# Neyman Construction

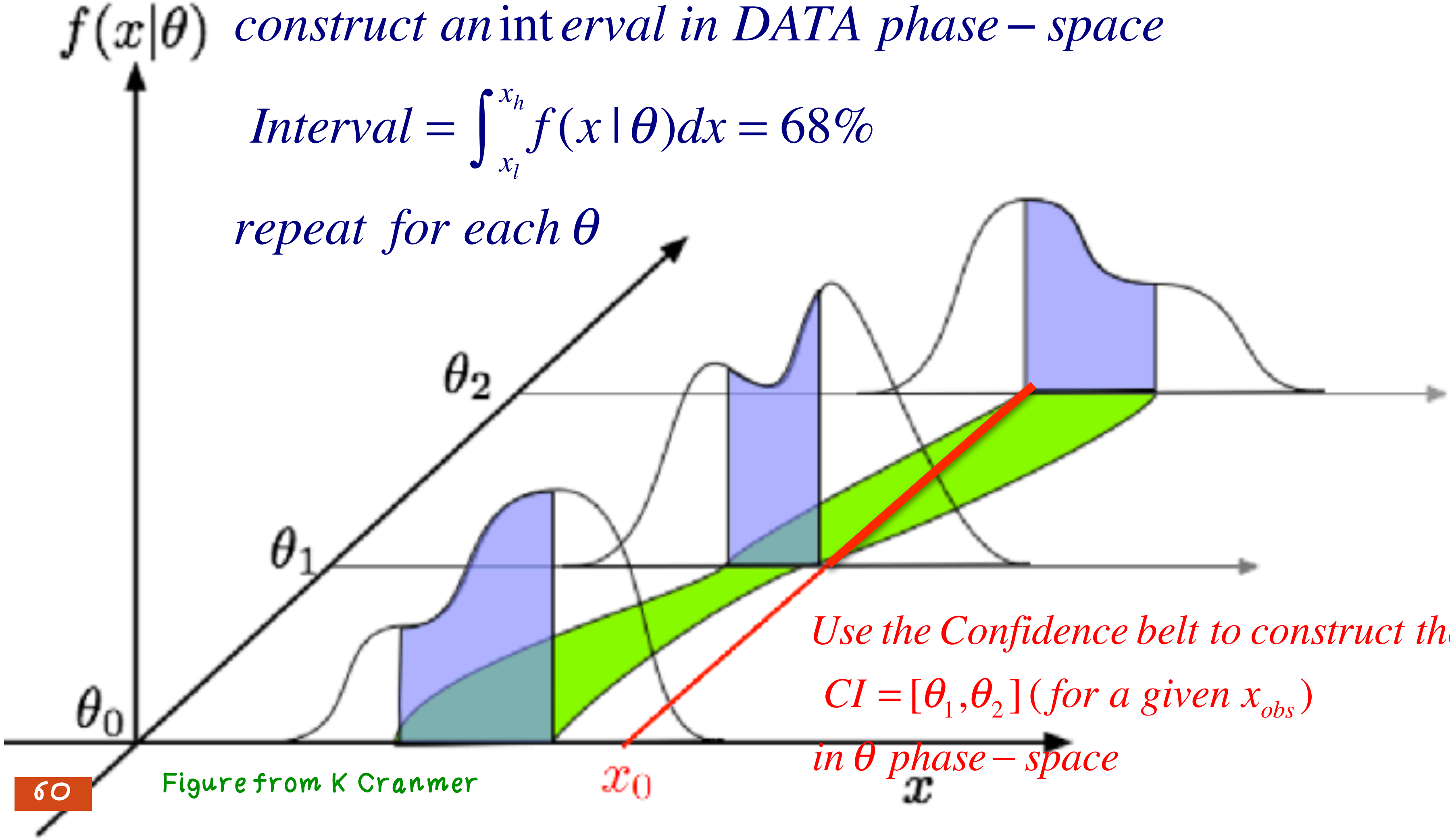
$\theta \equiv s_{true} \quad x \equiv s_{measured} \quad pdf \ f(x|\theta) \ is \ known$

*for each prospective* $\theta$ *generate* $x$

*construct an interval in DATA phase – space*

$$Interval = \int_{x_l}^{x_h} f(x|\theta)dx = 68\%$$

*repeat for each* $\theta$




*Use the Confidence belt to construct the*
$CI = [\theta_1, \theta_2]$ *(for a given* $x_{obs}$*)*
*in* $\theta$ *phase – space*

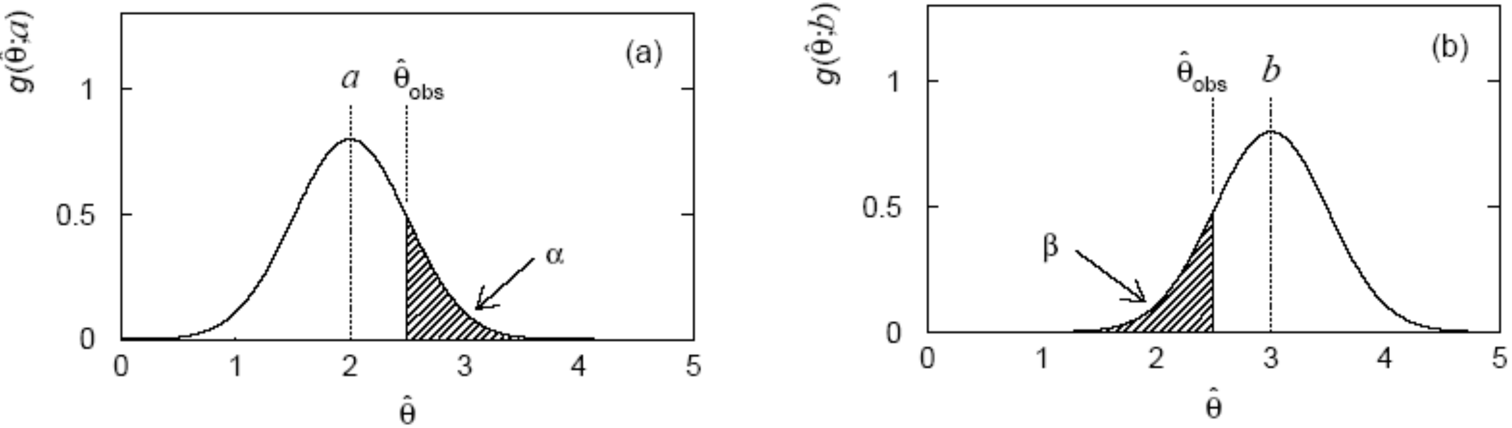
Figure from K Cranmer

# Confidence intervals in practice

The recipe to find the interval [*a*, *b*] boils down to solving

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\,d\hat{\theta} = \int_{\hat{\theta}_{\rm obs}}^{\infty} g(\hat{\theta};a)\,d\hat{\theta}\,,$$

$$\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\,d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{\rm obs}} g(\hat{\theta};b)\,d\hat{\theta}\,.$$




→ $a$ is hypothetical value of $\theta$ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha\,.$

→ $b$ is hypothetical value of $\theta$ such that $P(\hat{\theta} < \hat{\theta}_{\rm obs}) = \beta\,.$

# Meaning of a confidence interval

N.B. the interval is random, the true $\theta$ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

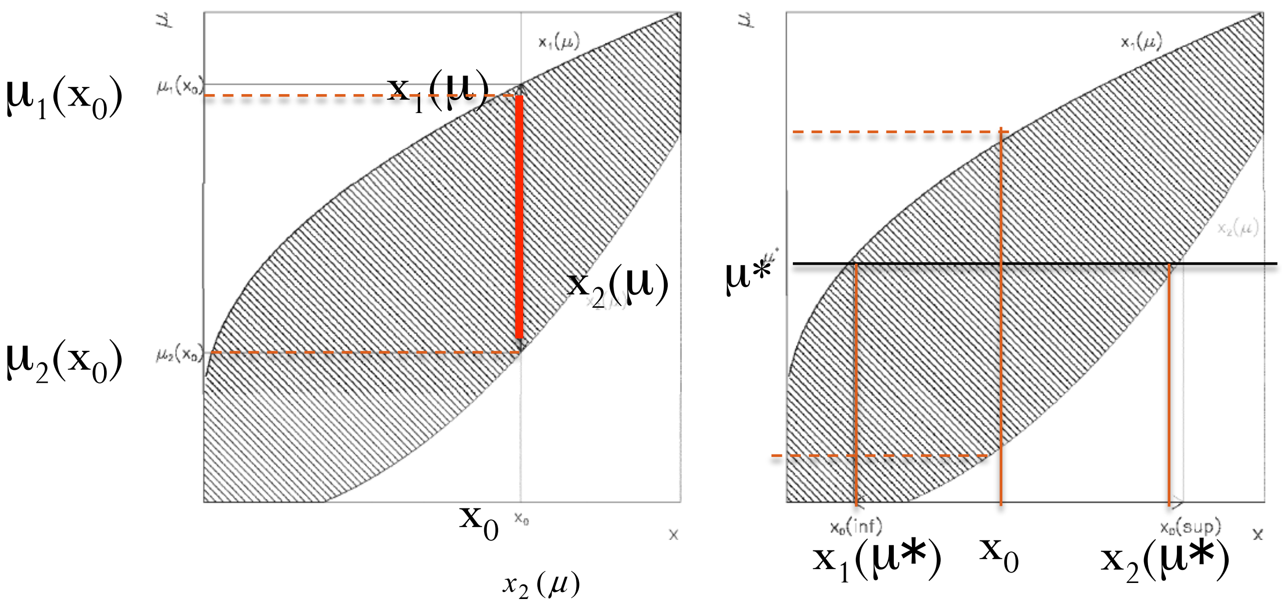
$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover $\theta$.

Neyman's construction

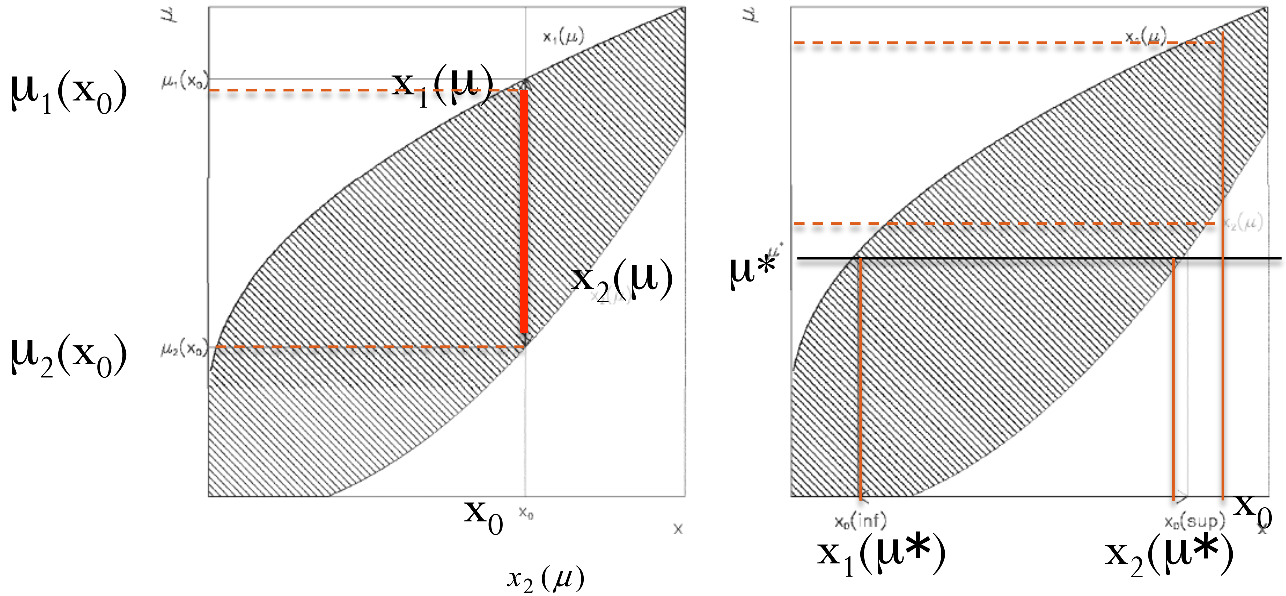



$$\int_{x_1(\mu)}^{x_2(\mu)} f(x/\mu)dx = \alpha$$

Coverage: suppose μ* the true value

$$P(x_1(\mu^*) < x_0 < x_2(\mu^*)) = \alpha$$

Neyman's construction




$$\int_{x_1(\mu)}^{x_2(\mu)} f(x/\mu)dx = \alpha$$

Coverage: suppose μ* the true value

$$P(x_1(\mu^*) < x_0 < x_2(\mu^*)) = \alpha$$

# “Logic” of an EPP experiment - VI

- End of the selection: CANDIDATES sample $N_{cand}$
- Which relation is there between $N_{cand}$ and $N_X$ ?
  - ***Efficiency***: not all searched final states are selected and go to the candidates sample.(Trigger efficiencies are particularly delicate to treat.) Efficiency includes also the **acceptance.**
  - ***Background***: few other final states are faking good ones and go in the candidates sample.

$$\varepsilon N_X = N_{cand} - N_b$$

- where:
  - $\varepsilon$ = efficiency ($0<\varepsilon<1$); $\boldsymbol{\varepsilon} = \boldsymbol{A} \times \boldsymbol{\varepsilon}_d$
  - $N_b$ = number of background events
- Estimate $\varepsilon$ and $N_b$ is a crucial work for the experimentalist and can be done either using simulation (this is tipically done before the experiment and updated later) or using data themselves.

# Binomial distribution

Consider $N$ independent experiments (Bernoulli trials):

outcome of each is 'success' or 'failure',

probability of success on any given trial is $p$.

Define discrete r.v. $n$ = number of successes ($0 \leq n \leq N$).

Probability of a specific outcome (in order), e.g. 'ssfsf' is

$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

But order not important; there are $\dfrac{N!}{n!(N-n)!}$

ways (permutations) to get $n$ successes in $N$ trials, total probability for $n$ is sum of probabilities for each permutation.

# Binomial distribution (2)

The binomial distribution is therefore

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

random variable → $n$

parameters → $N, p$

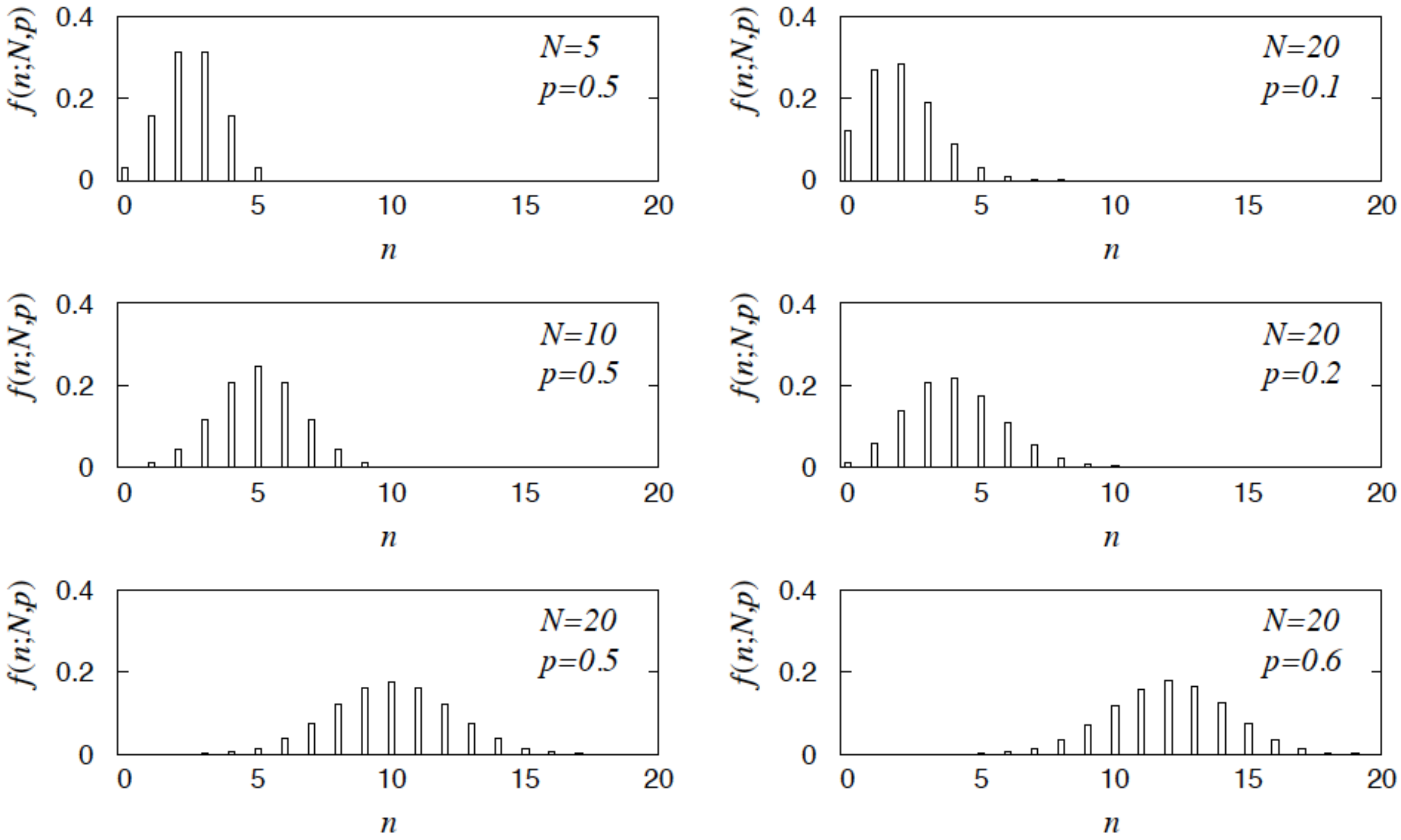
For the expectation value and variance we find:

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

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

# Binomial distribution (3)

Binomial distribution for several values of the parameters:




Example: observe $N$ decays of $W^{\pm}$, the number $n$ of which are $W \to \mu\nu$ is a binomial r.v., $p$ = branching ratio.

# Multinomial distribution

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

$$\vec{p} = (p_1, \ldots, p_m)\,, \quad \text{with} \sum_{i=1}^{m} p_i = 1\ .$$

For *N* trials we want the probability to obtain:

$n_1$ of outcome 1,
$n_2$ of outcome 2,
$\vdots$
$n_m$ of outcome *m*.

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

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

# Multinomial distribution (2)

Now consider outcome $i$ as 'success', all others as 'failure'.

→ all $n_i$ individually binomial with parameters $N$, $p_i$

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i) \quad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram

with $m$ bins, $N$ total entries, all entries independent.

# Poisson distribution

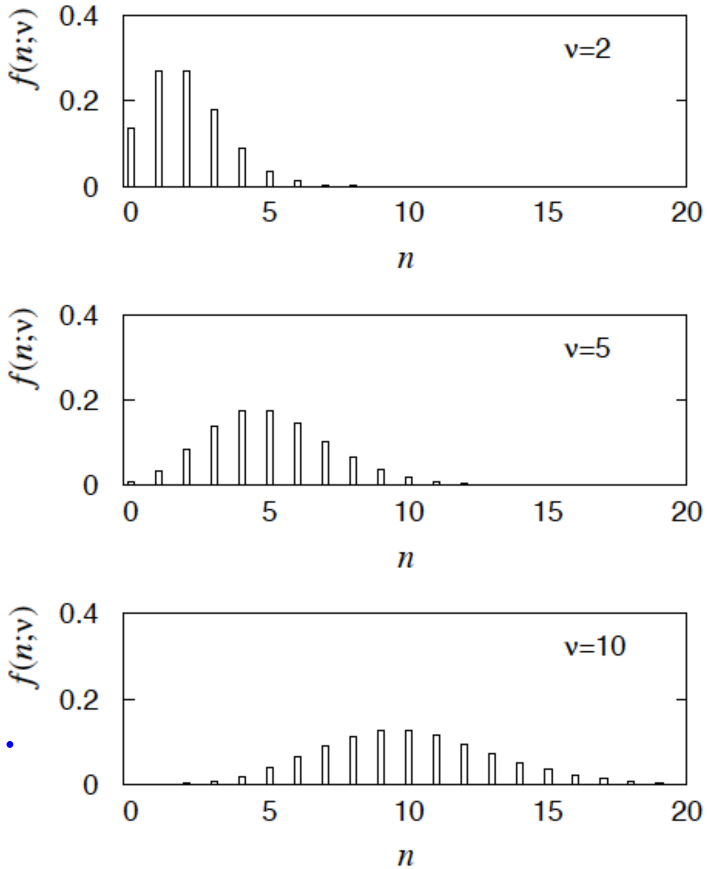
Consider binomial $n$ in the limit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \,.$$

$\to$ $n$ follows the Poisson distribution:

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

$$E[n] = \nu \,, \qquad V[n] = \nu \,.$$




Example: number of scattering events $n$ with cross section $\sigma$ found for a fixed integrated luminosity, with $\nu = \sigma \int L \, dt$ .

# In a Nut Shell

The binomial distribution with parameters n and p
is
the discrete probability distribution of the number of successes in a sequence of n independent experiments.
(Wikipedia)

$$P(k : n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

If X ~ B(n, p)

$$E[X] = np$$

$$P(k:n,p)=\binom{n}{k}p^k(1-p)^{n-k}$$

The Poisson distribution with parameter λ = np can be used as an approximation to B(n, p) of the binomial distribution if n is sufficiently large and p is sufficiently small.

$$P(k:n,p)\xrightarrow{n\to\infty,np=\lambda}Poiss(k;\lambda)=\frac{\lambda^k e^{-k}}{k!}$$

$$If\ X\sim Poiss(k;\lambda)$$

$$E[X]=Var[X]=\lambda$$

# From Binomial to Poisson to Gaussian

$$P(k:n,p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}$$

$$P(k:n,p) \xrightarrow{n\to\infty, np=\lambda} Poiss(k;\lambda) = \frac{\lambda^k e^{-k}}{k!}$$

$$\langle k \rangle = \lambda, \ \sigma_k = \sqrt{\lambda}$$

$$k \to \infty \Rightarrow x = k$$

Using Stirling Formula

$$\text{prob(x)=G(x},\sigma = \sqrt{\lambda}) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\lambda)^2/2\sigma^2}$$

*This is a Gaussian, or Normal distribution*

*with mean and variance of* $\lambda$

# Histograms

*N collisions*

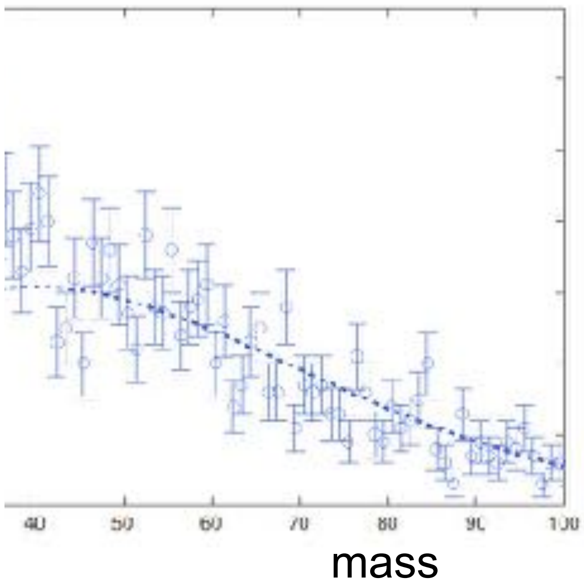
$$p(Higgs\ event) = \frac{\mathcal{L}\sigma(pp \to H)\, A\epsilon_{ff}}{\mathcal{L}\sigma(pp)}$$

Prob to see $n_H^{obs}$ in $N$ collisions is

$$P(n_H^{obs}) = \binom{N}{n_H^{obs}} p^{n_H^{obs}} (1-p)^{N-n_H^{obs}}$$

$$\ell im_{N\to\infty} P(n_H^{obs}) = Poiss(n_H^{obs}, \lambda) = \frac{e^{-\lambda}\lambda^{n_H^{obs}}}{n_H^{obs}!}$$

$$\lambda = Np = \mathcal{L}\sigma(pp) \cdot \frac{\mathcal{L}\sigma(pp \to H)\, A\epsilon_{ff}}{\mathcal{L}\sigma(pp)} = n_H^{exp}$$



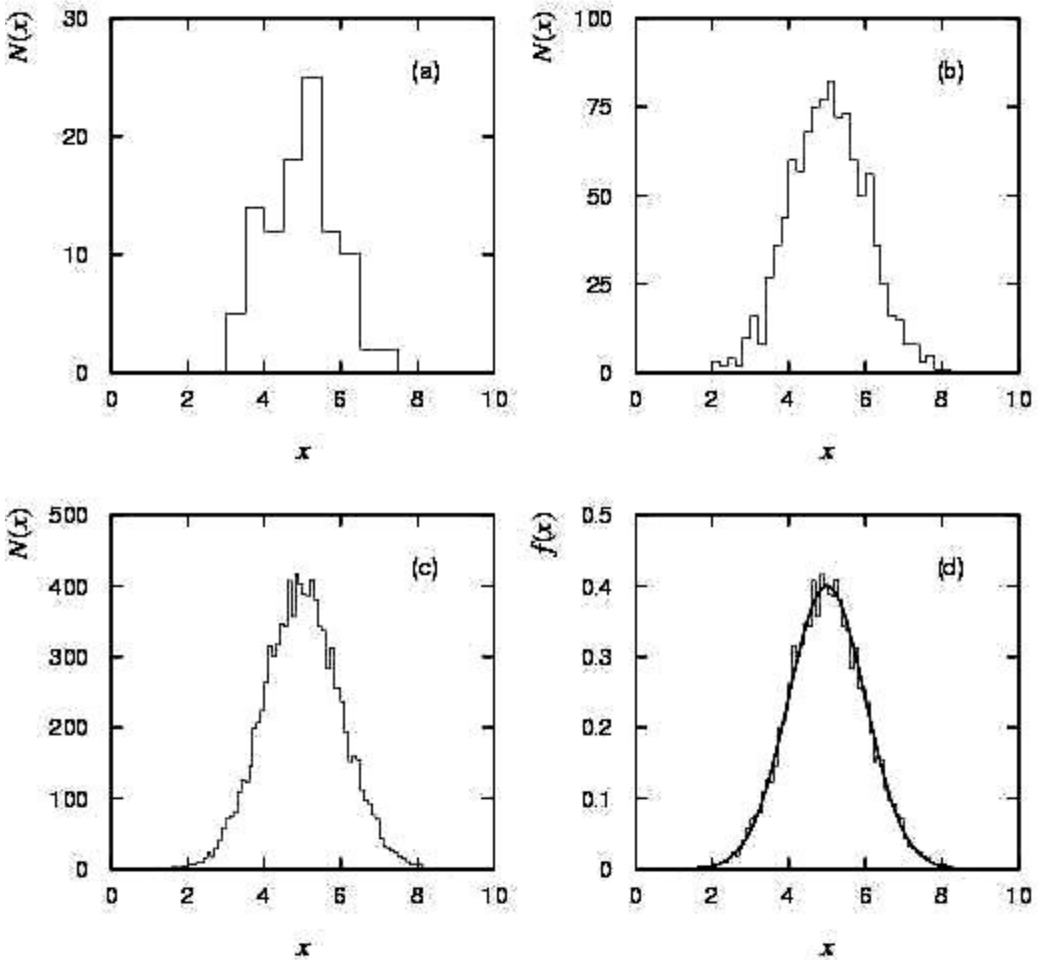
# Histograms

pdf = histogram with

infinite data sample,

zero bin width,

normalized to unit area.




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

$n =$ number of entries

$\Delta x =$ bin width

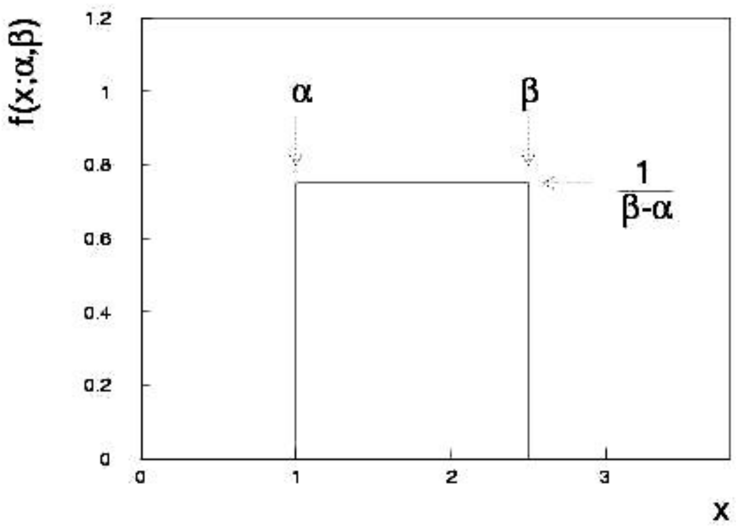
# Uniform distribution

Consider a continuous r.v. $x$ with $-\infty < x < \infty$ . Uniform pdf is:

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{1}{2}(\alpha + \beta)$$

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




N.B. For any r.v. $x$ with cumulative distribution $F(x)$,
$y = F(x)$ is uniform in [0,1].

Example: for $\pi^0 \to \gamma\gamma$, $E_\gamma$ is uniform in $[E_{min}, E_{max}]$, with

$$E_{\rm min} = \frac{1}{2}E_\pi(1-\beta)\,, \qquad E_{\rm max} = \frac{1}{2}E_\pi(1+\beta)$$

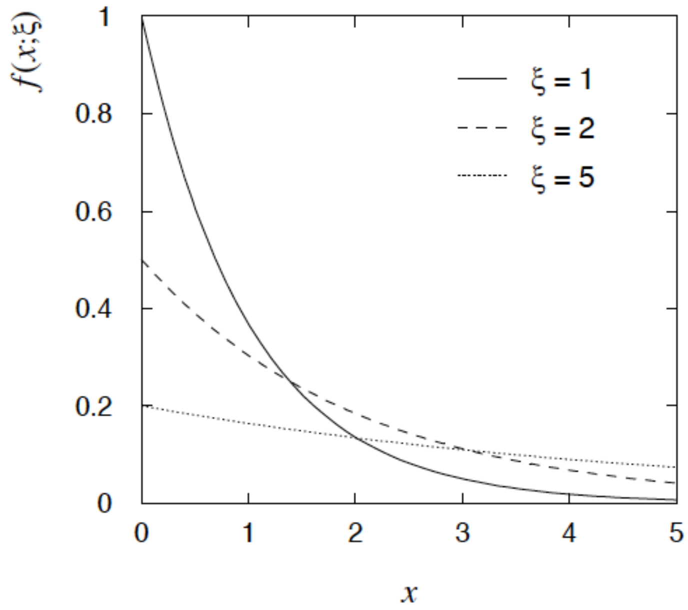
# Exponential distribution

The exponential pdf for the continuous r.v. *x* is defined by:

$$f(x;\xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \xi$$

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




Example: proper decay time *t* of an unstable particle

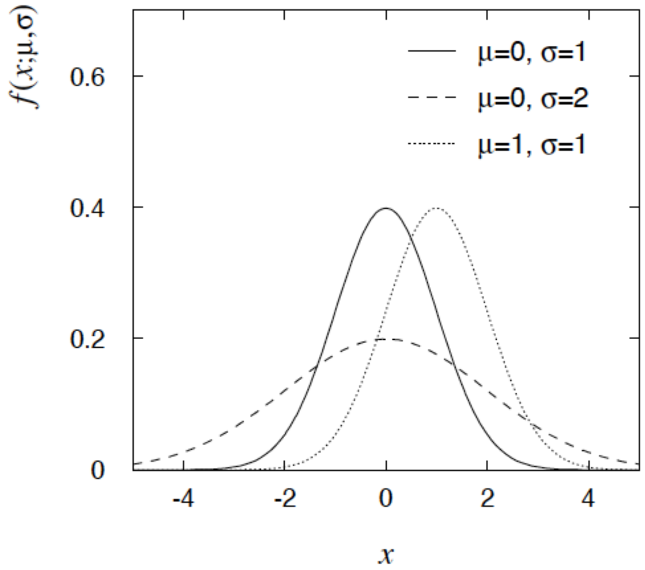
$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad (\tau = \text{mean lifetime})$$

Lack of memory (unique to exponential): $f(t - t_0 | t \geq t_0) = f(t)$

# Gaussian distribution

The Gaussian (normal) pdf for a continuous r.v. $x$ is defined by:

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




$E[x] = \mu$

$V[x] = \sigma^2$

(N.B. often $\mu$, $\sigma^2$ denote mean, variance of any r.v., not only Gaussian.)

Special case: $\mu = 0$, $\sigma^2 = 1$ (‘standard Gaussian’):

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(x')\,dx'$$

If $y \sim$ Gaussian with $\mu$, $\sigma^2$, then $x = (y - \mu)/\sigma$ follows $\varphi(x)$.

# Gaussian pdf and the Central Limit Theorem

The Gaussian pdf is so useful because almost any random variable that is a sum of a large number of small contributions follows it. This follows from the Central Limit Theorem:

For $n$ independent r.v.s $x_i$ with finite variances $\sigma_i^2$, otherwise arbitrary pdfs, consider the sum

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

In the limit $n \to \infty$, $y$ is a Gaussian r.v. with

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

Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian r.v.s.

# Quantities to measure

- In order to estimate $N_X$ we need to measure:
  - $N_{cand}$
  - $\varepsilon$
  - $N_b$
- We already know that each of these variables have a fluctuation model:
  - $N_{cand}$ is described by a Poisson process
  - $\varepsilon$ is described by a Bernoulli process
  - $N_b$

# $N_{cand}$: a Poisson variable

- If events come in a random way (without any time structure) the event count $N$ is a Poisson variable.
- → if I count $N$, the best estimate of $\lambda$ is $N$ itself and the uncertainty is $\sqrt{N}$
- If $N$ is large enough ($N>20$) Poisson → Gaussian. → $N \pm \sqrt{N}$ is a 68% probability interval for $N$.
- If $N$ is small (close to 0) the Gaussian limit is not ok, a specific treatment is required (see later in the course).

# $N_{cand}$: a Poisson variable

- If events come in a random way (without any time structure) the event count $N$ is a Poisson variable.
- ➔ if I count $N$, the best estimate of $\lambda$ is $N$ itself ( or better $N+1$) and the uncertainty is $\sqrt{N}$

$$P(N,\lambda) = \lambda^N e^{-\lambda}/N! \Rightarrow P(\lambda|N) = \lambda^N e^{-\lambda}/N!$$
$$E[\lambda] = N+1$$
$$\mathrm{var}[\lambda] = N+1$$

- If $N$ is large enough ($N>20$) Poisson → Gaussian. ➔ $N\pm\sqrt{N}$ is a 68% probability interval for $N$.
- If $N$ is small (close to 0) the Gaussian limit is not ok, a specific treatment is required (see later in the course).

# Efficiency: a binomial variable - I

- Bernoulli process: success/failure ***N*** proofs, 0<***n***<***N***, ***p*** = success probability. ***p*** == ***ε***

$$P(n/N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$E\left[n\right] = Np$$

$$\text{var}\left[n\right] = Np(1-p)$$

- Inference: given ***n*** and ***N*** which is the best estimate of ***p*** ?

And its uncertainty ? *(see previous lectures)*

$$\varepsilon = \hat{p} = \frac{n+1}{N+2}$$

$$\sigma\left(\varepsilon\right) = \frac{\sigma\left(n\right)}{N} = \frac{1}{\sqrt{N+2}}\sqrt{\hat{p}(1-\hat{p})}$$

# Efficiency: a binomial variable - I

- Bernoulli process: success/failure **N** proofs, 0<***n***<**N**, ***p*** = success probability. ***p*** == ***ε***

$$P(n/N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$E[n] = Np$$

$$\text{var}[n] = Np(1-p)$$

- Inference: given ***n*** and **N** which is the best estimate of ***p*** ?

And its uncertainty ? *(see previous lectures)*

$$\varepsilon = \hat{p} = \frac{n}{N}$$

$$\sigma(\varepsilon) = \frac{\sigma(n)}{N} = \frac{1}{\sqrt{N}}\sqrt{\hat{p}(1-\hat{p})}$$

# Efficiency: a binomial variable - II

- How measure it ?
  - From data: Sample of $N$ true particles and I measure how many, out of these give rise to a signal in my detector
  - From MC: I generate $N_{gen}$ "signal" events. If I select $N_{sel}$ of these events out of $N_{gen}$, the efficiency is (assume $N_{gen}$ and $N_{sel}$ large numbers):

$$\varepsilon = \frac{N_{sel}}{N_{gen}}$$

$$\sigma(\varepsilon) = \frac{\sigma(N_{sel})}{N_{gen}} = \frac{1}{\sqrt{N_{gen}}}\sqrt{\frac{N_{sel}}{N_{gen}}\left(1-\frac{N_{sel}}{N_{gen}}\right)}$$

# Background $N_b$

- Simulation of $N_{gen}$ "bad final states"; $N_{sel}$ are selected. What about $N_b$ ?
- We define the "rejection factor" $R = N_{gen}/N_{sel} > 1$
- We also need a correct normalization in this case: we need to know $N_{exp}$ = total number of expected "bad final states" in our sample ($N_{exp}$ related to luminosity and cross-section).

$$N_b = N_{sel}\frac{N_{\exp}}{N_{gen}} = \frac{N_{\exp}}{R}$$

$$\sigma(N_b) = \sigma(N_{sel})\frac{N_{\exp}}{N_{gen}} = \sqrt{N_{sel}}\frac{N_{\exp}}{N_{gen}} = \frac{N_{\exp}}{\sqrt{RN_{gen}}}$$

# Statistical Errors

- In alla cases there is an unreducible error on $\boldsymbol{N_X}$ given by limited statistics. It is a random error, coming from the procedure of "sampling" that is intrinsic in our experiments.
- In all cases increasing the statistics, the error decreases

$$\frac{\sigma(N_{cand})}{N_{cand}} = \frac{1}{\sqrt{N_{cand}}}$$

$$\sigma(\varepsilon) \approx \frac{1}{\sqrt{N_{gen}}}$$

$$\sigma(N_b) \approx \frac{1}{\sqrt{N_{gen}}}$$