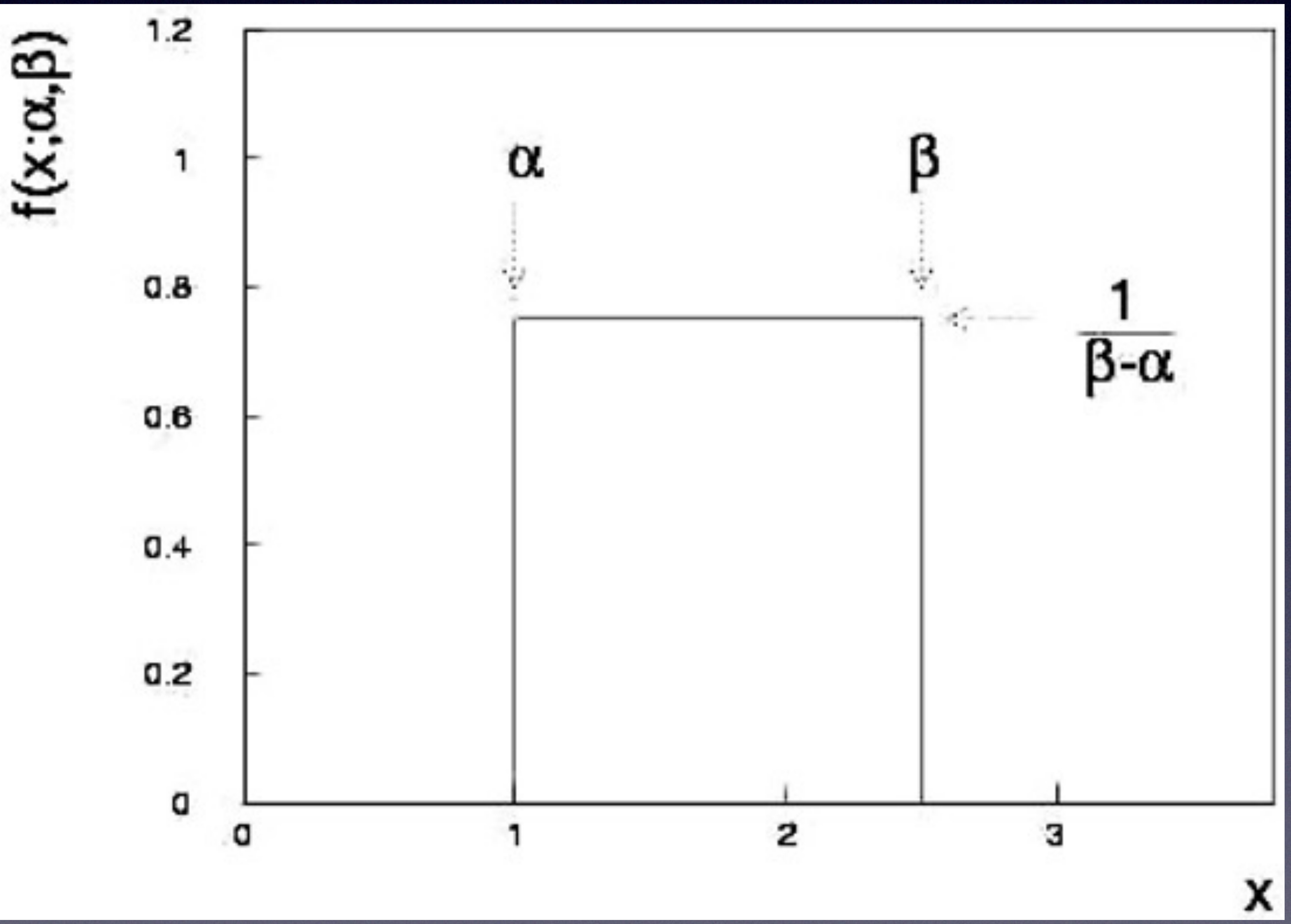
- Example: number of expected events with the cross section σ with integrated luminosity $\mathscr{L}$

$$\nu = \sigma \int \mathscr{L} dt$$

Uniform distribution

- Consider a continuous random variable x with $-\infty < x < \infty$

$$f(x;\alpha,\beta) = \frac{1}{\beta-\alpha} \; (\alpha < x < \beta)$$



$$E[x] = \frac{1}{2}(\alpha+\beta), \qquad V[x] = \frac{1}{12}(\beta-\alpha)^2$$

- Example: measure the spatial resolution of the detector using a scintillator as a trigger with cosmic muons

$$\sigma_t = \sqrt{\sigma_d^2 + \frac{w}{\sqrt{12}}}$$ (w is the width of the scintillator)

Exponential distribution

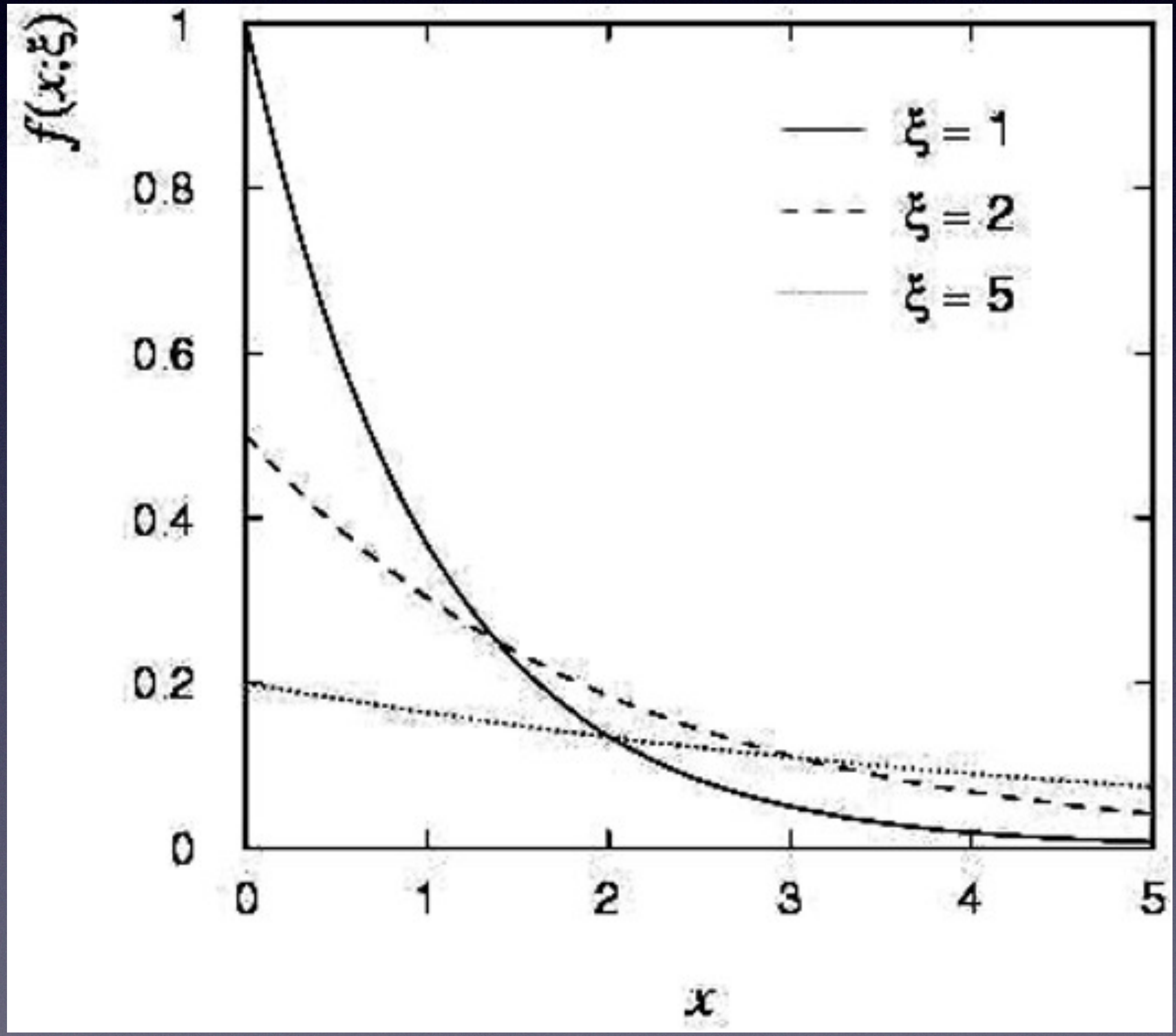
- Exponential PDF

$$f(x;\xi) = \frac{1}{\xi}e^{-x/\xi}\ (x \geq 0)$$

$$E[x] = \xi, \qquad V[x] = \xi^2$$

- Example: decay time t of an unstable particle

$$f(t;\tau) = \frac{1}{\tau}e^{-t/\tau}\ (\tau\text{=mean lifetime})$$



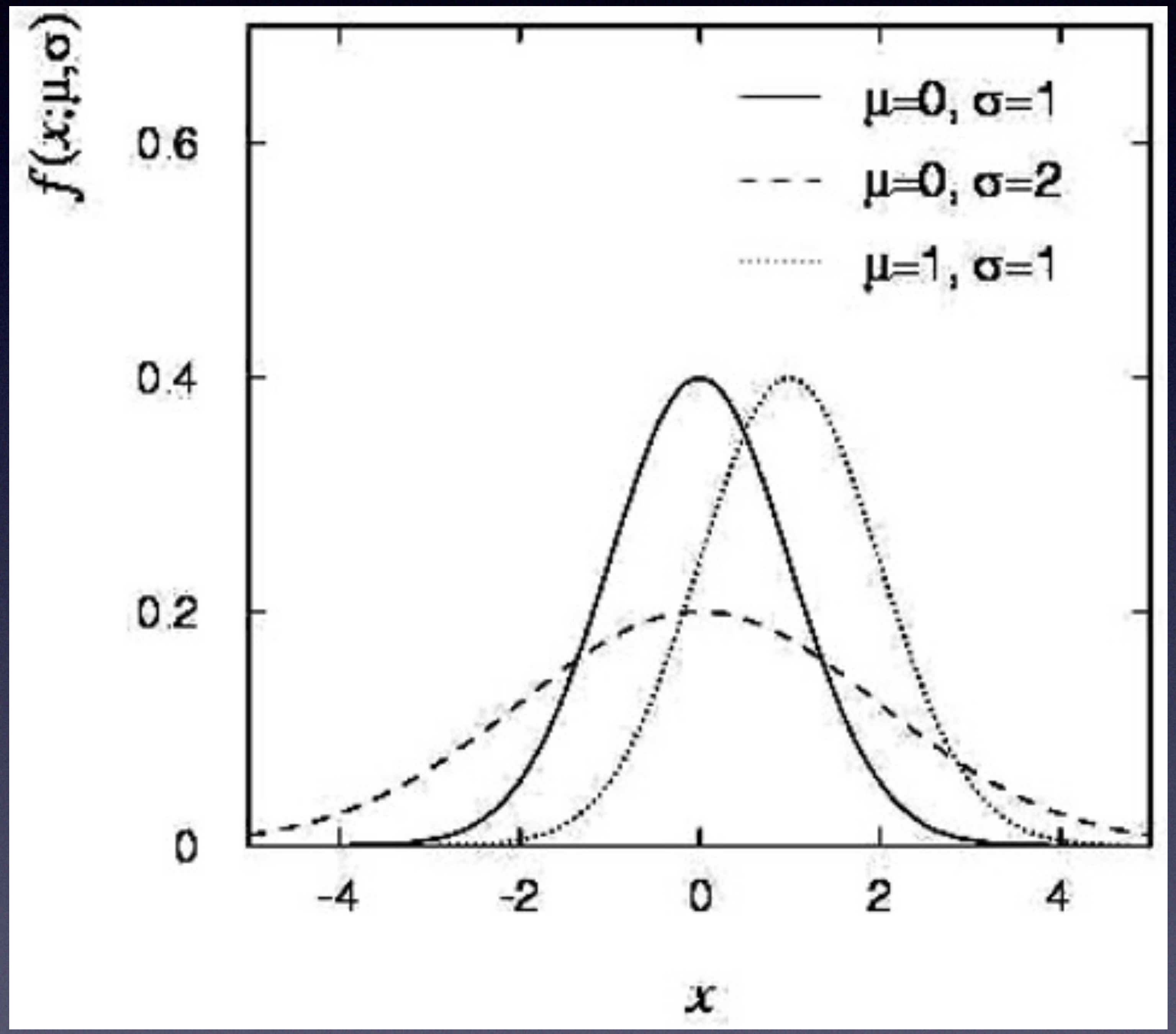
Gaussian distribution

- Gaussian PDF

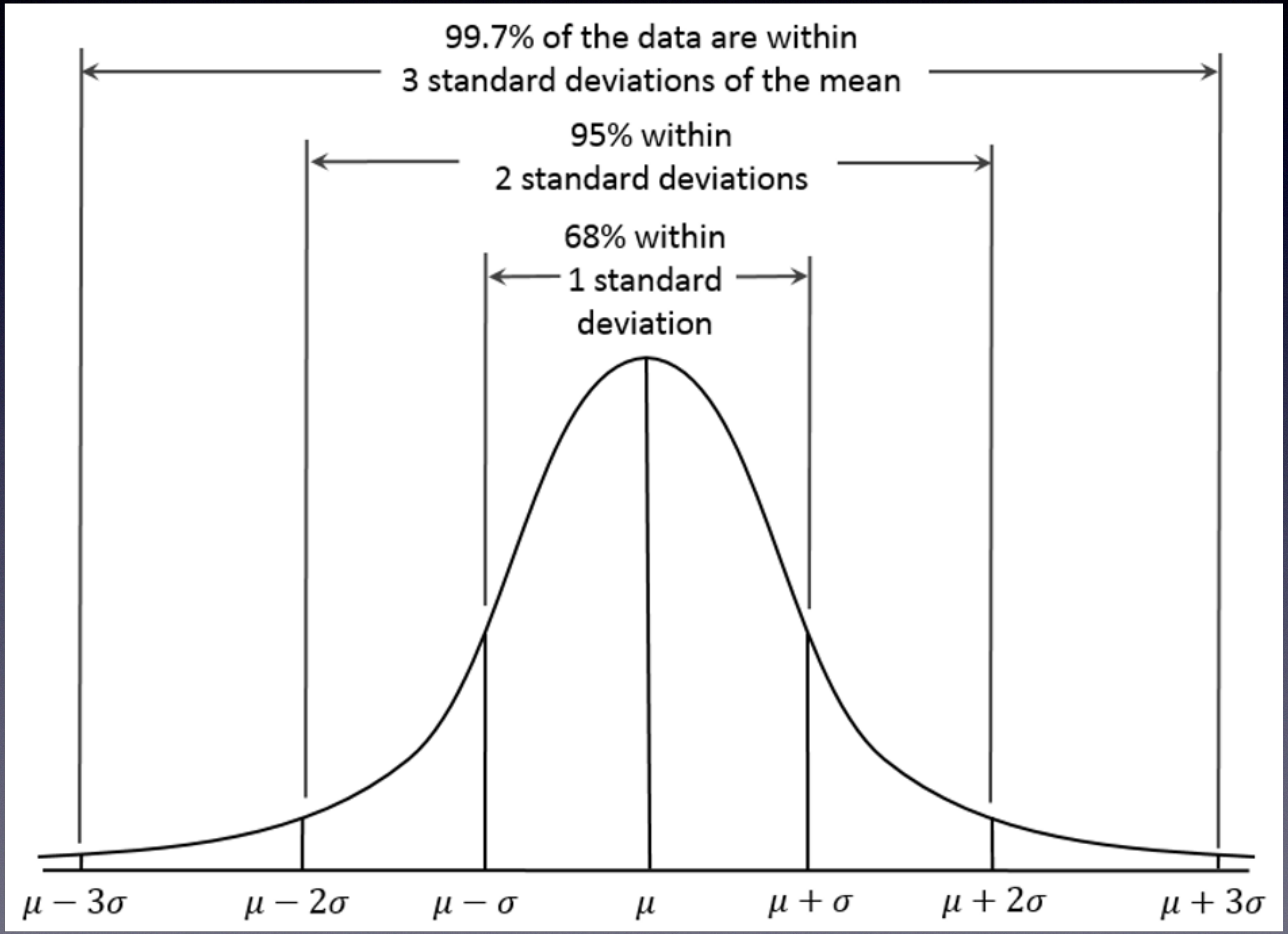
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[x] = \mu, \quad V[x] = \sigma^2$$

- If y is Gaussian with m, then $x = (y-m)/\sigma$ follows the standard Gaussian with $m = 0$, $\sigma = 1$



Gaussian distribution



$4\ \sigma \sim 99.994\%$

$5\ \sigma \sim 99.99994\%$

Gaussian PDF and the Central Limit Theorem

- For arbitrary PDFs with n independent random variable x_i with finite variance σ_i^2, consider the sum

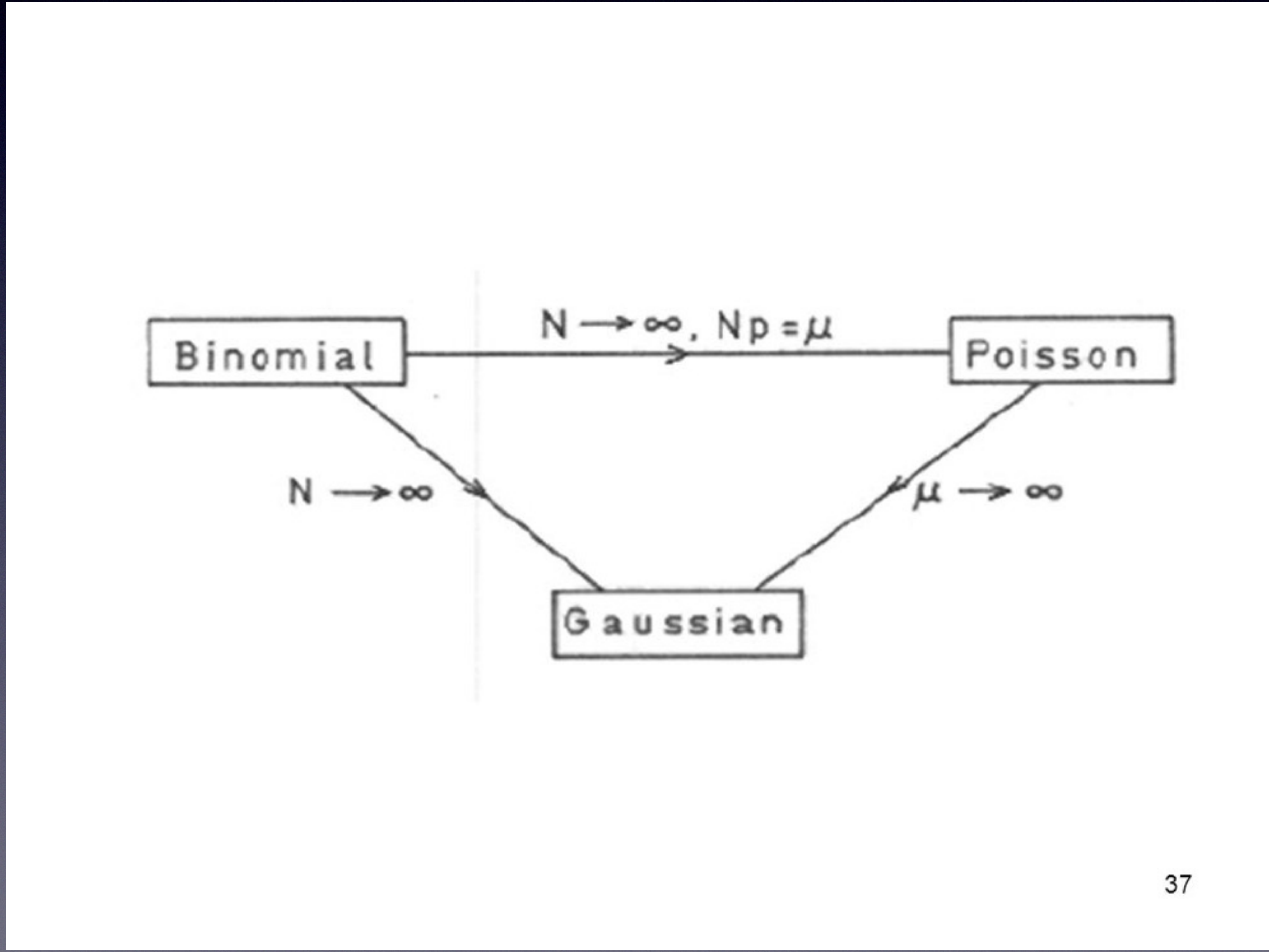
$$y = \sum_{i=1}^{n} x_i$$

- In the limit $n \to \infty$, y is a Gaussian random variable with

$$E[y] = \sum_{i=1}^{n} \mu_i \quad V[n] = \sum_{i=1}^{n} \sigma_i^2$$

- Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian random variable x

Relationships among Binomial, Poisson and Gaussian distributions



Chi-square (χ^2) distribution

- The Chi-square PDF for the continuous variable z is defined by

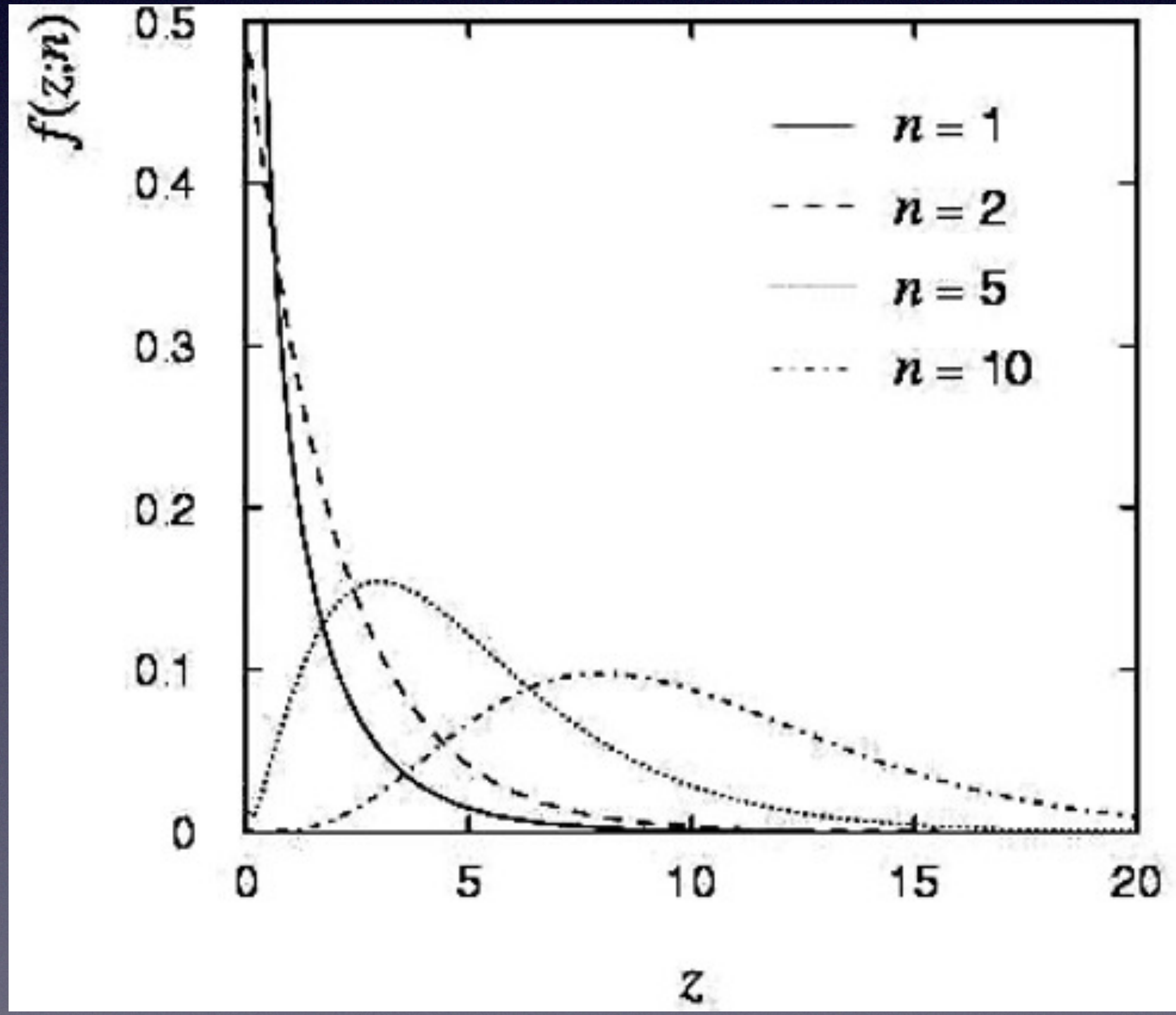
$$f(z;n) = \frac{1}{2^{n/2}\Gamma(k/2)} z^{(n/2)-1} e^{-z/2}$$

$$E[z] = n, \quad V[z] = 2n$$

- For independent Gaussian x_i,

$$\chi^2_{min} = z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

follows χ^2 PDF with n degree of freedom
expected to have χ^2/dof ~ 1



Breit-Wigner distribution

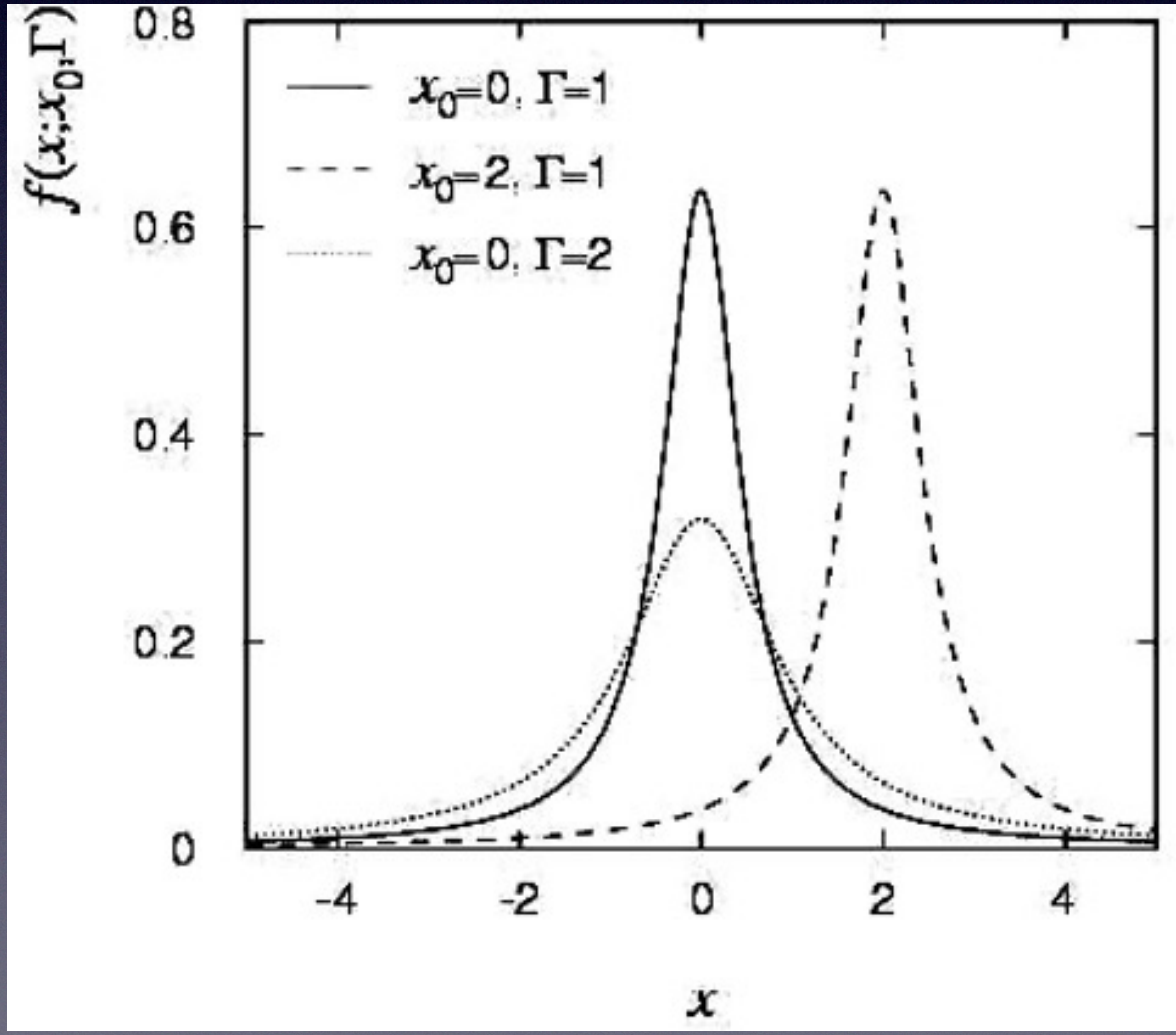
- The Breit-Wigner PDF for the continuous variable x is defined by

$$f(x;\Gamma,x_0) = \frac{1}{\pi}\frac{\Gamma/2}{\Gamma^2/4+(x-x_0)^2}$$

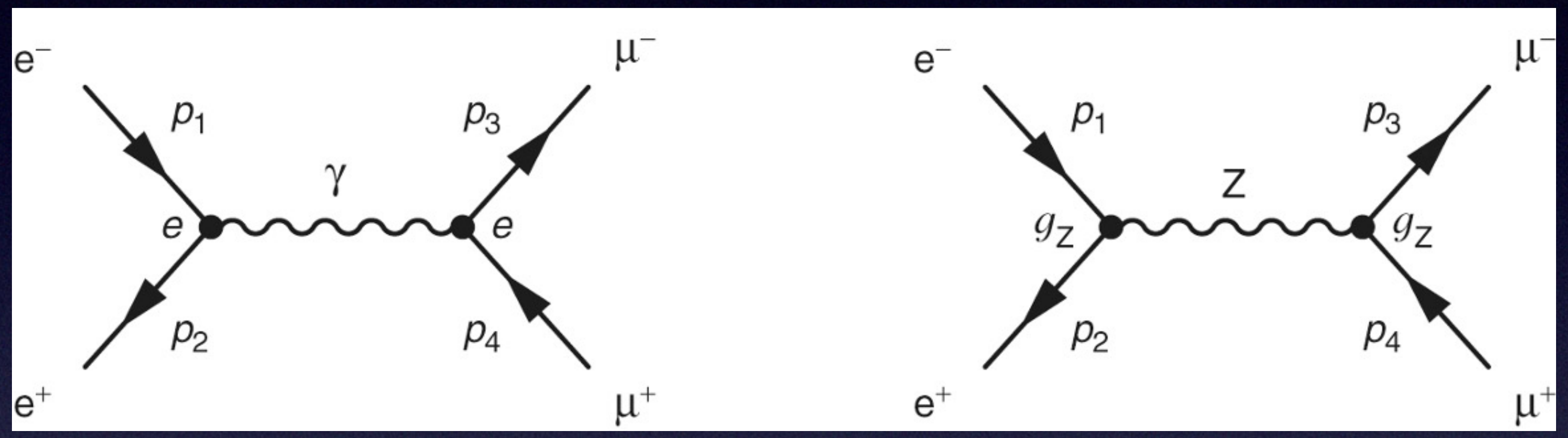
x_0 = mode (most probable value)
Γ=full width at half maximum

- Example: mass of resonance particle, e.g. Z boson, Higgs boson,…

- Γ=decay rate (inverse of mean lifetime)

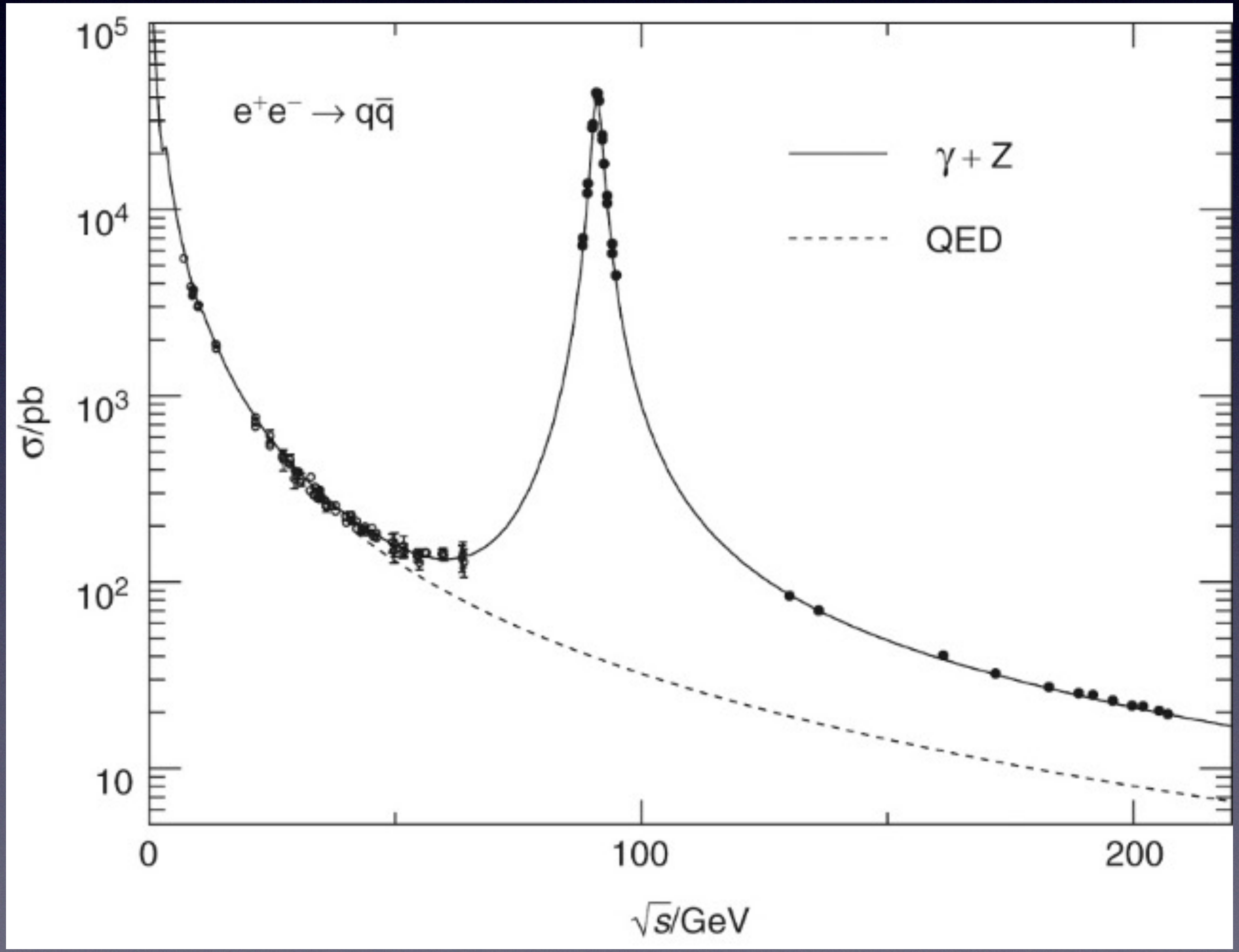


The Briet-Wigner resonance



- Cross-section for $e^+e^- \rightarrow \mu^+\mu^-(q\bar{q})$ is proportional to

$$\sigma \propto |M|^2 \propto \frac{1}{m_Z^2\Gamma_Z^2 + (s - m_Z^2)^2}$$



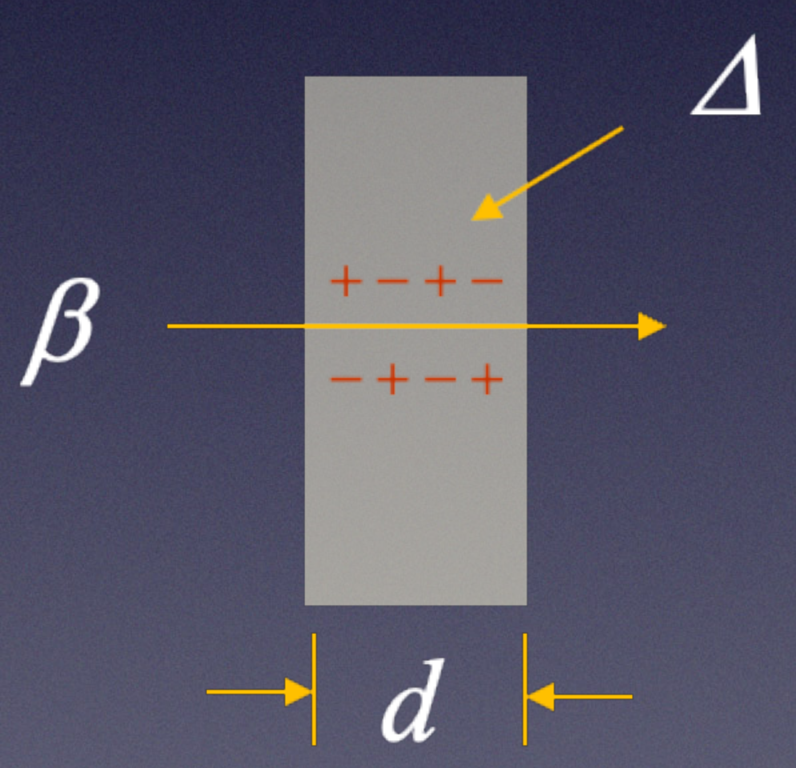
Landau distribution

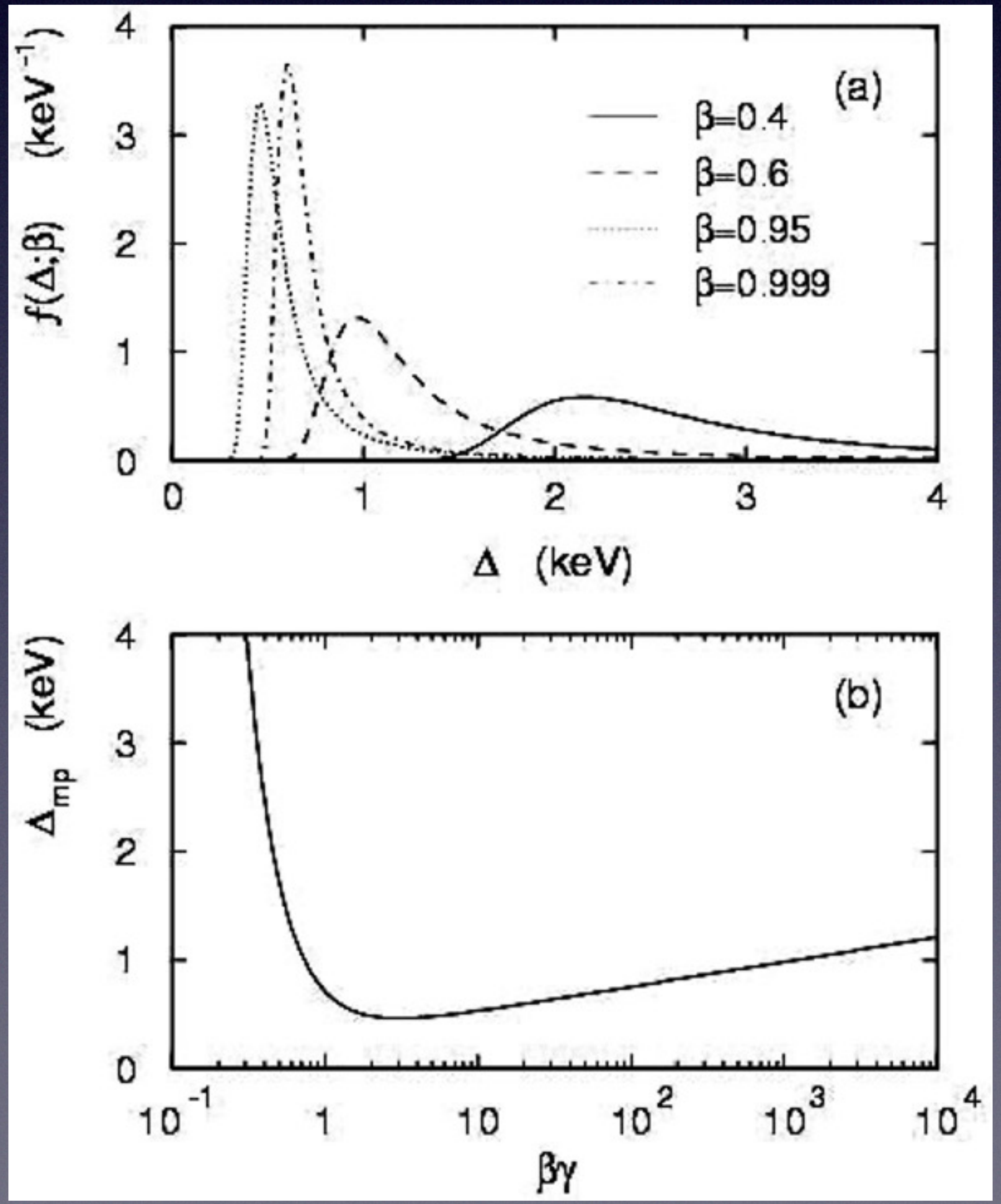
- For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau PDF

$$f(\Delta;\beta) = \frac{1}{\xi}\phi(\lambda)$$

$$\phi(\lambda) = \frac{1}{\pi}\int_0^{\infty} e^{-t\ln t-\lambda t} sin(\pi t)dt$$

$$\lambda = \frac{1}{\xi}[\Delta - \Delta_0]$$





The Monte Carlo method

- A numerical technique for calculating probabilities and related quantities using sequences of random numbers

- Usual steps

  1. Generate random numbers $r_1,\ r_2, \ldots, r_m$ uniform betwen 0 and 1

  2. Use these random numbers to produce another sequence $x_1,\ x_2, \ldots,\ x_m$ that follows some PDF $f(x)$

  3. Use these x values to estimate some property of $f(x)$

     A. Integration between $a < x < b$: $\int_a^b f(x)dx$

     B. We can also use MC for testing statistical procedures

Random number generator

- Goal is to generate uniformly distributed values in [0,1]
- We can achieve this goal manually by tossing coin
  - e.g. 32 bit number: toss 32 times to get one number (too tiring)
  - This would be truely random number
- How do we implement this in computer algorithem to generate $r_1, r_2, \ldots, r_n$

multiplicative linear congruential generator (MLCG) $\longrightarrow$

- The sequence is periodic!

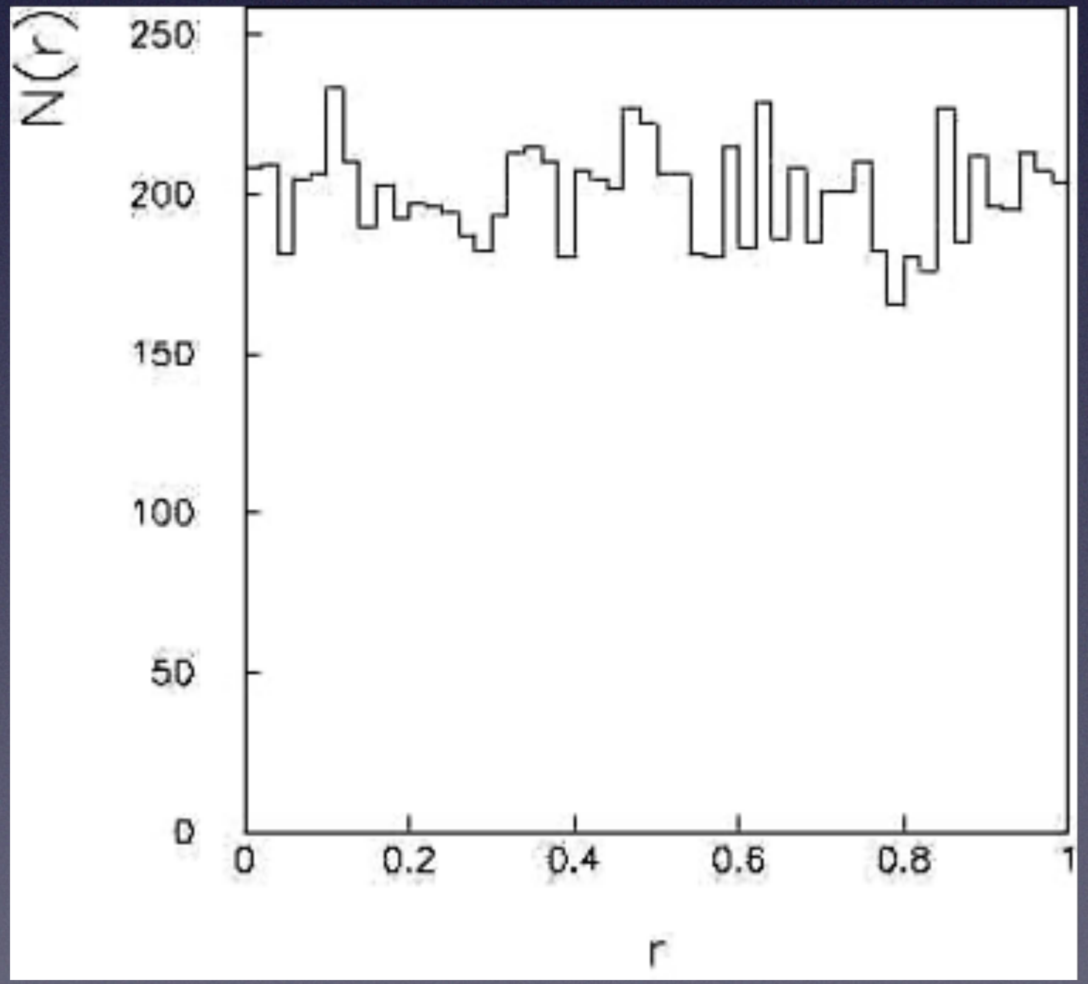
a=3, m=7, n_0=1

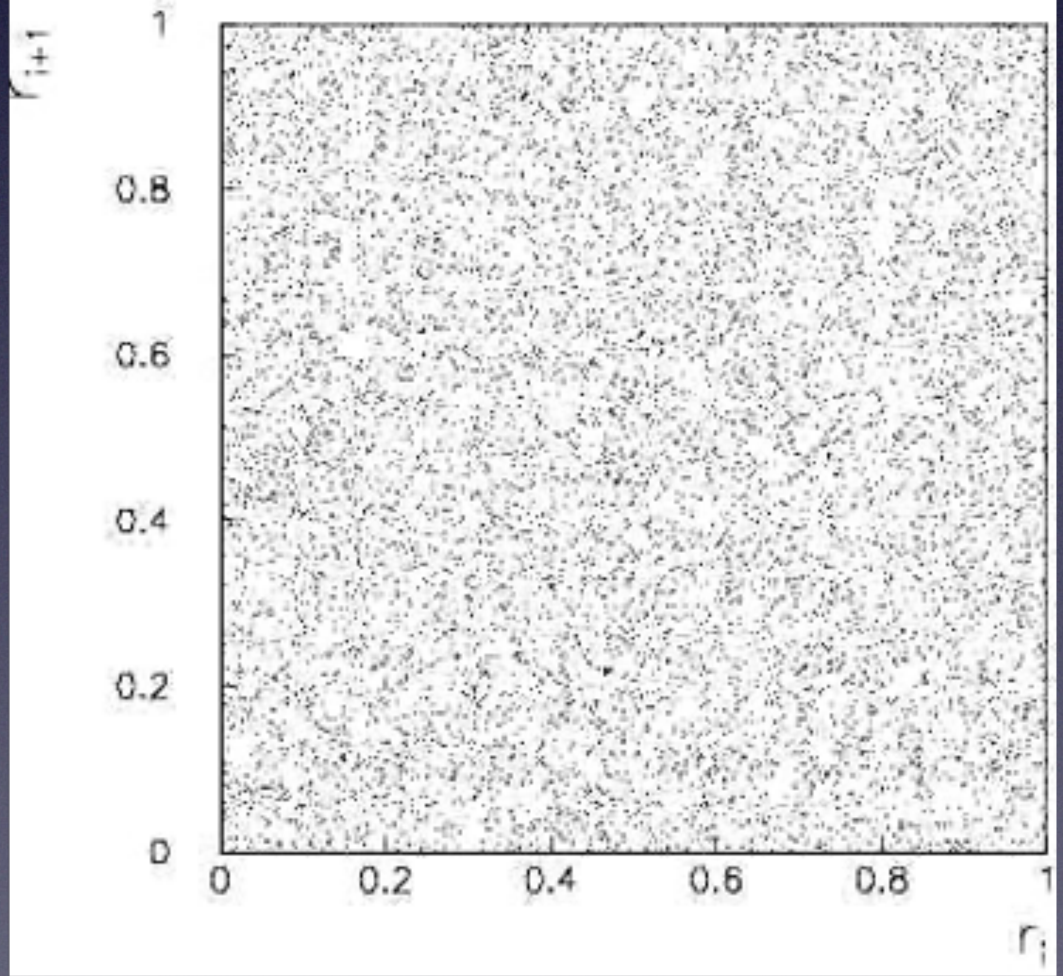
$$\begin{aligned} n_1 &= (3 \cdot 1) \bmod 7 = 3 \\ n_2 &= (3 \cdot 3) \bmod 7 = 2 \\ n_3 &= (3 \cdot 2) \bmod 7 = 6 \\ n_4 &= (3 \cdot 6) \bmod 7 = 4 \\ n_5 &= (3 \cdot 4) \bmod 7 = 5 \\ n_6 &= (3 \cdot 5) \bmod 7 = 1 \end{aligned}$$

- Choose a, m to obtain long period
- Still only use a subset of a single period of the sequence

Random number generator

- Can we say the variable $r_i = n_i/m$ are random in [0,1]?
- uniform distribution in [0,1] and all values are independent as long as subset is used
- a=40692, m=2147483399 (maximum period: $m-1$)

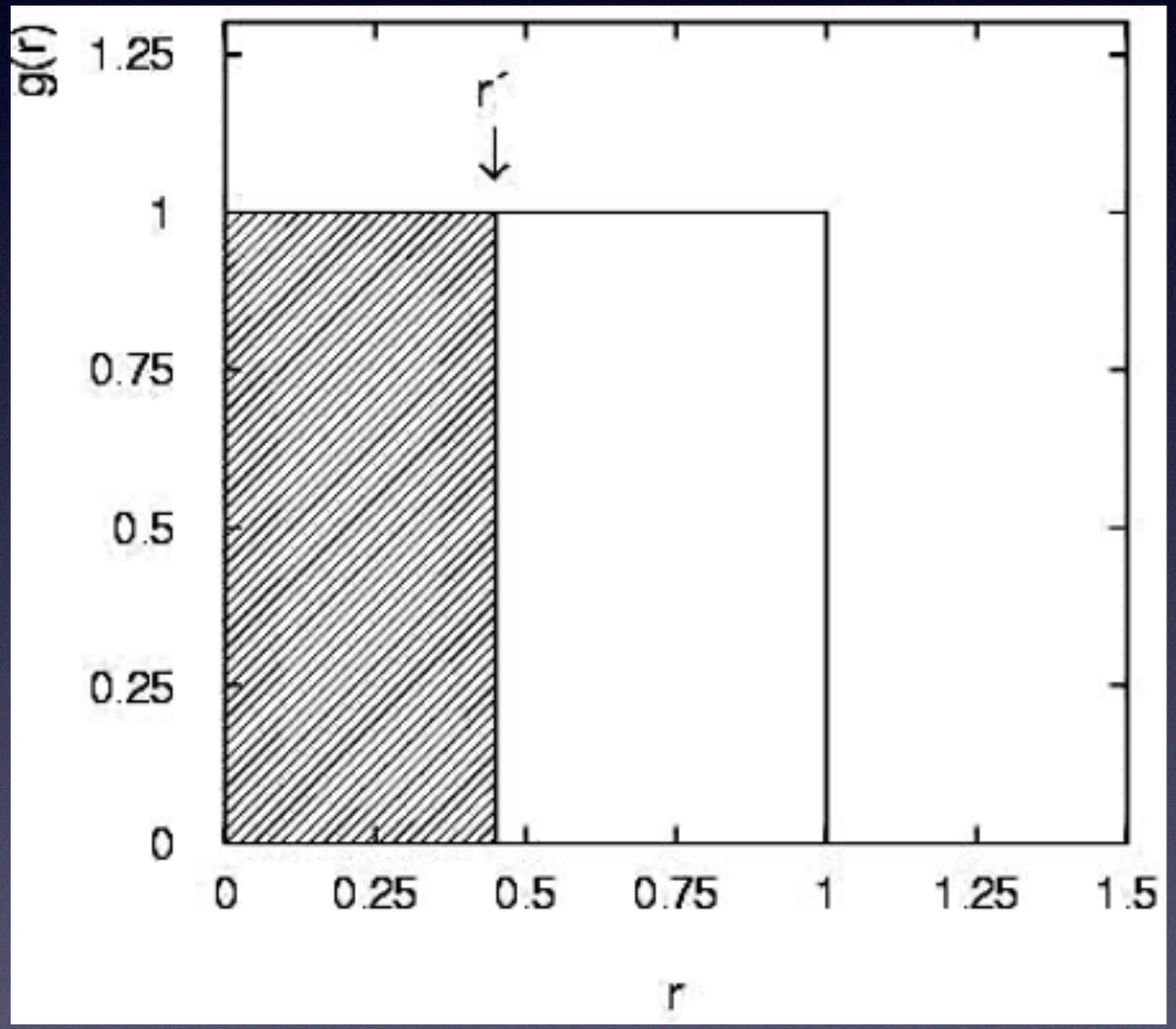


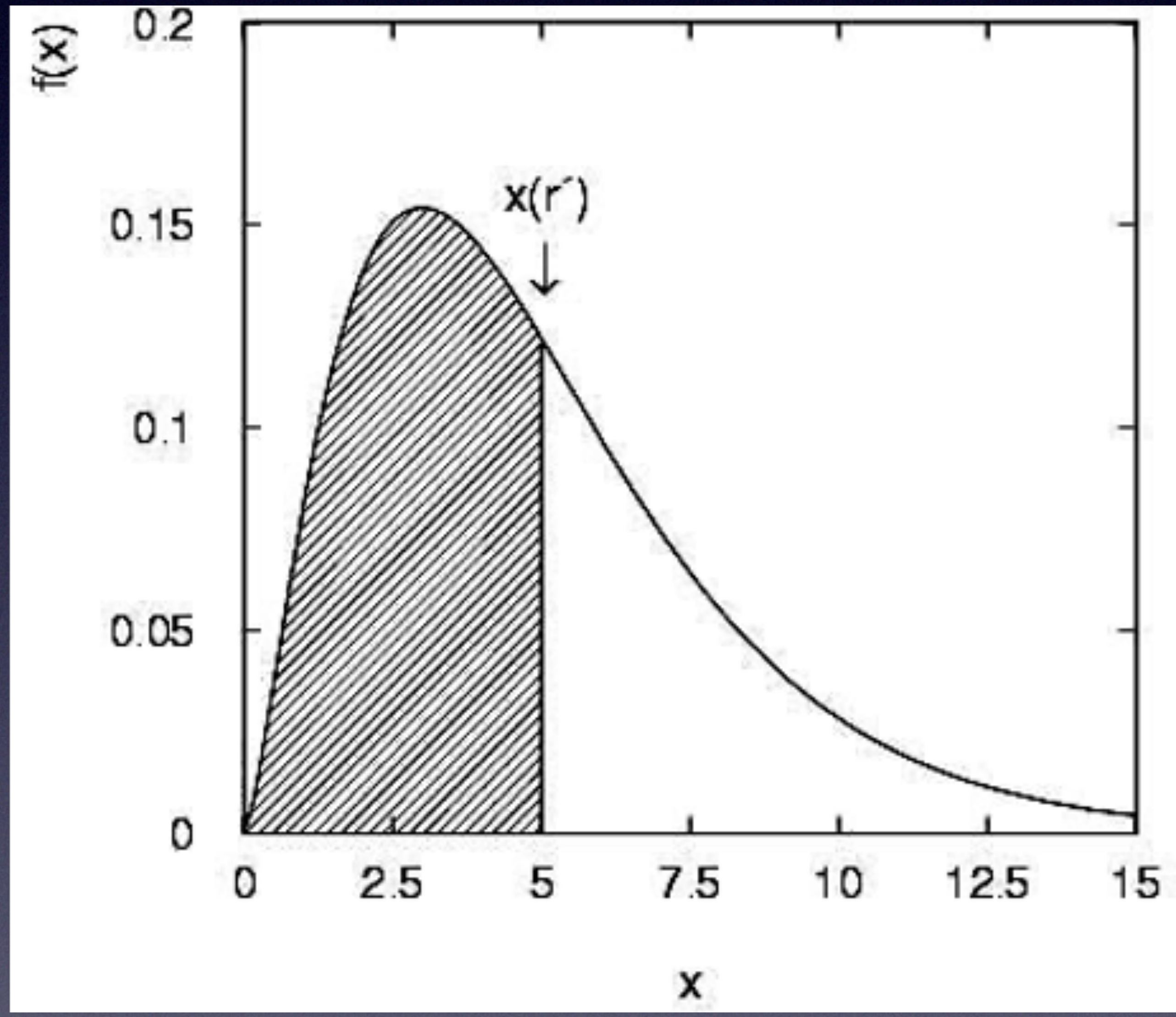


- Far better algorithm is available with TRandom3 in ROOT, period $2^{19937}-1$

Transformation method

- Given $r_1, r_2, \ldots, r_n$ uniform in [0,1], find $x_1, x_2, \ldots, x_n$ that follow $f(x)$ by finding a suitable transformation $x(r)$





- Use the fact that the probability should be the same $P(r \le r') = P(x \le x(r'))$

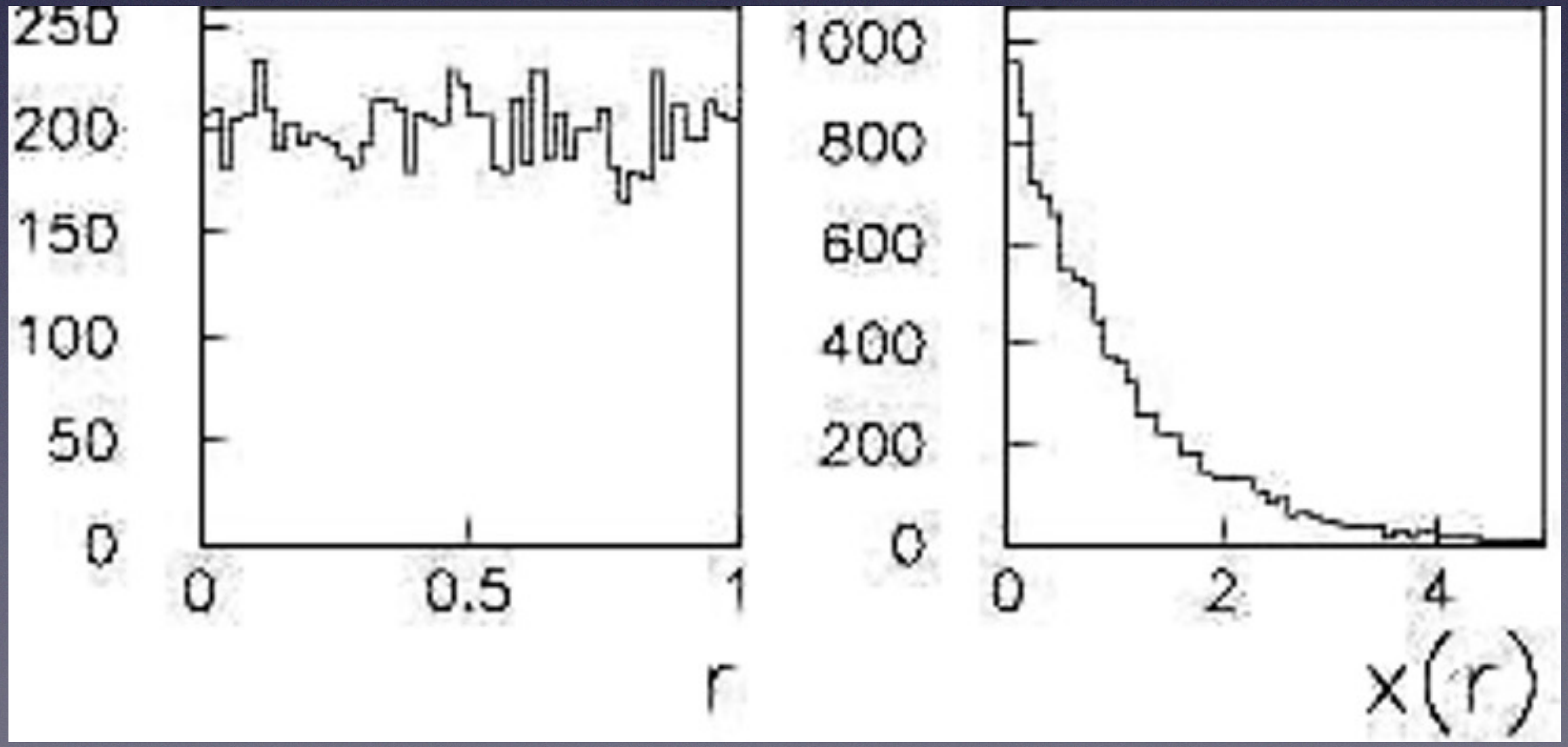
$$\boxed{r' = \int_{-\infty}^{x(r')} f(x')dx' = F(x)}$$

transformation method

- Exponential PDF: $f(x;\xi) = \frac{1}{\xi} e^{-x/\xi} \quad (x \geq 0)$

from the previous slide, the condition of $\int_0^x \frac{1}{\xi} e^{-x'/\xi} dx' = r$

$$\longrightarrow \boxed{x = -\xi \ln r}$$

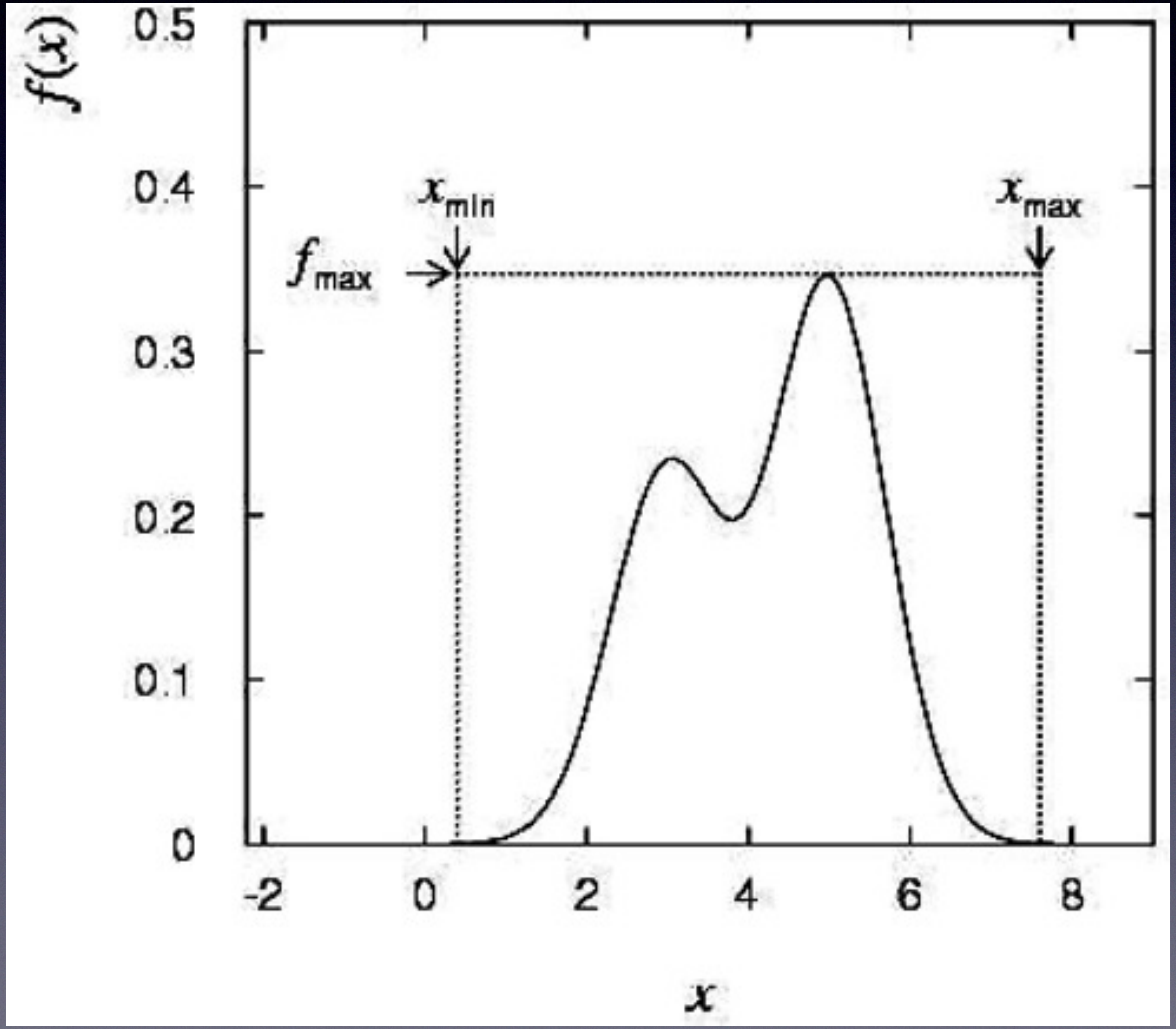


The acceptance-rejection method

1. Generate a random number x, uniform in $[x_{min}, x_{max}]$, i.e. $x = x_{min} + r_1(x_{max} - x_{min})$ where r_1 is random number in [0,1]

2. Generate 2nd independent random number y uniformly distributed between 0 and f_{max} i.e. $y = r_2 f_{max}$

3. If $y > f(x)$, then accept x, If not, reject x and repeat

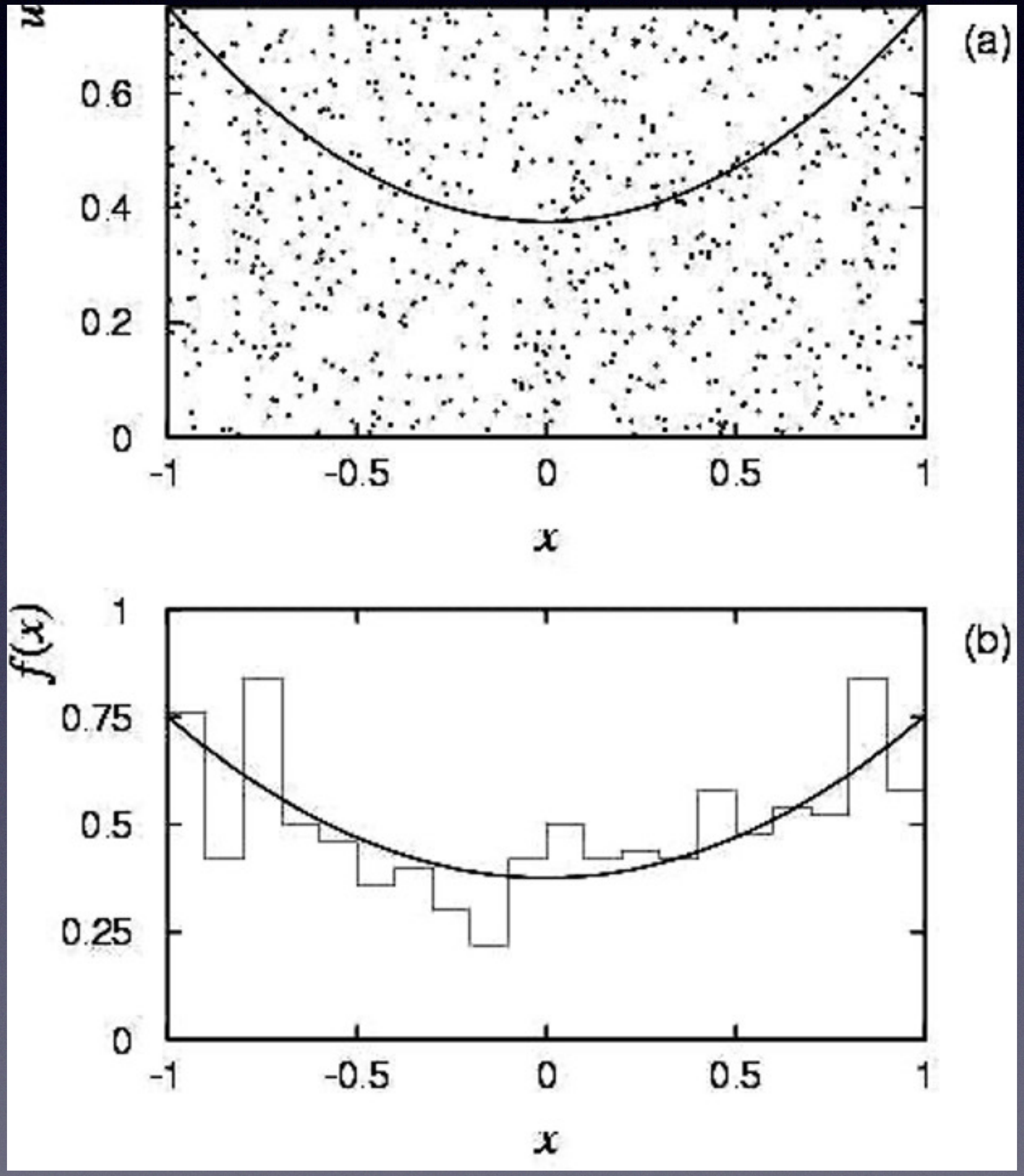


Example with acceptance-rejection method

$$f(x) = \frac{3}{8}(1 + x^2)$$

$$(-1 \leq x \leq 1)$$

- If dot below curve, use x value in histogram



Monte Carlo event generators

- Simple example: $e^+e^- \rightarrow \mu^+\mu^-$
- Generate $\cos\theta$ and ϕ

$$f(\cos\theta; \mathrm{A_{FB}}) \propto (1 + \frac{8}{3}\mathrm{A_{FB}}\cos\theta + \cos^2\theta)$$

$$g(\phi) = \frac{1}{2\pi}(0 \leq \phi \leq 2\pi)$$

- Less simple: event generators for a variety of reactions: ($pp \rightarrow$ hadrons, DY, SUSY)
  - PYTHIA, HERWIG,…
- Output = events; for each event, we get a list of generated particles and their momentum vectors, type, etc

Example: a simulated event

-------- End PYTHIA Info Listing ------------------------------------

-------- PYTHIA Event Listing (hard process) --

no	id	name	status	mothers		daughters		colours		p_x	p_y	p_z	e	m
0	90	(system)	-11	0	0	0	0	0	0	0.000	0.000	0.000	13000.000	13000.000
1	2212	(p+)	-12	0	0	3	0	0	0	0.000	0.000	6500.000	6500.000	0.938
2	2212	(p+)	-12	0	0	4	0	0	0	0.000	0.000	-6500.000	6500.000	0.938
3	21	(g)	-21	1	0	5	6	503	502	0.000	0.000	466.997	466.997	0.000
4	21	(g)	-21	2	0	5	6	501	503	-0.000	-0.000	-173.068	173.068	0.000
5	6	(t)	-22	3	4	7	8	501	0	-105.493	47.504	364.981	420.149	173.000
6	-6	(tbar)	-22	3	4	9	10	0	502	105.493	-47.504	-71.053	219.916	173.000
7	24	(W+)	-22	5	0	11	12	0	0	-5.406	0.381	234.618	249.107	83.542
8	5	b	23	5	0	0	0	501	0	-100.087	47.123	130.363	171.043	4.800
9	-24	(W-)	-22	6	0	13	14	0	0	36.525	18.640	-80.971	123.723	84.080
10	-5	bbar	23	6	0	0	0	0	502	68.969	-66.144	9.918	96.193	4.800
11	-1	dbar	23	7	0	0	0	0	504	-36.167	12.183	57.925	69.368	0.330
12	2	u	23											
13	3	s	23											
14	-4	cbar	23											

Charge su

-------- End PYTHIA Event Listing --------

-------- PYTHIA Event Listing (complete event) --

no	id	name	status	mothers		daughters		colours		p_x	p_y	p_z	e	m
0	90	(system)	-11	0	0	0	0	0	0	0.000	0.000	0.000	13000.000	13000.000
1	2212	(p+)	-12	0	0	654	0	0	0	0.000	0.000	6500.000	6500.000	0.938
2	2212	(p+)	-12	0	0	655	0	0	0	0.000	0.000	-6500.000	6500.000	0.938
3	21	(g)	-21	7	7	5	6	503	502	0.000	0.000	466.997	466.997	0.000
4	21	(g)	-21	8	0	5	6	501	503	-0.000	-0.000	-173.068	173.068	0.000
5	6	(t)	-22	3	4	9	9	501	0	-105.493	47.504	364.981	420.149	173.000
6	-6	(tbar)	-22	3	4	10	10	0	502	105.493	-47.504	-71.053	219.916	173.000
7	21	(g)	-42	12	0	3	3	503	502	0.000	0.000	466.997	466.997	0.000
8	21	(g)	-41	13	13	11	4	501	506	-0.000	-0.000	-242.115	242.115	0.000
9	6	(t)	-44	5	5	14	14	501	0	-108.783	32.916	278.766	347.213	173.000
10	-6	(tbar)	-44	6	6	15	15	0	502	88.142	-124.441	-106.839	254.162	173.000
11	21	(g)	-43	8	0	16	16	503	506	20.641	91.525	52.955	107.737	0.000
12	21	(g)	-41	18	0	17	7	507	502	-0.000	-0.000	1062.750	1062.750	0.000
13	21	(g)	-42	19	19	8	8	501	506	0.000	0.000	-242.115	242.115	0.000
14	6	(t)	-44	9	9	20	20	501	0	-126.871	-1.368	280.882	353.443	173.000
15	-6	(tbar)	-44	10	10	21	21	0	502	83.885	-132.510	-107.973	257.260	173.000
16	21	(g)	-44	11	11	22	22	503	506	15.998	82.724	59.618	103.216	0.000
17	21	(g)	-43	12	0	23	23	507	503	26.987	51.154	588.108	590.945	0.000
18	1	(d)	-41	28	28	24	12	507	0	-0.000	0.000	1697.748	1697.748	0.000
19	21	(g)	-42	29	0	13	13	501	506	0.000	0.000	-242.115	242.115	0.000
20	6	(t)	-44	14	14	30	30	501	0	-128.407	-7.574	280.952	354.131	173.000
21	-6	(tbar)	-44	15	15	25	26	0	502	83.524	-133.971	-108.766	258.231	173.000
22	21	(g)	-44	16	16	32	32	503	506	15.604	81.131	60.580	102.449	0.000
23	21	(g)	-44	17	17	33	33	507	503	24.132	39.619	589.319	591.142	0.000
24	1	(d)	-43	18	0	27	27	502	0	5.148	20.795	633.548	633.910	0.330
25	-6	(tbar)	-51	21	0	31	31	0	508	81.168	-135.687	-107.600	257.887	173.000
26	21	(g)	-51	21	0	35	35	508	502	2.384	1.827	2.229	3.740	0.000
27	1	(d)	-52	24	24	34	34	502	0	5.120	20.684	630.153	630.514	0.330
28	1	(d)	-42	86	0	18	18	507	0	0.000	-0.000	1697.748	1697.748	0.000
29	21	(g)	-41	87	87	36	19	501	509	-0.000	0.000	-260.112	260.112	0.000
30	6	(t)	-44	20	20	88	88	501	0	-129.051	-5.592	281.299	354.603	173.000
31	-6	(tbar)	-44	25	25	89	89	0	508	77.953	-125.785	-112.276	253.837	173.000
32	21	(g)	-44	22	22	37	38	503	506	15.236	82.266	62.480	104.419	0.000

- PYTHIA Monte Carlo

$$pp \to t\bar{t}$$

Monte Carlo detector simulation

- Take the particle list and momenta from generator as input
- Simulates detector response:
    - multiple Coulomb scattering (generate scattering angle)
    - particle decays (generate lifetime)
    - ionization energy loss (generate Δ)
    - electromagnetic and hadronic showers
    - production of signal, electronics response,…
- Output = simulated raw data
    - input to reconstruction algorithm: tracking finding, fitting, etc
- We usually compare distributions at detector level given a certain hypothesis with the real data
- Also can estimate acceptance = #events found/#events generated
- Programming package: GEANT4