Probability Distributions and Monte Carlo Techniques

Common probability distributions

Binomial, Poisson, Gaussian, Exponential

Sums and differences

Characteristic functions
Central Limits Theorem

Generation of random distributions

Elton S. Smith, Jefferson Lab

Selected references

- Particle Data Group
  - http://pdg.lbl.gov/2009/reviews/contents_sports.html
- CDF Statistics Committee
  - http://www-cdf.fnal.gov/physics/statistics/
- G. Cowan, *Statistical Data Analysis* (1998).
- F. James, *Statistical Methods in Experimental Physics*, 2nd ed. 2006 (W.T. Eadie et al.1971)
- R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences* (2002)
- H. Cramer, *Mathematical Methods of Statistics* (1946)

Uncertainties

- Theory of quantum mechanics is not deterministic
  - Present even for “perfect” measurements
  - Example: Lifetime of a radioactive nucleus
- Random measurement uncertainties or “errors”
  - Present even without quantum effects
  - Example: limited accuracy of measurement
- Things we know in principle, but don’t in practice
  - Example: uncontrolled parameters during measurement

We quantify all these uncertainties using the concept of PROBABILITY

Interpretation of probability

- Relative frequency (classical)
  - If 'A' is the outcome of a repeatable experiment

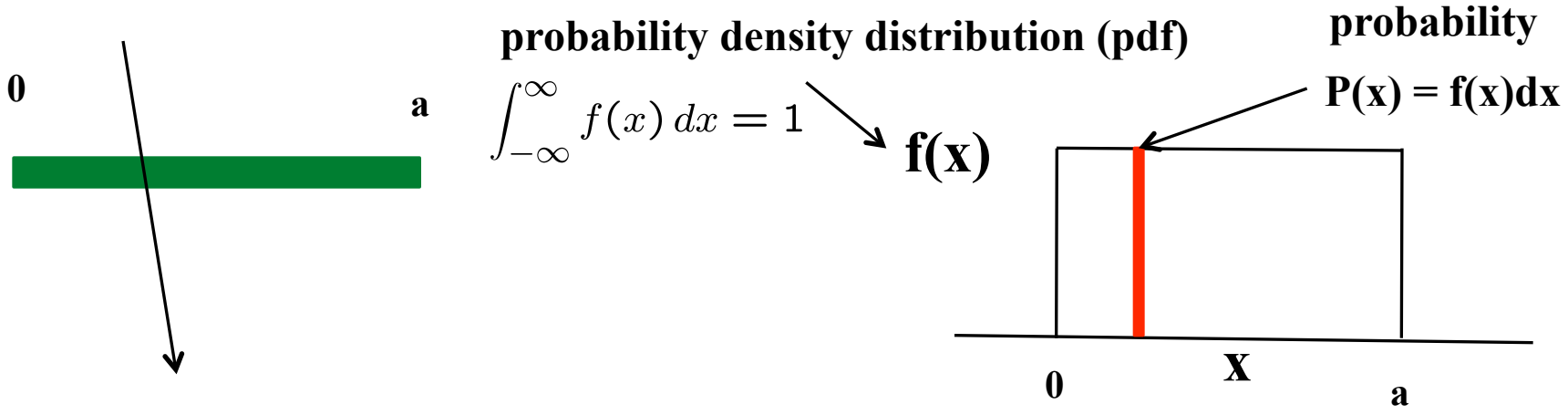
$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A}{n}$$

- Subjective probability (Bayesian)
  - If 'A' is a hypothesis (statement that is true or false)

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

- In particle physics the classical or 'frequentist' interpretation is most common, but the Bayesian approach can be useful for non-repeatable phenomena, e.g. probability the Higgs boson exists

Uniform random variable



- Particle counter detects a particle if it hits anywhere over its sensitive length ‘a’.
- Absent any knowledge about the source of particles, we would assign a uniform probability distribution to the position *x*, of the particle whenever the counter registers a hit.

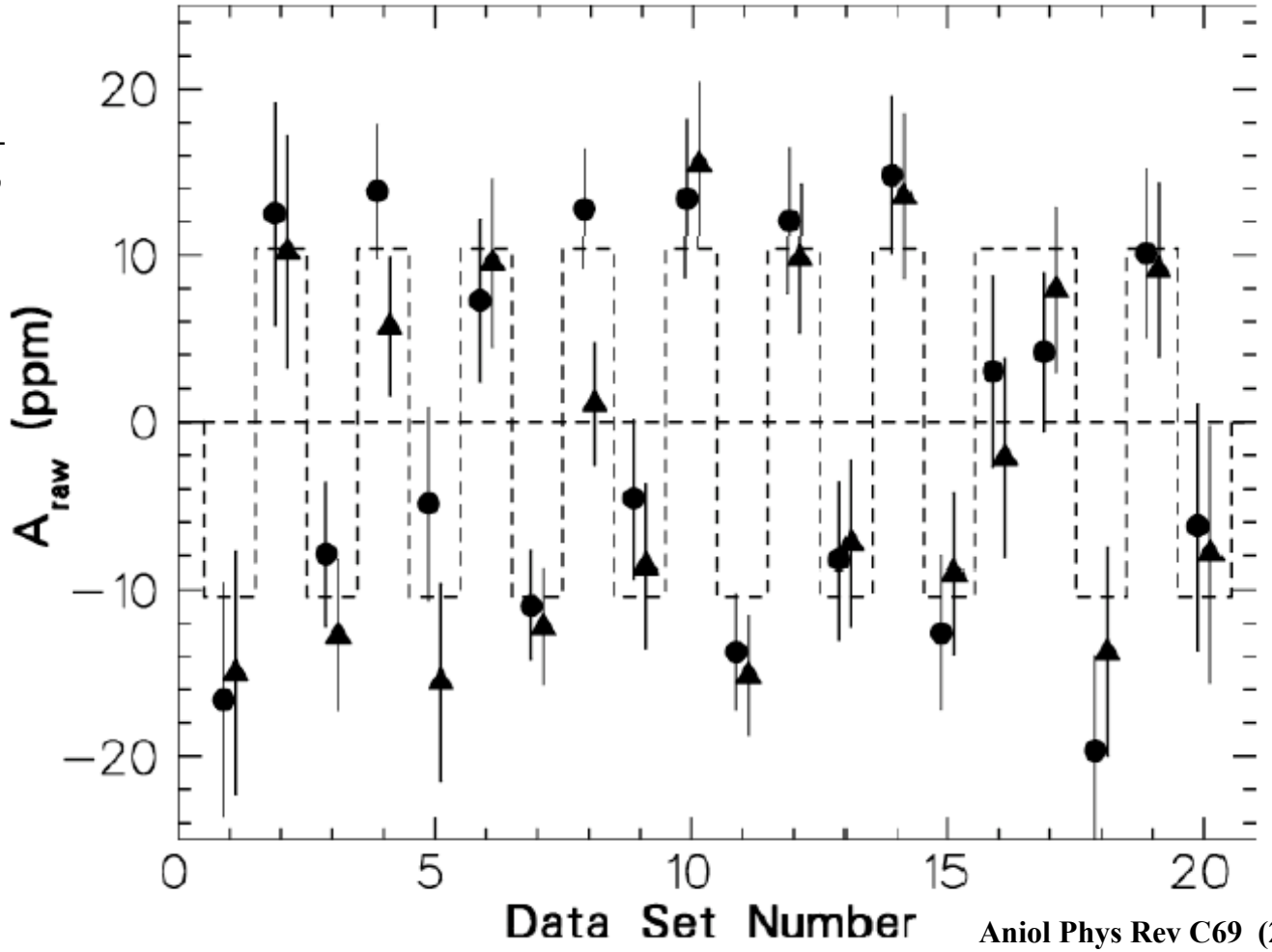
Experimental uncertainties

- Assume a measurement is made of a quantity μ. Let x be the value of a single measurement and this measurement has an uncertainty σ.
- Now, assume this measurement is repeated many times, and assume that each measurement is made independently of other measurements. In that case, the measurement x can be considered a random variable, with a probability distribution that will be centered on μ, but with a value that differs from μ by an amount that is approximately σ. But what is this probability distribution?

Distribution of measurements

Raw asymmetries for 1999 HAPPEX running period, in ppm, broken down by data set. Circles are for left spectrometer, triangles for right. Dashed line is the average for the entire run.

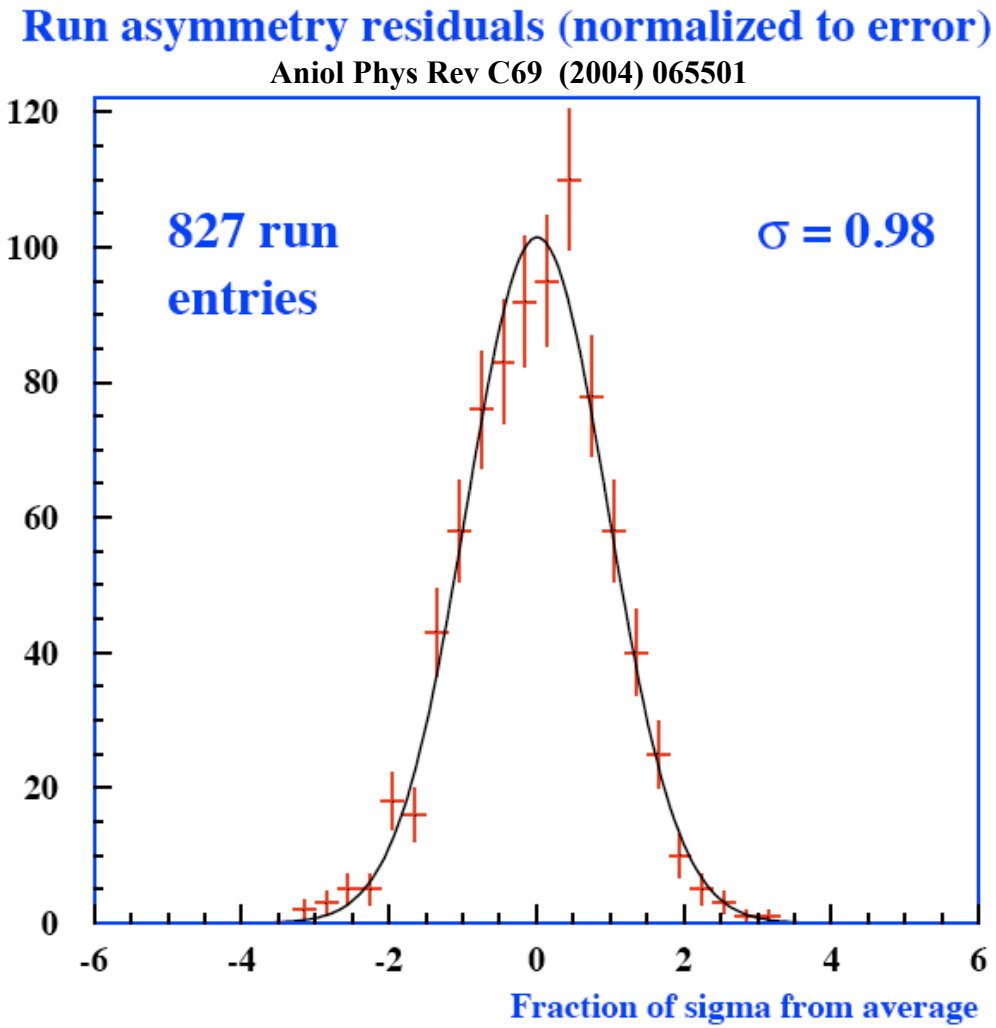
$$A_{raw} = \frac{\vec{e}p - \overleftarrow{e}p}{\vec{e}p + \overleftarrow{e}p}$$



Aniol Phys Rev C69 (2004) 065501

Distribution of experimental measurements

Run asymmetries for 1999 HAPPEX running period, with mean subtracted off and normalized by statistical error



Gaussian (Normal) distribution

- Typical of experimental random uncertainties

$$F(a) = \int_{-\infty}^{a} f(x;\mu,\sigma)\,dx$$

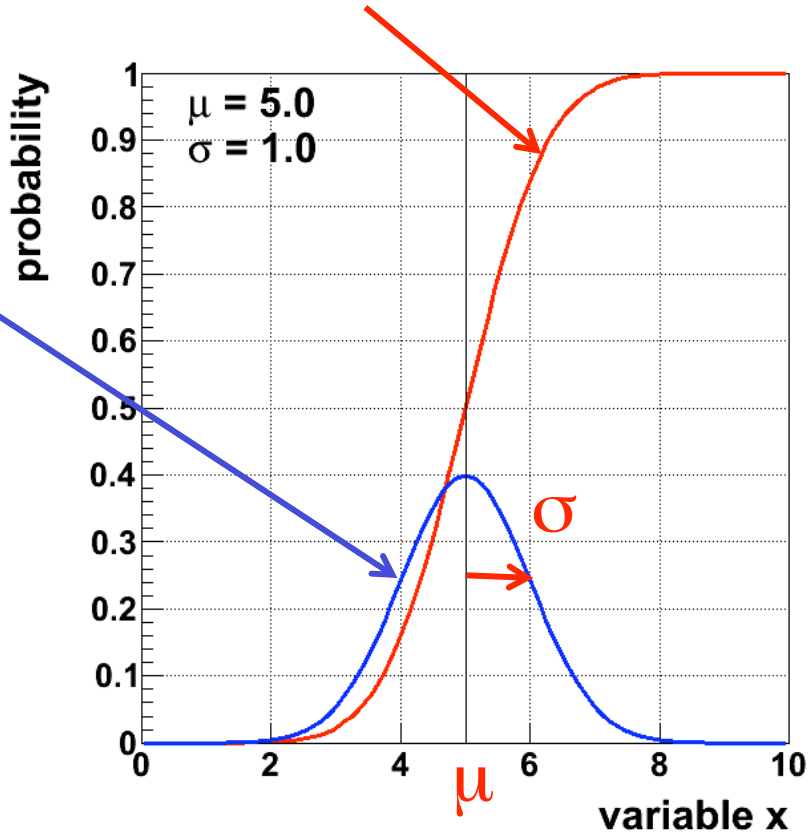
Cumulative distribution function

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

(units ~ 1/σ)

$$P = \int_{-\infty}^{\infty} f(x;\mu,\sigma)\,dx = 1$$

(dimensionless)



Named "standard Gaussian" when μ=0 and σ=1

Moments: Defined for all distributions

Expectation value of x

$$E[x] = a_1 = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

n^{th} moment of a random variable x

$$\alpha_n = E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

n^{th} central moment of a random variable x

$$m_n = E[(x - \alpha_1)^n]$$

mean

$$\mu \equiv \alpha_1$$

variance

$$\sigma^2 \equiv V[x] \equiv m_2 = \alpha_2 - \mu^2$$

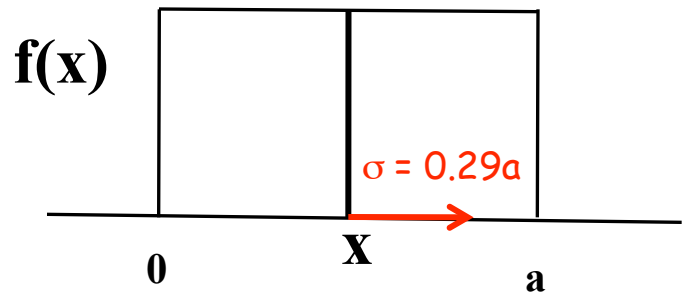
Same units!

root-mean-square

$$\sigma \equiv \sqrt{V[x]}$$

Example: uniform distribution

$$f(x) = \frac{1}{a}$$



$$\mu \equiv \alpha_1 = \int_0^a x f(x)\,dx = \frac{1}{a}\int_0^a x\,dx = \frac{a}{2}$$

$\mu = a/2$

$$\sigma^2 \equiv m_2 = \alpha_2 - \mu^2 = \frac{1}{a}\int_0^a x^2 dx - \frac{a^2}{4} = \frac{a^2}{12}$$

$\sigma = a/\sqrt{12}$

For a Gaussian distribution

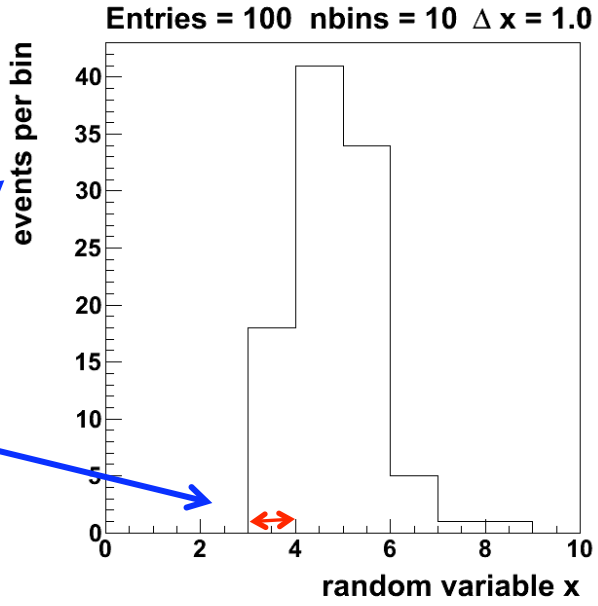
μ = "μ"

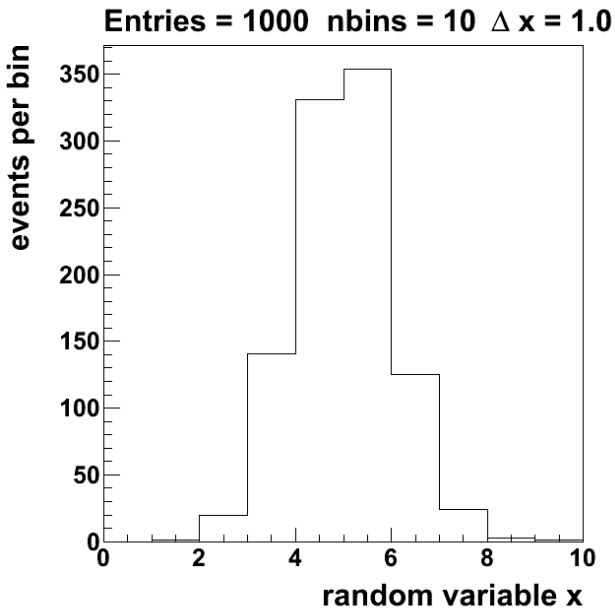
σ = "σ"

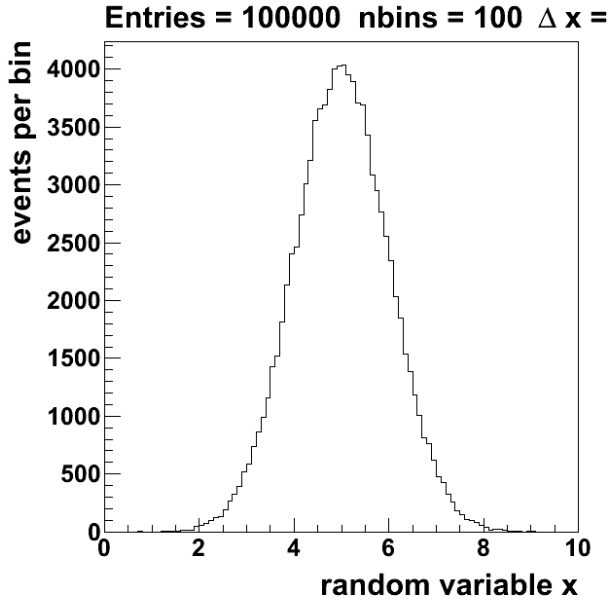
Histograms: representation of pdfs using data

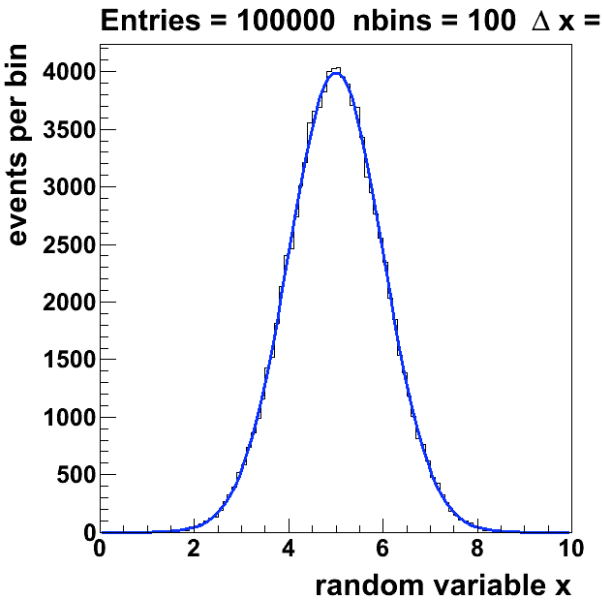
$$f(x) = \frac{N_x}{N\,\Delta x}$$

normalization









Discrete distributions: binomial

- Biased coin toss
  - N trials
  - probability of success p
  - probability of failure (1-p)
  - $0 \le p \le 1$

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

random variable

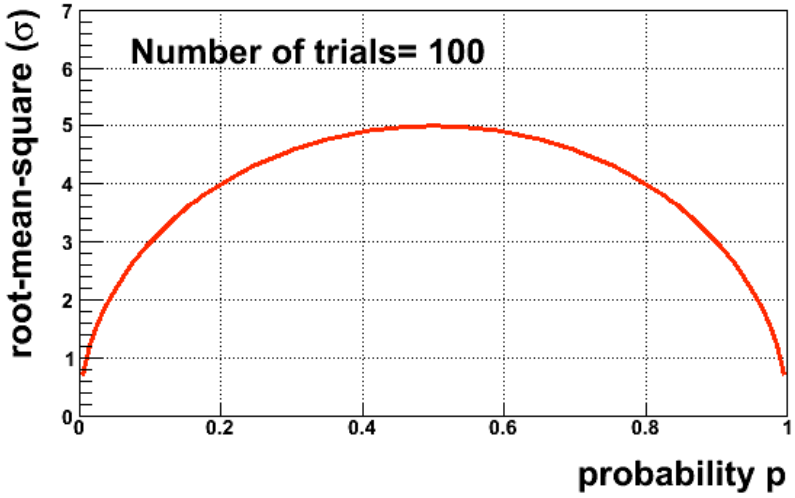
parameters

mean

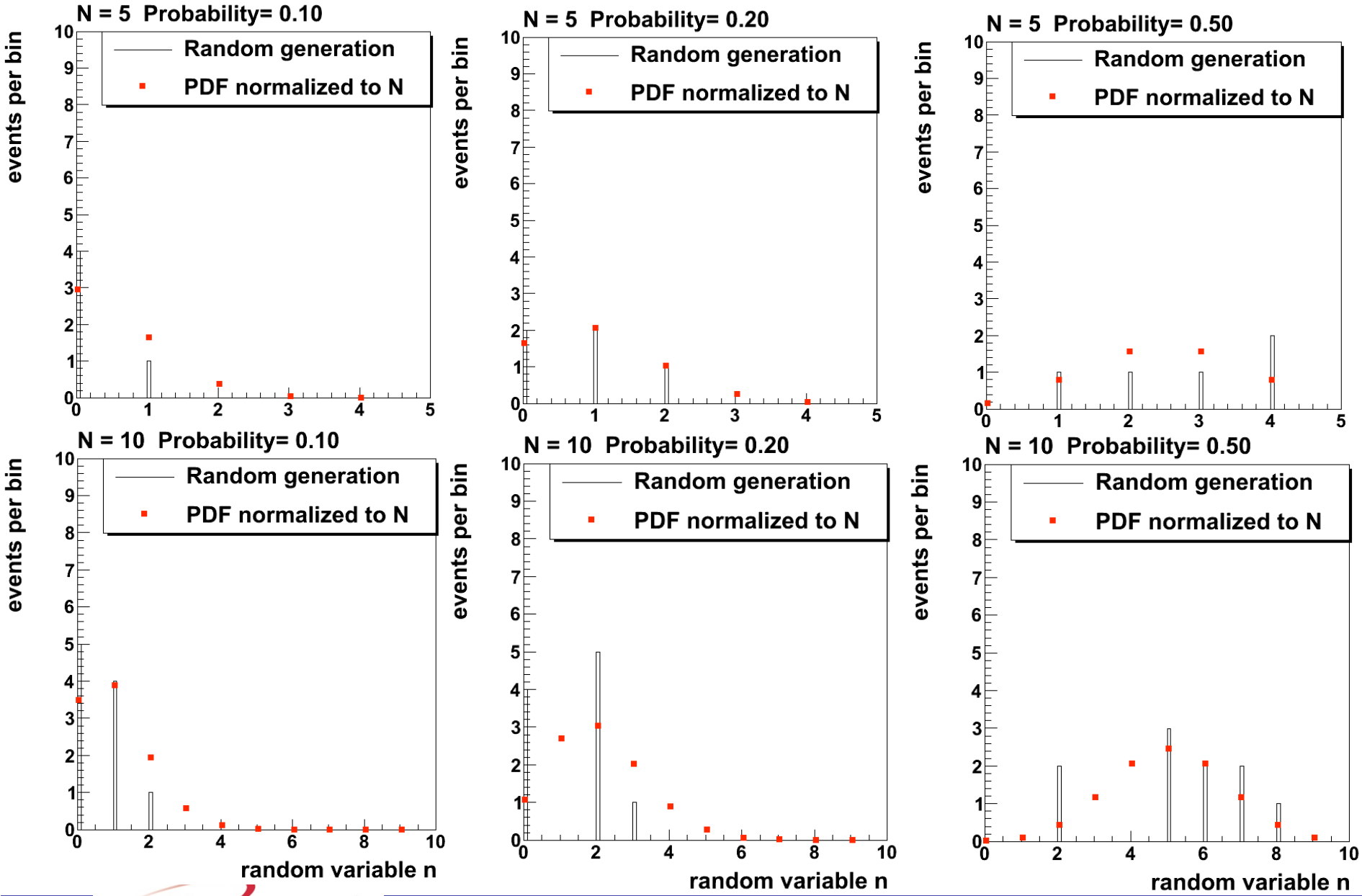
$$E[n] = \sum_{n=0}^{N} n f(n;N,p) = Np$$

variance

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$



Binomial distribution examples



Discrete distributions: Poisson

- Parent distribution for counting experiments
- Limiting case of binomial distribution
  - Limit for $p \to 0$
  - Mean successes $Np \to \mu$
  - $n \geq 0$

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

random variable

parameter

mean

$$E[n] = \sum_{n=0}^{\infty} n \frac{\mu^n}{n!} e^{-\mu} = \mu$$

$\sigma = \sqrt{\mu}$

variance

$$V[n] = E[n^2] - (E[n])^2 = \mu$$

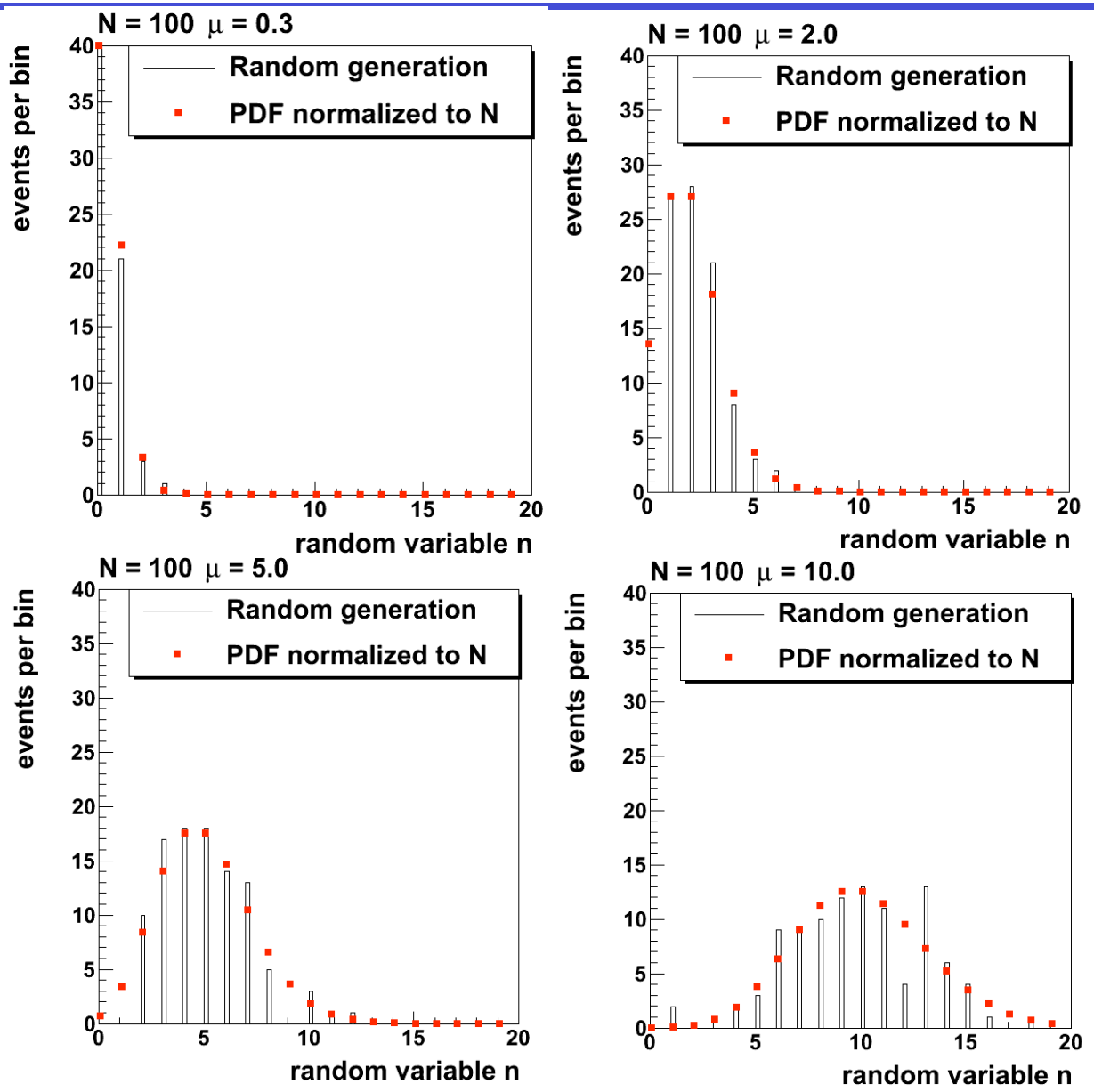
$\sigma/\mu = 1/\sqrt{\mu}$

Poisson distribution examples

The mean μ can be any positive real number

For small values of μ, there is a significant probability n=0

The distribution approaches Gaussian for values of $\mu \geq 10$



Counting experiments

- The Poisson approximation is appropriate for counting experiments where the data represent the number of items observed per unit time.
- Example: 10 nA beam (10^{11} particles/s) produce 10^4 interactions/s, i.e p ~ 10^{-7}
- Here $\mu = 10^4$ (1 s of data), and $\sigma = \sqrt{\mu} = 10^2$
- The uncertainty σ is called the *statistical* error or uncertainty
- Note also: 10% chance to get zero events is

$$f(0;\mu) = e^{-\mu} = 0.1 \qquad \mu_{10} = -\ln(0.1) = 2.3$$

Another continuous pdf: exponential

- Proper decay time for unstable particle
- Population growth

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

random variable

parameter

mean

$$E[x] = \mu$$

variance

$$V[x] = \mu^2$$

- Lack of memory:
  - $f(t\text{-}t_0 \mid t \geq t_0) = f(t)$

$\sigma = \mu$

Characteristic functions

$$\phi(u) = E[e^{iux}] = \int_{-\infty}^{\infty} e^{iux} f(x)\, dx$$

$$\alpha_n = \int_{-\infty}^{\infty} x^n f(x)\, dx = i^{-n} \left. \frac{d^n \phi}{du^n} \right|_{u=0}$$

$\phi_1(u)$ <--> $f_1(x)$

$\phi_2(u)$ <--> $f_2(y)$

Form a new random variable z = ax + by

For independent variables x and y, $f(x,y) = f_1(x)\, f_2(y)$

Then $\phi(u) = \phi_1(au)\phi_2(bu)$ <--> $g(z)$

Allows computation of pdfs for sums and differences of random variables with known distributions

Example: rules for sums

Gaussian sum of $G(\mu_1,\sigma_1)$ and $G(\mu_2,\sigma_2)$ gives $G(\mu=\mu_1+\mu_2,\sigma=\sqrt{\sigma_1^2+\sigma_2^2})$

$$\phi_1(u) = e^{i\mu_1 u - \frac{1}{2}\sigma_1^2 u^2} \qquad \phi_2(u) = e^{i\mu_2 u - \frac{1}{2}\sigma_2^2 u^2}$$

$$\phi(u) = \phi_1(u)\phi_2(u) = e^{i(\mu_1+\mu_2)u - \frac{1}{2}(\sigma_2^2+\sigma_1^2)u^2} = e^{i\mu u - \frac{1}{2}\sigma^2 u^2}$$

Poisson sum of $P(\mu_1)$ and $P(\mu_2)$ gives $P(\mu=\mu_1+\mu_2)$

$$\phi_1(u) = e^{\mu_1(e^{iu}-1)} \qquad \phi_2(u) = e^{\mu_2(e^{iu}-1)}$$

$$\phi(u) = \phi_1(u)\phi_2(u) = e^{\mu_1(e^{iu}-1)+\mu_2(e^{iu}-1)} = e^{(\mu_1+\mu_2)(e^{iu}-1)}$$

Note: Difference also works for G, not P!

Central Limit Theorem

Any random variable that is the sum of many small contributions (with arbitrary pdfs) has a Gaussian pdf.

x_i are (independent) random variables with means μ_i and variances σ_i^2

Variable of interest is $X = \sum_i^n x_i$

For large n: $E[X] = \sum_i^n \mu_i$ $V[X] = \sum_i^n \sigma_i^2$

Example: Multiple scattering distributions are approximately Gaussian because they result from the sum of many individual scatters

Correlations

Let x,y be two random variables with a joint probability distribution *f(x,y)*

Marginal probability distributions (integrate over y)

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Conditional probability distributions (fix $y=y_0$)

$$f_4(x|y_0) = \frac{f(x,y_0)}{f_2(y_0)} \qquad f_3(y|x_0) = \frac{f(x_0,y)}{f_1(x_0)}$$

Averages

$$\mu_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dx\,dy \qquad \mu_y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\,dx\,dy$$

Covariance

Define covariance cov[x,y] (also use matrix notation V_{xy}) as

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

Correlation coefficient (dimensionless) defined as

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

If x, y, independent, i.e., $f(x, y) = f_x(x) f_y(y)$, then

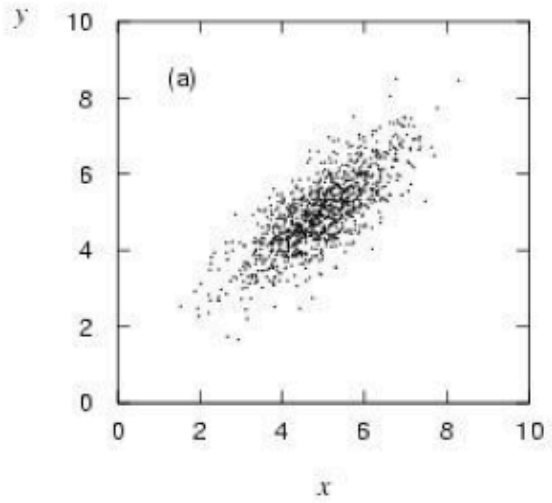
$$E[xy] = \int\int xy\, f(x, y)\, dxdy = \mu_x \mu_y$$

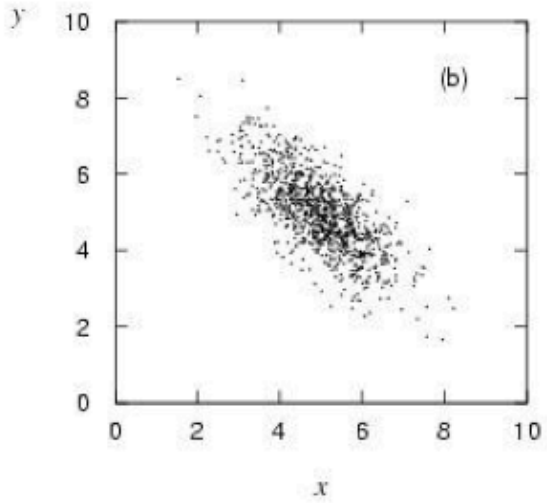
$\rightarrow$ $\mathrm{cov}[x, y] = 0$ x and y, 'uncorrelated'

N.B. converse not always true.

Examples

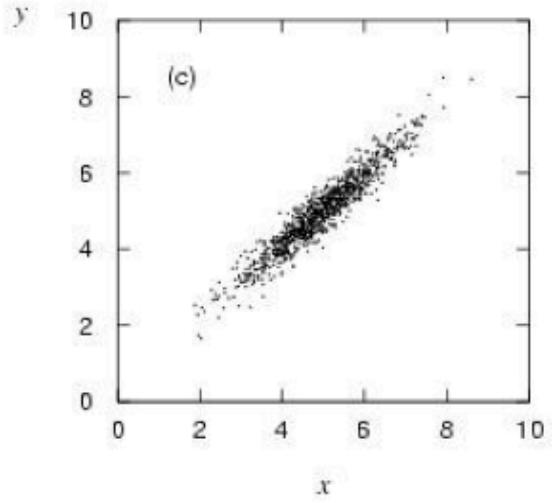
$\rho = 0.75$

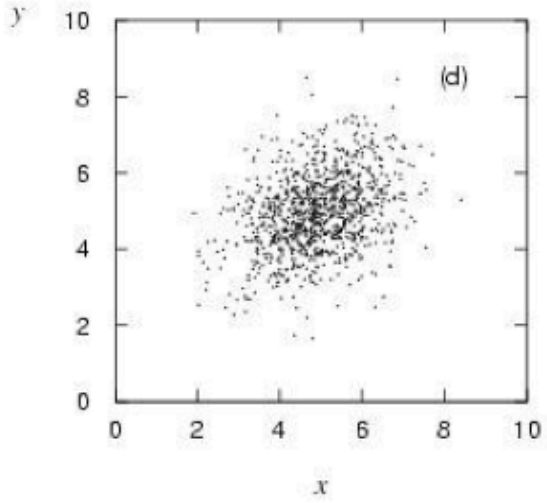




$\rho = -0.75$

$\rho = 0.95$





$\rho = 0.25$

Propagation of ~~errors~~ uncertainties

Physical quantities of interest are often combinations of more than one measurements

sums

$$E[x+y] = E[x] + E[y] \qquad \mu = \mu_x + \mu_y$$

$$V[x+y] = V[x] + V[y] + 2\mathrm{cov}(x,y)$$

$$\sigma^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_x\sigma_y\rho_{xy} \qquad (-1 \le \rho_{xy} \le 1)$$

products

$$E[xy] = E[x]E[y] \qquad \mu = \mu_x\,\mu_y$$

$$\frac{V[xy]}{(xy)^2} = \frac{V[x]}{x^2} + \frac{V[y]}{y^2} + \frac{2\mathrm{cov}(x,y)}{xy}$$

$$\sigma^2/(x^2y^2) = \sigma_x^2/x^2 + \sigma_y^2/y^2 + 2\sigma_x\sigma_y\rho_{xy}/(xy)$$

If x and y are independent, then $\rho_{xy} = 0$

Uncertainty in the average

If x_i are independent measurements of a quantity with mean μ and distributed with the same sigma σ

$$E[x_i] = \mu \qquad V[x_i] = \sigma^2$$

Then the average of x_i

$$E[\frac{1}{N}\sum_1^N x_i] = \frac{1}{N}\sum_1^N \mu_i = \mu \qquad V[\frac{1}{N}\sum_1^N x_i] = \frac{1}{N^2}\sum_1^N \sigma_i^2 = \frac{\sigma^2}{N}$$

average $\longrightarrow$ μ $\qquad$ σ of average $\longrightarrow$ $\sigma/\sqrt{N}$

Application 1

Daniel Kahneman, "Thinking Fast and Slow"

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

1. The larger hospital
2. The smaller hospital
3. About the same (that is, within 5% of each other)

56% of subjects chose option 3, and 22% of subjects respectively chose options 1 or 2. However, according to sampling theory the larger hospital is much more likely to report a sex ratio close to 50% on a given day than the smaller hospital

Application 2

A study of the incidence of kidney cancer in the 3,141 counties of the United States reveals a remarkable pattern. The counties in which the incidence of kidney cancer is lowest are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. Now, what do you make of this information?

1. Republican politics provide protection against kidney cancer.
2. The low cancer rates are directly due to the clean living of the rural lifestyle - no air pollution, no water pollution, access to fresh food without additives.

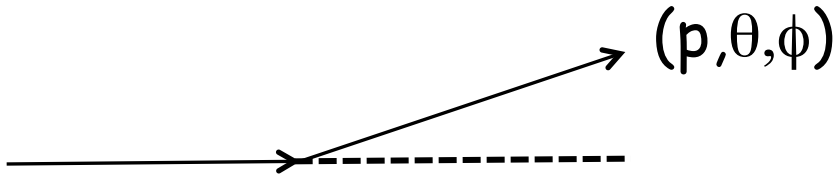
Now consider the counties in which the incidence of kidney cancer is highest. These ailing counties tend to be mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. Do we infer that

1. Republican politics produces high cancer rates.
2. High cancer rates are due to poverty of the rural lifestyle - no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco.

Logic is wrong

Monte Carlo Method

- Numerical technique for computing the distribution of particle interactions
- Each interaction is assumed to be governed by the conservation of energy and momentum and the probabilistic laws of quantum mechanics
- Perfect modeling of the interaction will require the use of correct probability distribution for each variable (e.g. momentum and angles of each particle) including all correlations, although much can be learned with reasonable approximations
- Sequences of random numbers are used to generate Monte Carlo or "simulated" data to be compared to actual measurements. Differences between the true and simulated data can be used to improve understanding the process under study

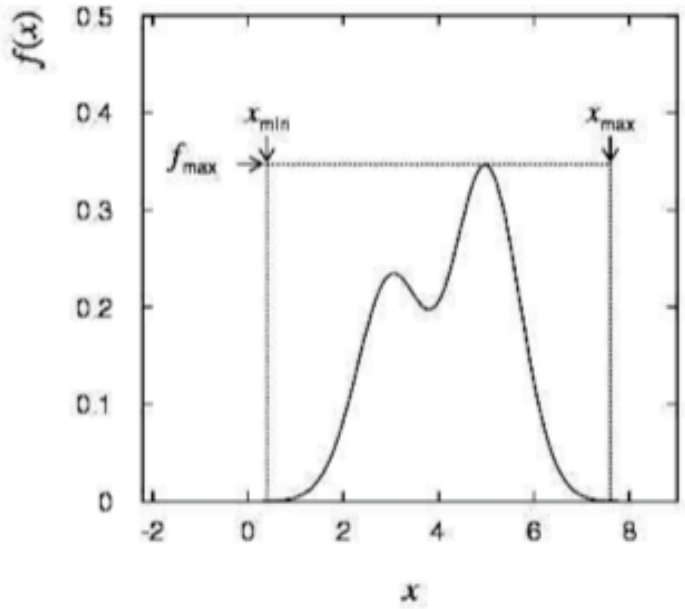


Random number generators

- Computer generates “pseudo-random” numbers, which are deterministic by depend on an input “seed” (often Unix time used)
- Many interactions are simulated
- Each interaction requires the generation a series of random numbers (p, θ and ϕ in the present example)
- Poor random number generators will repeat themselves and/or have periodic correlations (e.g. between the first and third number generated)
- Very good algorithms are available, e.g. **`TRandom3`** from ROOT have a periodicity of 2^{19937}-1.

The acceptance-rejection method

Enclose the pdf in a box:



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

(2) Generate a 2nd independent random number u uniformly distributed between 0 and f_{max}, i.e. $u = r_2 f_{max}$.

(3) If $u < f(x)$, then accept x. If not, reject x and repeat.

The transform method

Uses every single random number produced to generate the distribution f(x) of interest

Integrate distribution $$F(x) = \int_{-\infty}^{x} f(x)dx$$

Normalize distribution $$F_N(x) = \frac{\int_{-\infty}^{x} f(x)dx}{\int_{-\infty}^{\infty} f(x)dx}$$

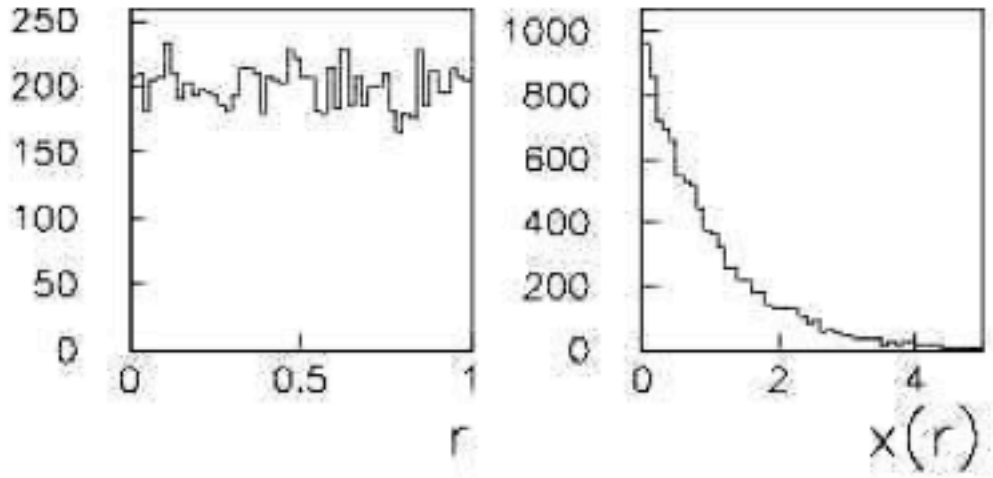
Generate random number r on [0,1]. Then compute $x = F_N^{-1}(r)$. The variable x will be distributed according to the function f(x).

Example of the transform method

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

Set $\int_0^x \frac{1}{\xi} e^{-x'/\xi}\, dx' = r$ and solve for $x\,(r)$.

$\rightarrow \quad x(r) = -\xi \ln(1-r) \qquad (\, x(r) = -\xi \ln r$ works too.)



Many other examples in the Review of Particle Physics

Summary of first lecture

- Defined probability
- Described various common probability distributions
- Demonstrated how new distributions could be generated from combinations of known distributions
- Described the Monte Carlo method for numerically simulating physical processes.
- Next lecture will focus on interpreting data to extract information about the parent distributions, namely statistics.

Backup slides

Variance of sample mean

Population mean = μ, variance = σ $\qquad E(x_i) = E(x_j) = \mu$

$$\bar{x} \equiv \frac{1}{n}\sum x_i \qquad E(\bar{x}) = \frac{1}{n}\sum E(x_i) = \mu$$

$$V(\bar{x}) = E(\bar{x}^2) - E^2(\bar{x}) = E\left(\frac{1}{n}\sum_i x_i \frac{1}{n}\sum_j x_j\right) - E\left(\frac{1}{n}\sum_i x_i\right)E\left(\frac{1}{n}\sum_j x_j\right)$$

$$V(\bar{x}) = \frac{1}{n^2}\left(\sum_i E(x_i^2) + \sum_{i\neq j} E(x_i)E(x_j)\right) - \mu^2 = \frac{1}{n^2}\left(nE(x_i^2) + (n^2-n)\mu^2 - n^2\mu^2\right)$$

$$V(\bar{x}) = \frac{1}{n}\left(E(x_i^2) - E^2(x_i)\right) = \frac{V(x_i)}{n}$$

$$\boxed{V(\bar{x}) = \frac{\sigma^2}{n}}$$

Cramér Math Meth of Statistics 27.2

Expectation value of sample variance

Population mean = μ, variance = σ

$$s^2 \equiv \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$E(s^2) = \frac{1}{n}\sum E(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n}\sum E(x_i^2) - 2E(x_i\bar{x}) + E(\bar{x}^2)$$

$$E(x_i\bar{x}) = E\left(x_i \frac{1}{n}\sum x_j\right) = \frac{1}{n}\sum E(x_i x_j) = \frac{1}{n}E(x_i^2) + \frac{1}{n}\sum_{i\neq j} E(x_i)E(x_j)$$

$$E(x_i) = E(x_j) = \mu \qquad E(x_i\bar{x}) = \frac{1}{n}\left(E(x_i^2) + (n-1)\mu\right)$$

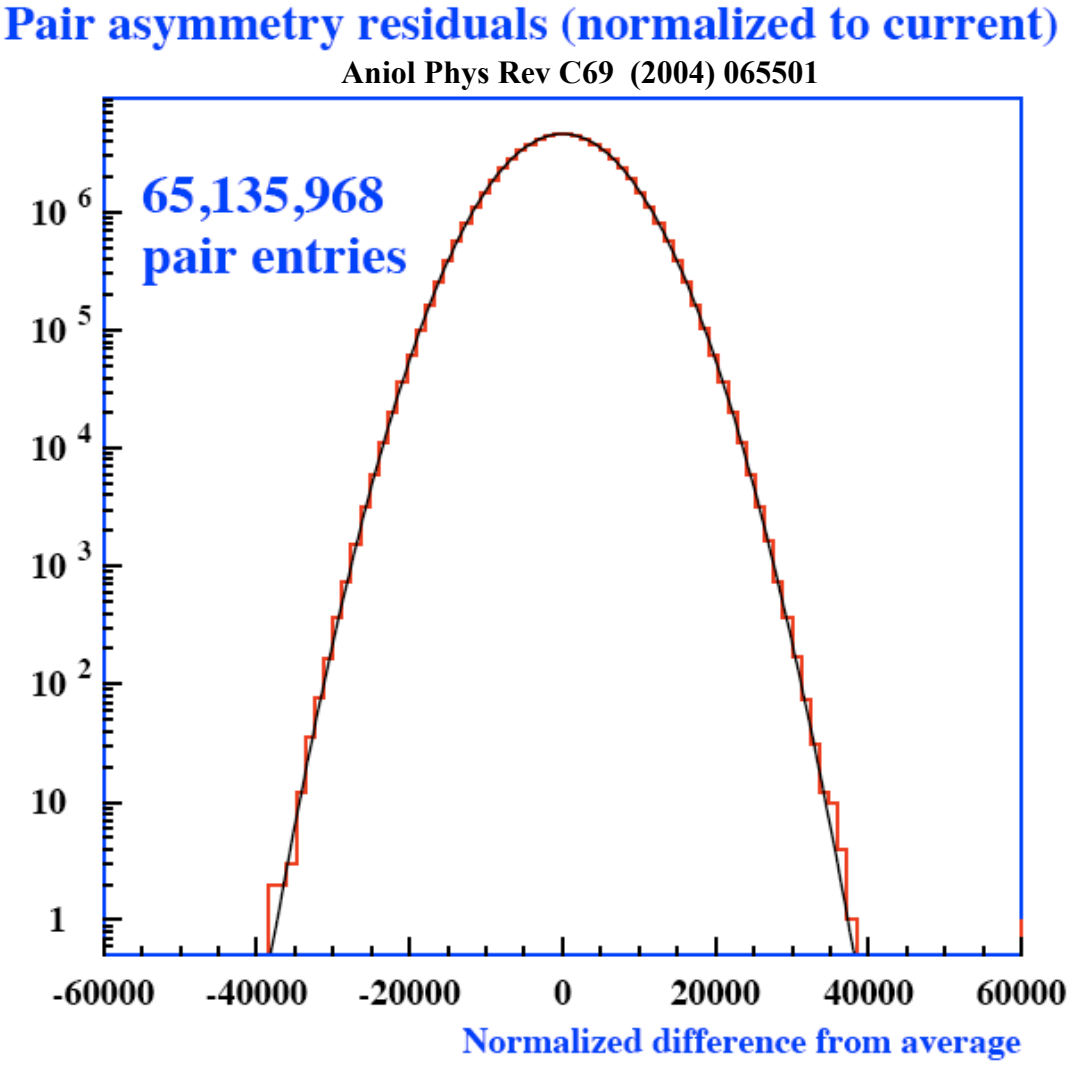
$$E(\bar{x}^2) = E\left(\frac{1}{n}\sum_i x_i \frac{1}{n}\sum_j x_j\right) = \frac{1}{n^2}\sum_{i,j} E(x_i x_j) = \frac{1}{n^2}\left(nE(x_i^2) + \sum_{i\neq j} E(x_i)E(x_j)\right)$$

$$E(\bar{x}^2) = \frac{1}{n^2}\left(nE(x_i^2) + (n^2-n)\mu^2\right) = \frac{1}{n}\left(E(x_i^2) + (n-1)\mu\right) = E(x_i\bar{x})$$

$$E(s^2) = \frac{1}{n}\sum E(x_i^2) - \mu - E(\bar{x}^2) + \mu = \frac{1}{n}\sum E(x_i^2) - E^2(x_i) - E(\bar{x}^2) + E^2(\bar{x})$$

$$E(s^2) = \frac{1}{n}\sum\left(\sigma^2 - \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2$$

Cramér Math Meth of Statistics 27.4

Window pair asymmetries

Window pair asymmetries for 1999 HAPPEX running period, normalized by square root of beam intensity, with mean value subtracted off, in ppm.