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

parameters

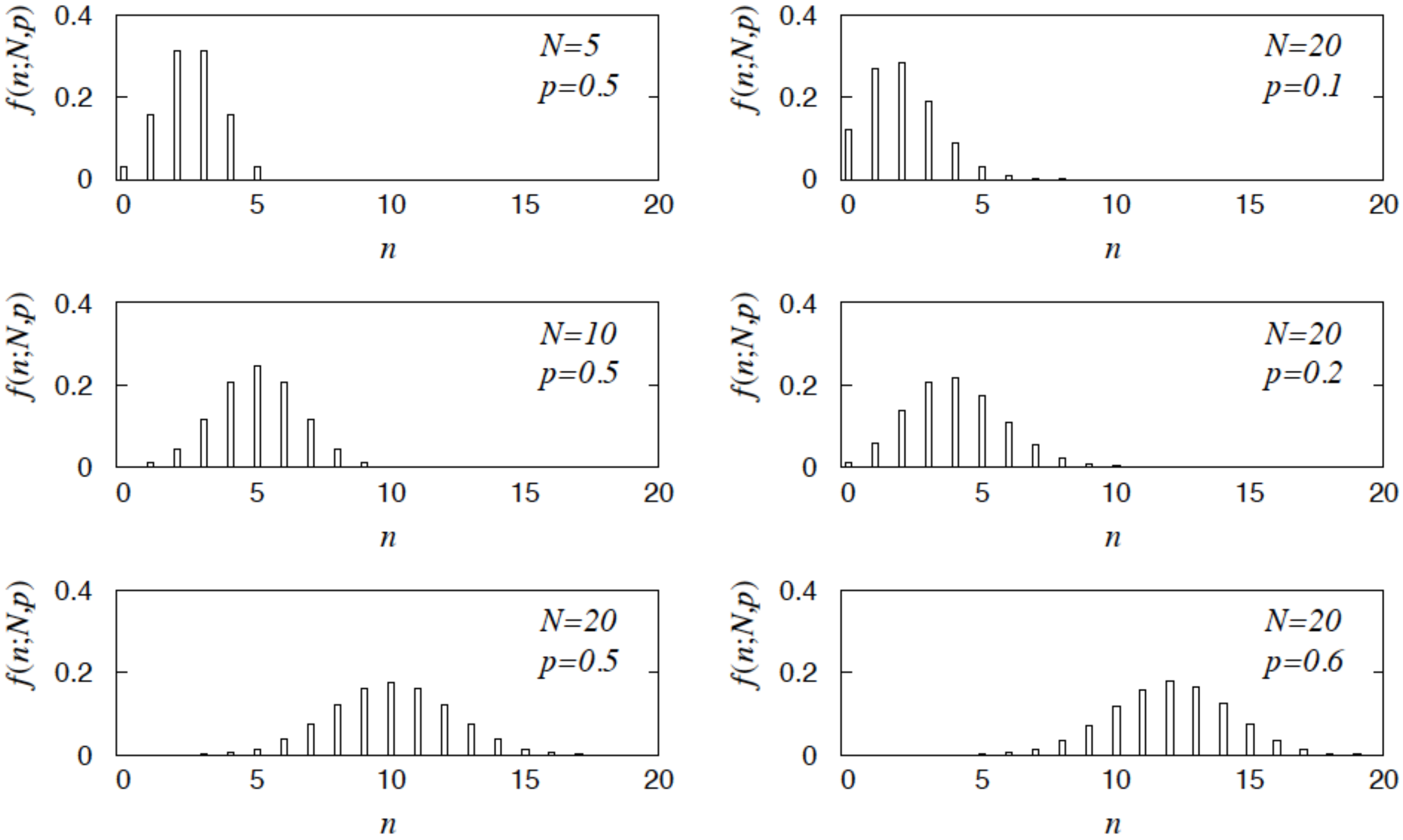
For the expectation value and variance we find:

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

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

Binomial distribution (3)

Binomial distribution for several values of the parameters:



Example: observe N decays of $W^{\pm}$, the number n of which are $W \rightarrow \mu\nu$ 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:

$$\begin{aligned} &n_1 \text{ of outcome } 1, \\ &n_2 \text{ of outcome } 2, \\ &\quad\vdots \\ &n_m \text{ of outcome } m. \end{aligned}$$

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

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

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

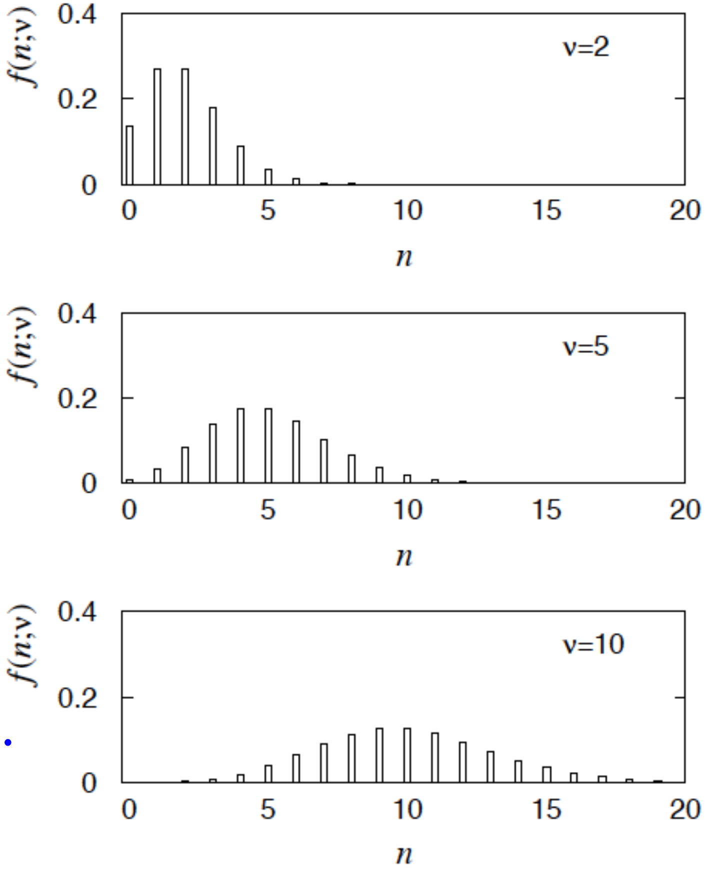
Consider binomial n in the limit

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

→ 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 σ found for a fixed integrated luminosity, with $\nu = \sigma \int L \, dt$.

From Binomial to Poisson to Gaussian

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

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

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

$$k\rightarrow\infty\Rightarrow x=k$$

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 λ

Histograms

N collisions

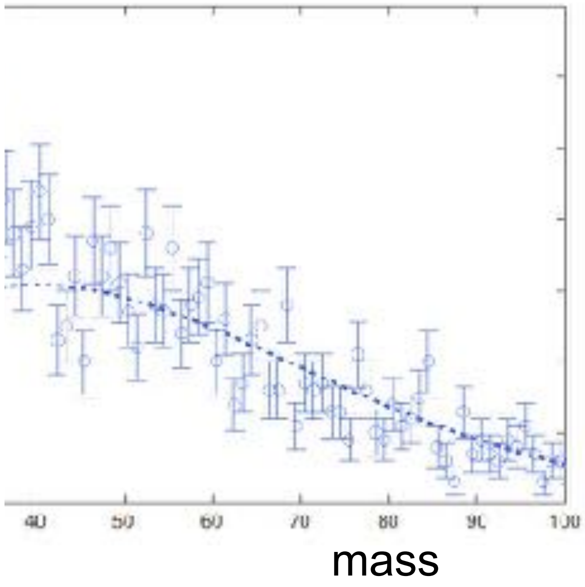
$$p(Higgs\ event) = \frac{\mathcal{L}\sigma(pp \to H)\, A\epsilon_{ff}}{\mathcal{L}\sigma(pp)}$$

Pr*ob to see* n_H^{obs} *in N collisions is*

$$P(n_H^{obs}) = \begin{pmatrix} N \\ n_H^{obs} \end{pmatrix} p^{n_H^{obs}} (1-p)^{N-n_H^{obs}}$$

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

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



Histograms

pdf = histogram with

infinite data sample,

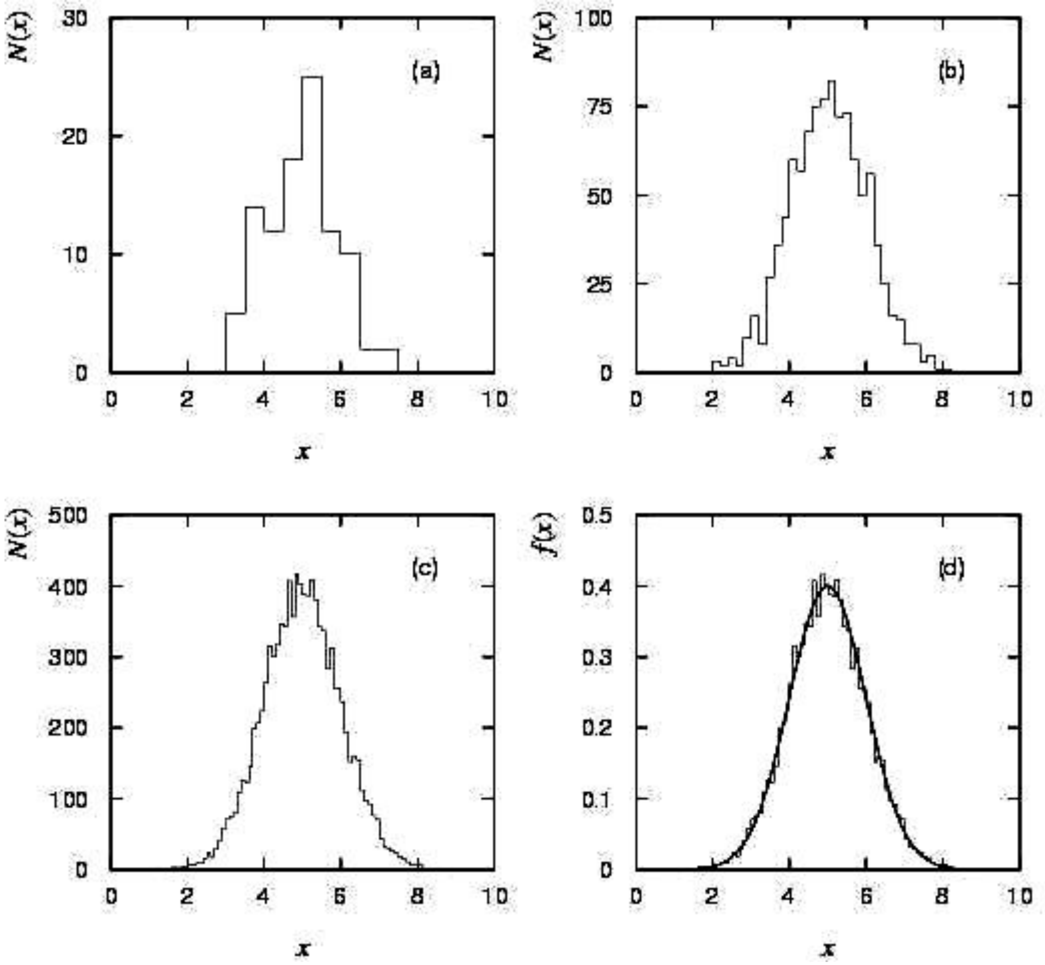
zero bin width,

normalized to unit area.

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

$n =$ number of entries

$\Delta x =$ bin width



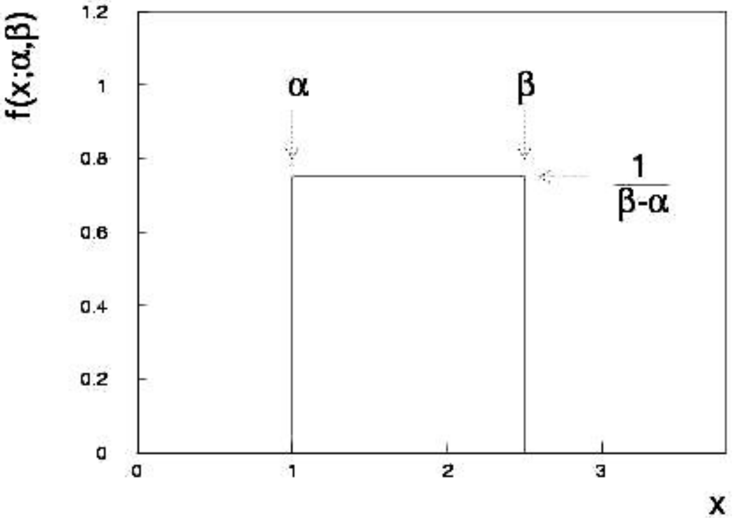
Uniform distribution

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_γ is uniform in $[E_{\min}, E_{\max}]$, with

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

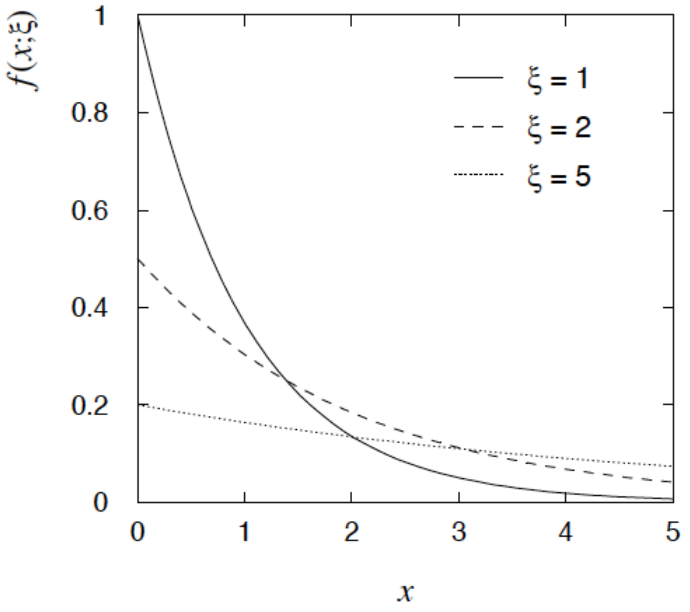
Exponential distribution

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

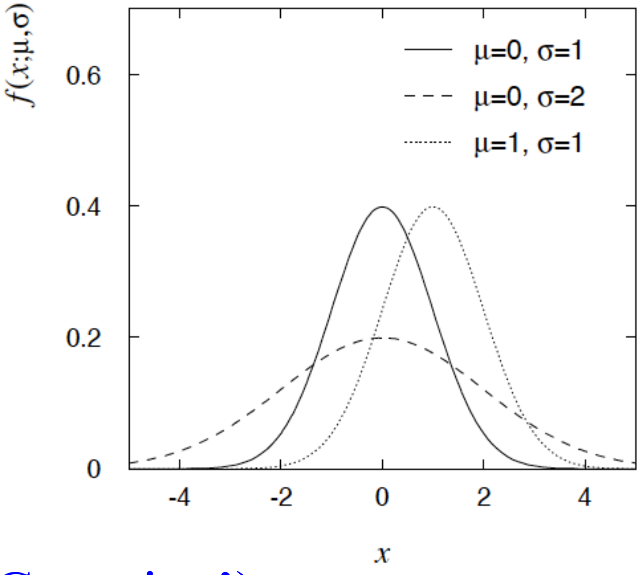
$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad (\tau = \text{mean lifetime})$$

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 μ, σ^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 μ, σ^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 σ_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

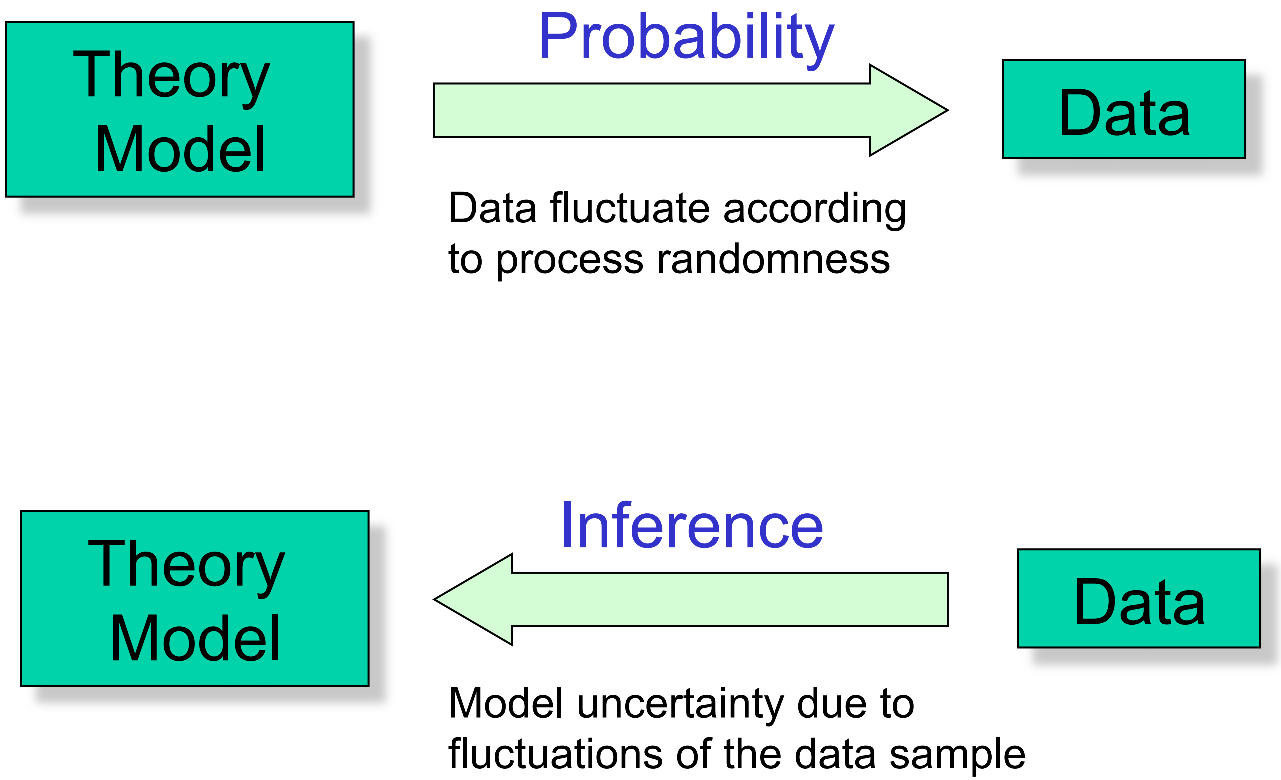
$$E[y] = \sum_{i=1}^{n} \mu_i \qquad V[y] = \sum_{i=1}^{n} \sigma_i^2$$

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

Meaning of parameter estimate

- We are interested in some physical unknown parameters
- Experiments provide samplings of some PDF which has among its parameters the physical unknowns we are interested in
- Experiment's results are statistically "related" to the unknown PDF
    - PDF parameters can be determined from the sample within some approximation or uncertainty
- Knowing a parameter within some error may mean different things:
    - **Frequentist**: a large fraction (68% or 95%, usually) of the experiments will contain, in the limit of large number of experiments, the (fixed) unknown true value within the quoted confidence interval, usually $[\mu - \sigma, \mu + \sigma]$ ('coverage')
    - **Bayesian**: we determine a degree of belief that the unknown parameter is contained in a specified interval can be quantified as 68% or 95%
- We will see that there is still some more degree of arbitrariness in the definition of confidence intervals…

Statistical inference



Theory Model

Probability

Data

Data fluctuate according to process randomness

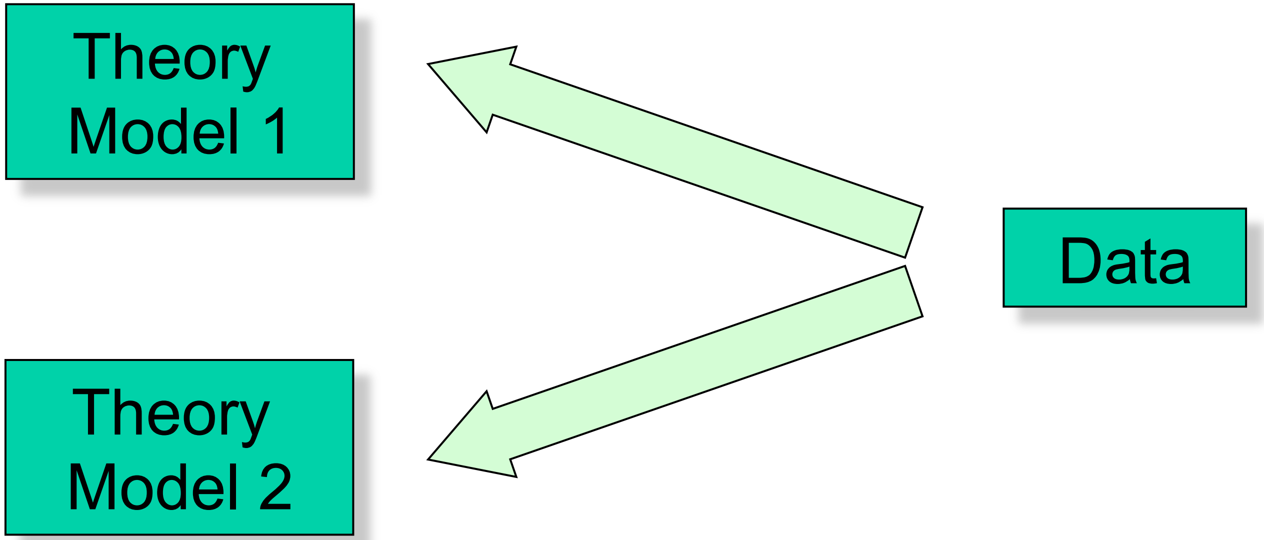
Theory Model

Inference

Data

Model uncertainty due to fluctuations of the data sample

Hypothesis tests



Which hypothesis is the most consistent with the experimental data?

Parameter estimators

- An estimator is a function of a given sample whose statistical properties are known and related to some PDF parameters
  - “Best fit”
- Simplest example:
  - Assume we have a Gaussian PDF with a *known* σ and an *unknown* μ
  - A single experiment will provide a measurement x
  - We estimate μ as $\mu^{\mathrm{est}} = x$
  - The distribution of μ^{est} (repeating the experiment many times) is the original Gaussian
  - 68.27%, *on average*, of the experiments will provide an estimate within: $\mu - \sigma < \mu^{\mathrm{est}} < \mu + \sigma$
- We can determine: $\mu = \mu^{\mathrm{est}} \pm \sigma$

Likelihood function

- Given a sample of N events each with variables $(x_1, \ldots, x_n)$, the likelihood function expresses the probability density of the sample, as a function of the unknown parameters:

$$L = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

- Sometimes the used notation for parameters is the same as for conditional probability:

$$f(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)$$

- If the size N of the sample is also a random variable, the extended likelihood function is also used:

$$L = p(N; \theta_1, \cdots, \theta_m) \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

  - Where p is most of the times a Poisson distribution whose average is a function of the unknown parameters

- In many cases it is convenient to use $-\ln L$ or $-2\ln L$: $\prod_i \rightarrow \sum_i$

Maximum likelihood estimates

- ML is the widest used parameter estimator
- The “best fit” parameters are the set that maximizes the likelihood function
  - “Very good” statistical properties

- The maximization can be performed analytically, for the simplest cases, and numerically for most of the cases
- Minuit is historically the most used minimization engine in High Energy Physics
  - F. James, 1970’s; rewritten in C++ recently

CL & CI

$$\text{measurement } \hat{\mu} = 1.1 \pm 0.3$$

$$L(\mu) = G(\mu; \hat{\mu}, \sigma_{\hat{\mu}})$$

$$\Rightarrow CI \text{ of } \mu = [0.8, 1.4] \text{ at } 68\% \ CL$$

- A confidence interval (CI) is a particular kind of interval estimate of a population parameter.

- Instead of estimating the parameter by a single value, an interval likely to include the parameter is given.

- How likely the interval is to contain the parameter is determined by the confidence level

- Increasing the desired confidence level will widen the confidence interval.

Confidence Interval & Coverage

- Say you have a measurement μ_{meas} of μ with μ_{true} being the unknown true value of μ

- Assume you know the probability distribution function $p(\mu_{meas}|\mu)$

- based on your statistical method you deduce that there is a 95% Confidence interval $[\mu_1,\mu_2]$.

  (it is 95% likely that the μ_{true} is in the quoted interval)

The correct statement:

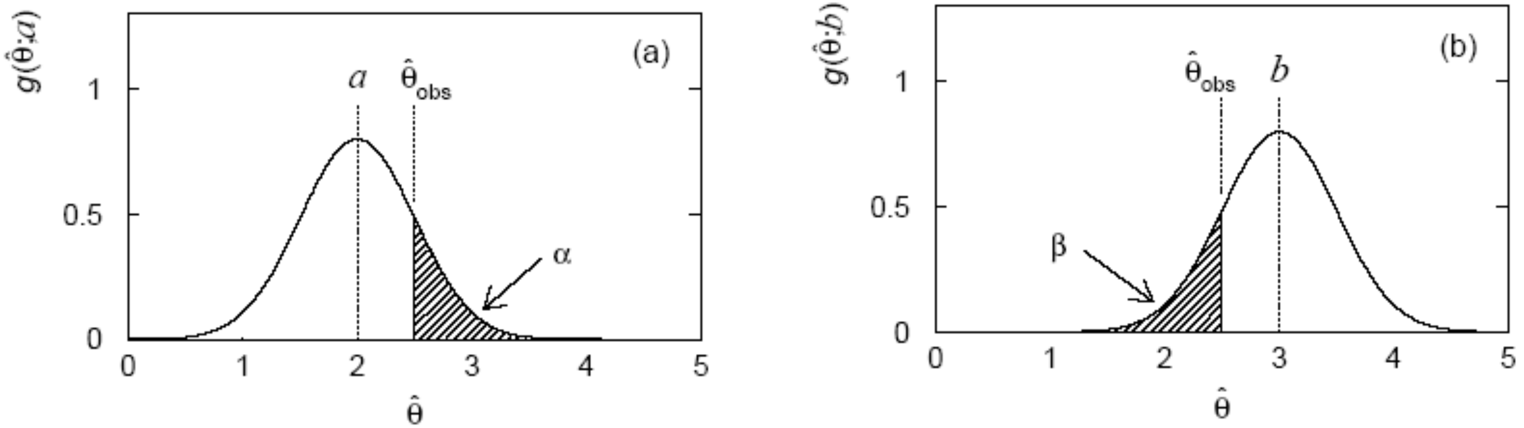
- In an ensemble of experiments 95% of the obtained confidence intervals will contain the true value of μ.

Confidence intervals in practice

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

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

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



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

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

Meaning of a confidence interval

N.B. the interval is random, the true θ 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 θ.

Confidence Interval & Coverage

- You claim, $CI_\mu=[\mu_1,\mu_2]$ at the 95% C.L
  i.e. In an ensemble of experiments CL (95%) of the obtained confidence intervals will contain the true value of μ.

  - If your statement is accurate, you have full coverage
  - If the true C.L is>95%, your interval has an over coverage
  - If the true C.L is <95%, your interval has an undercoverage

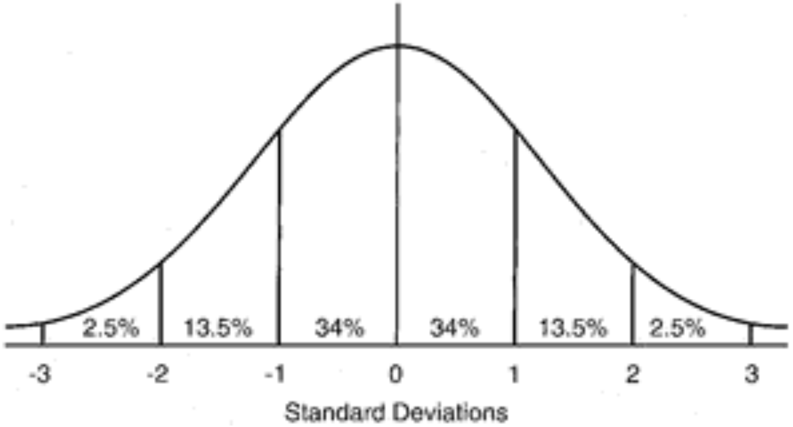
How to deduce a CI

- One can show that if the data is distributed normal around the average i.e. P(data|μ)=normal

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

- then one can construct a 68% CI around the estimator of μ to be

$$\hat{x} \pm \sigma \qquad \boxed{i.e.\, x_{true} \in \left[\hat{x} - \sigma_{\hat{x}}, \hat{x} + \sigma_{\hat{x}}\right] @\, 68\%\, CL}$$



Side Note:
A CI is an interval in the true parameters phase-space

- However, not all distributions are normal, many distributions are even unknown and coverage might be a real issue

- One can guarantee a coverage with the Neyman Construction (1937)

Neyman, J. (1937) "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability" Philosophical Transactions of the Royal Society of London A, 236, 333-380.

The Frequentist Game a 'la Neyman

Or

How to ensure a Coverage with Neyman construction

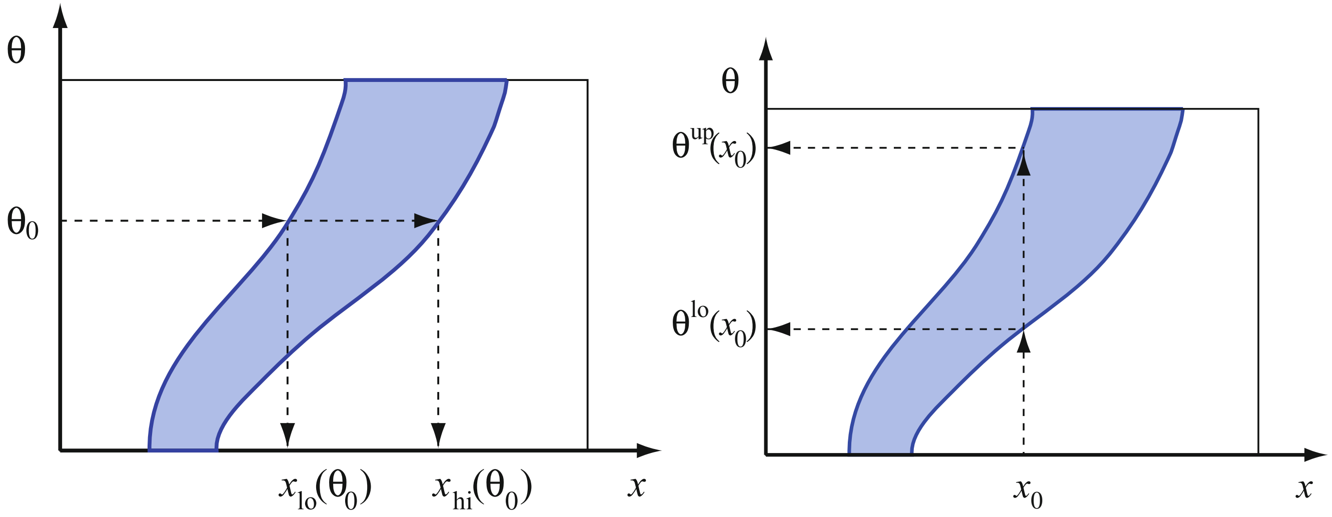
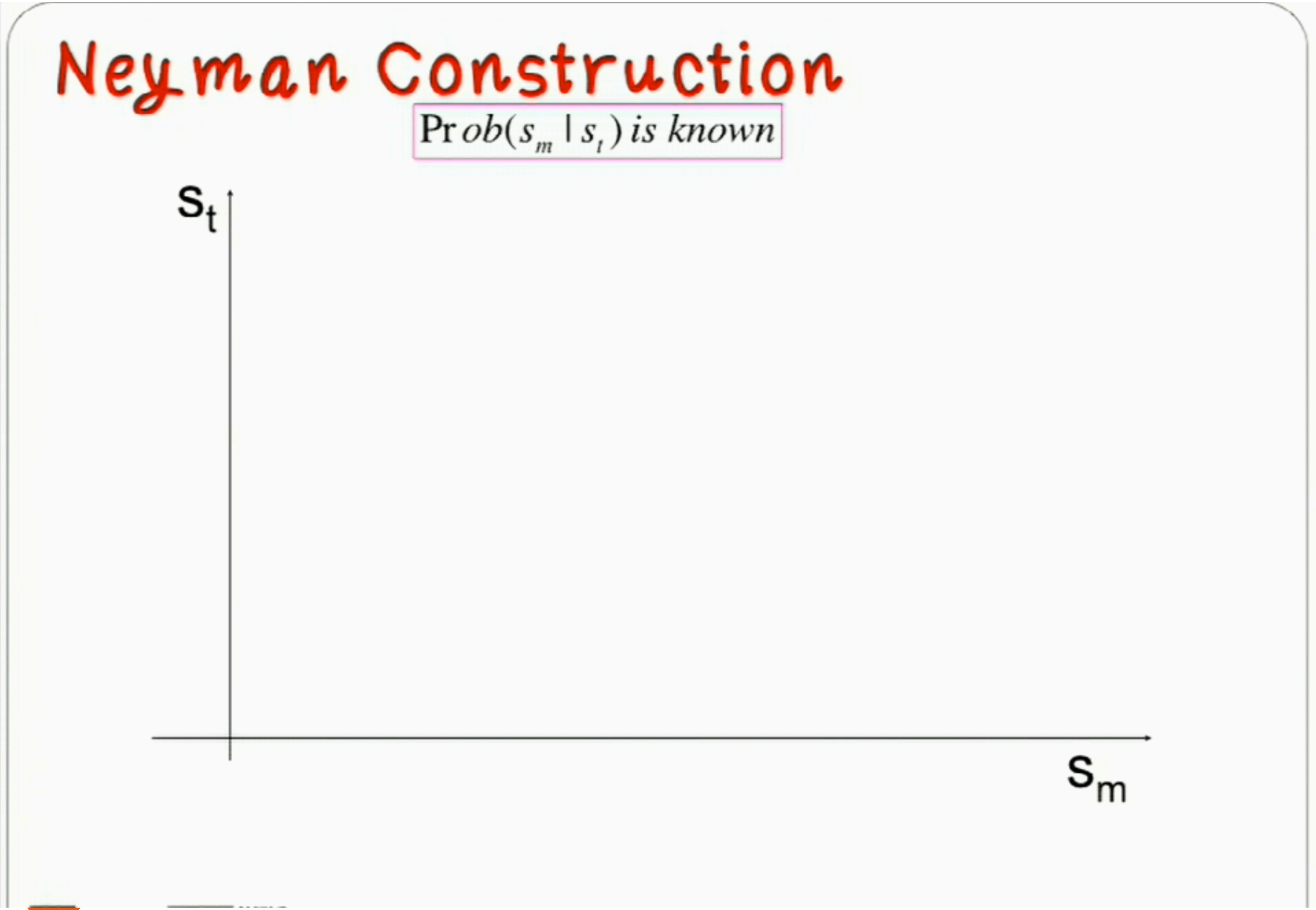


Fig. 7.1 Graphical illustration of Neyman belt construction (*left*) and inversion (*right*)

$$1 - \alpha = \int_{x^{\mathrm{lo}}(\theta_0)}^{x^{\mathrm{up}}(\theta_0)} f(x \,|\, \theta_0)\,\mathrm{d}x$$

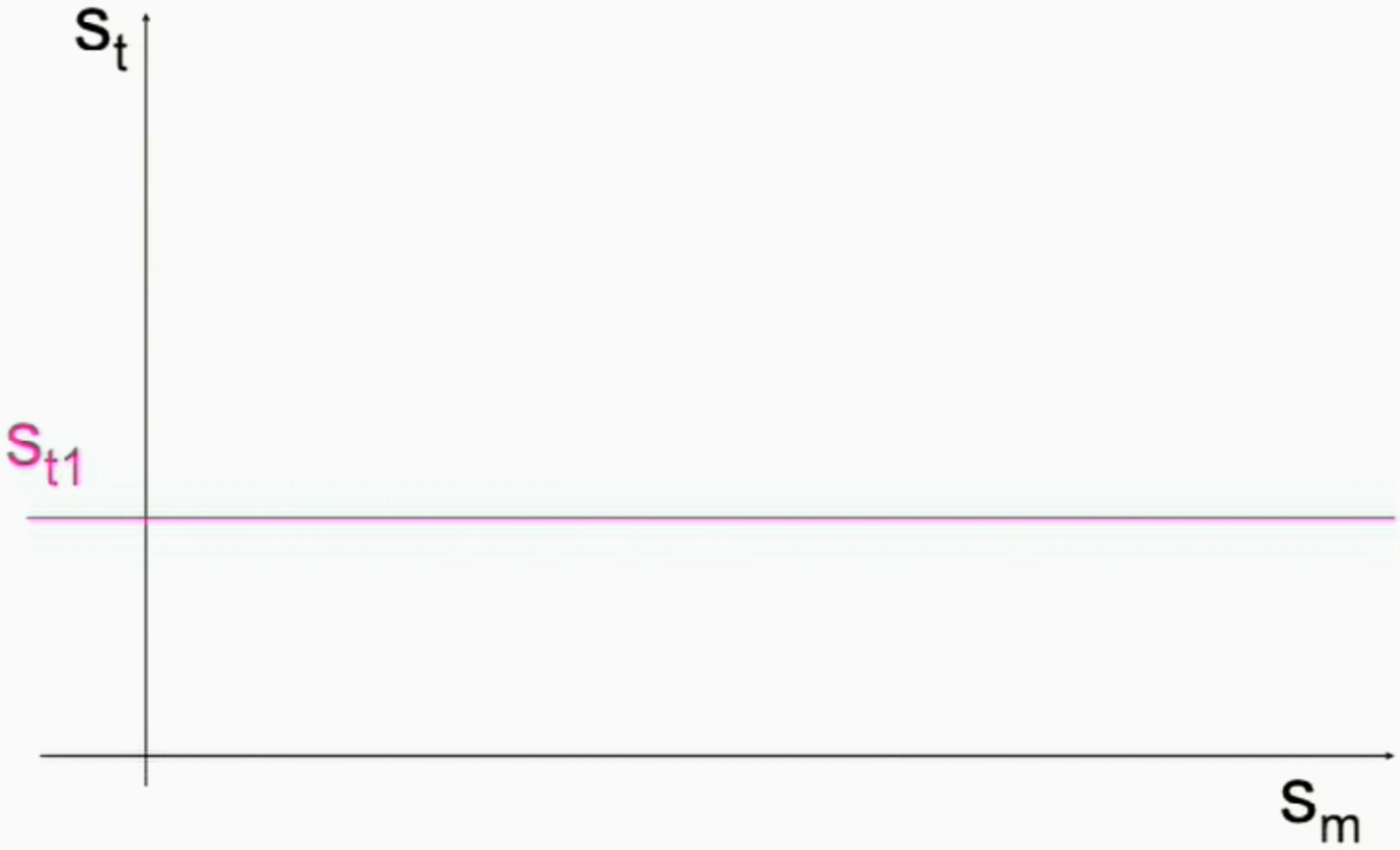


Neyman Construction

$\boxed{\Pr ob(s_m \mid s_t)\ is\ known}$

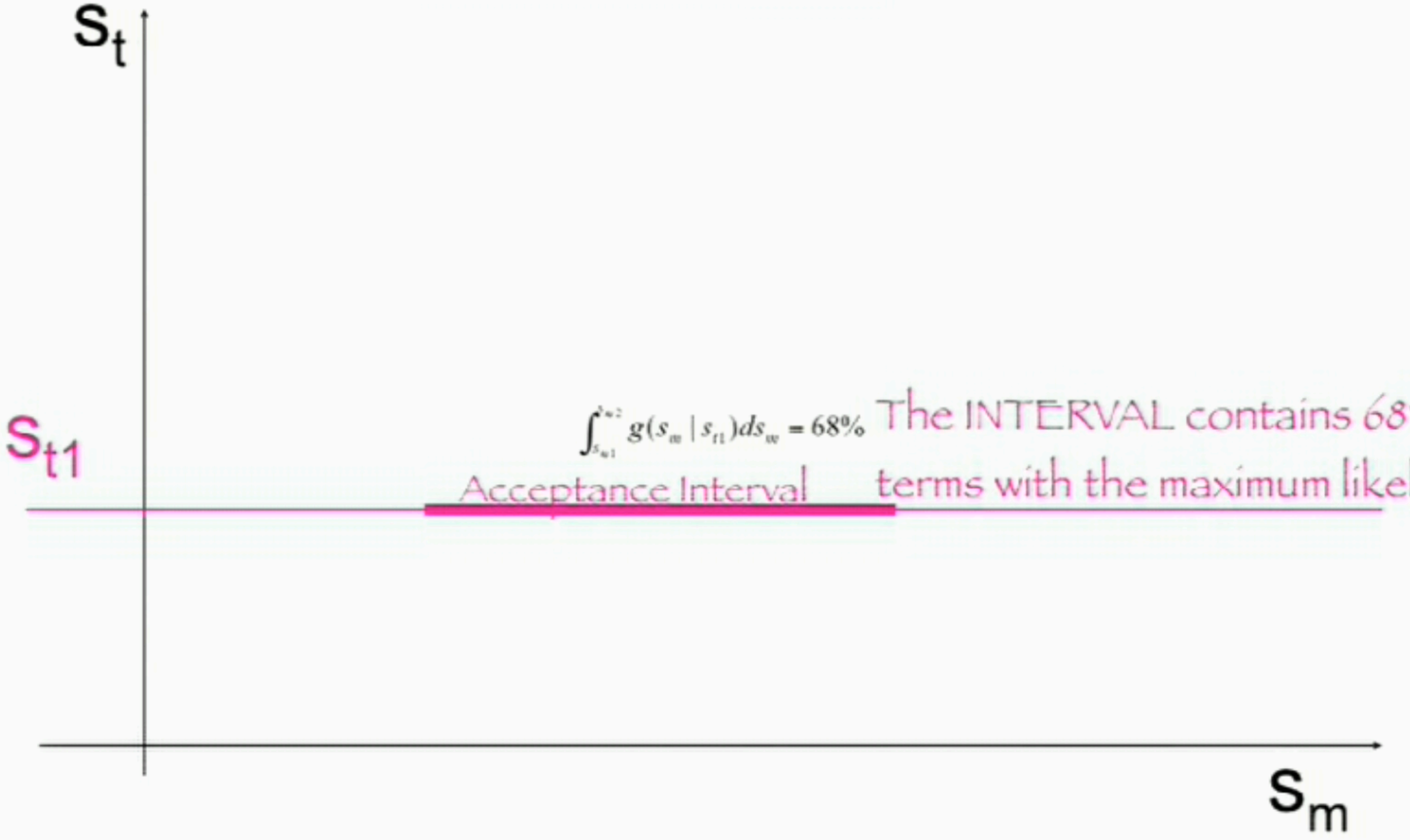
Neyman Construction

$\boxed{\Pr ob(s_m \mid s_t)\ is\ known}$



Neyman Construction

$\boxed{\Pr ob(s_m \mid s_t)\ is\ known}$

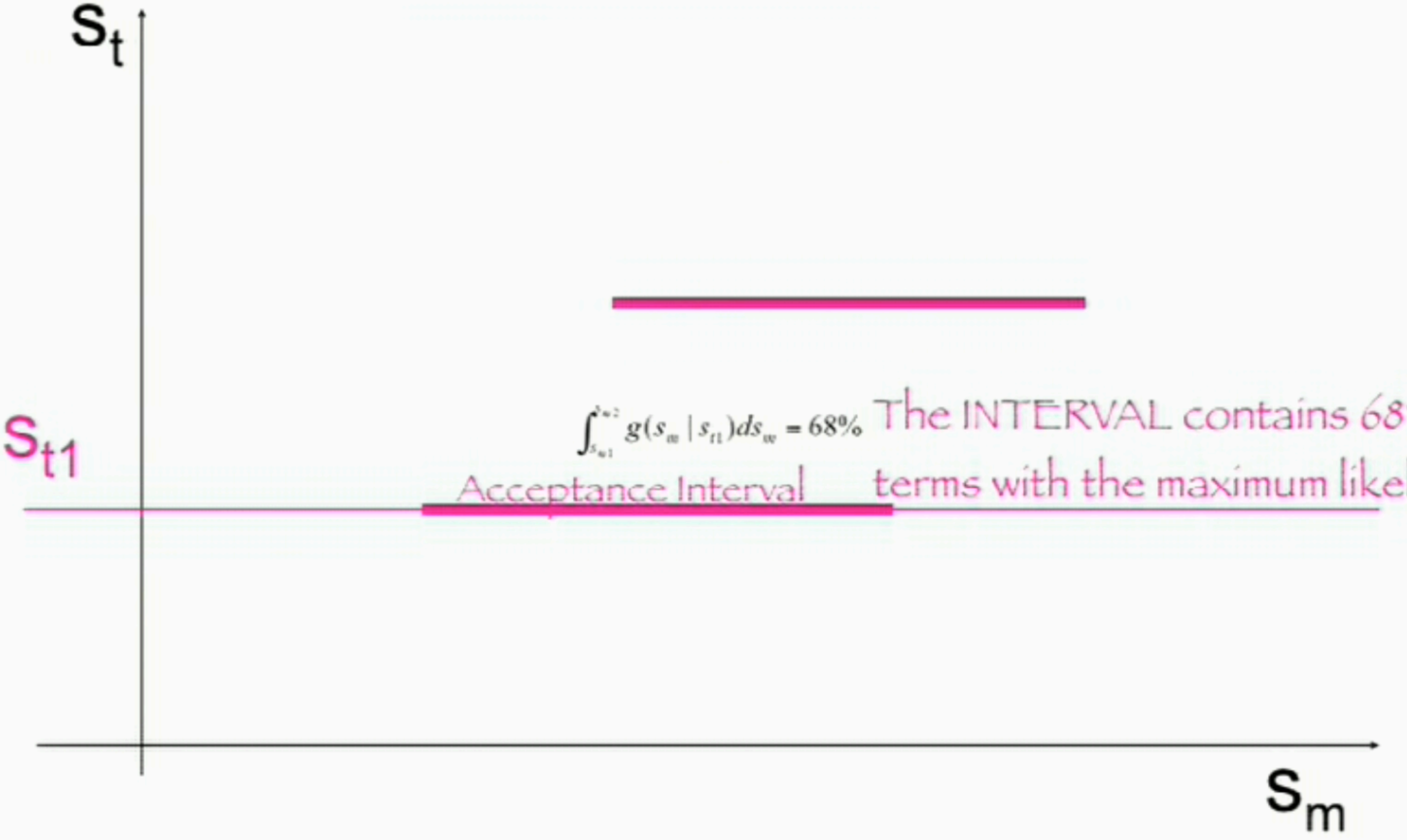


$$\int_{s_{m1}}^{s_{m2}} g(s_m \mid s_{t1})ds_m = 68\%$$

The INTERVAL contains 68% of the terms with the maximum likelihood

Neyman Construction

$\boxed{\Pr ob(s_m \mid s_t)\ is\ known}$

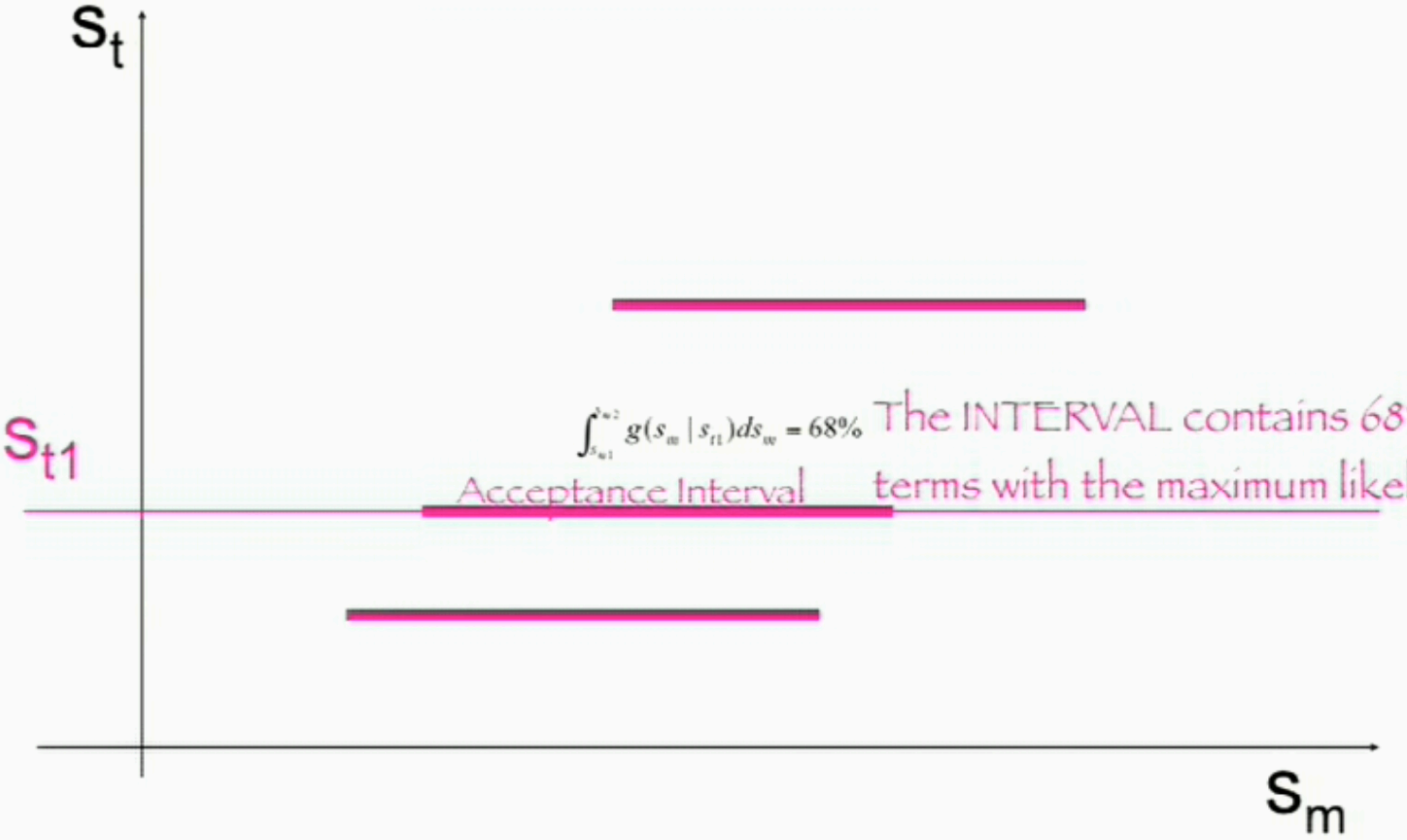


$$\int_{s_{m1}}^{s_{m2}} g(s_m \mid s_{t1})ds_m = 68\%$$

The INTERVAL contains 68% of the terms with the maximum likelihood

Neyman Construction

$\boxed{\Pr ob(s_m \mid s_t)\ is\ known}$

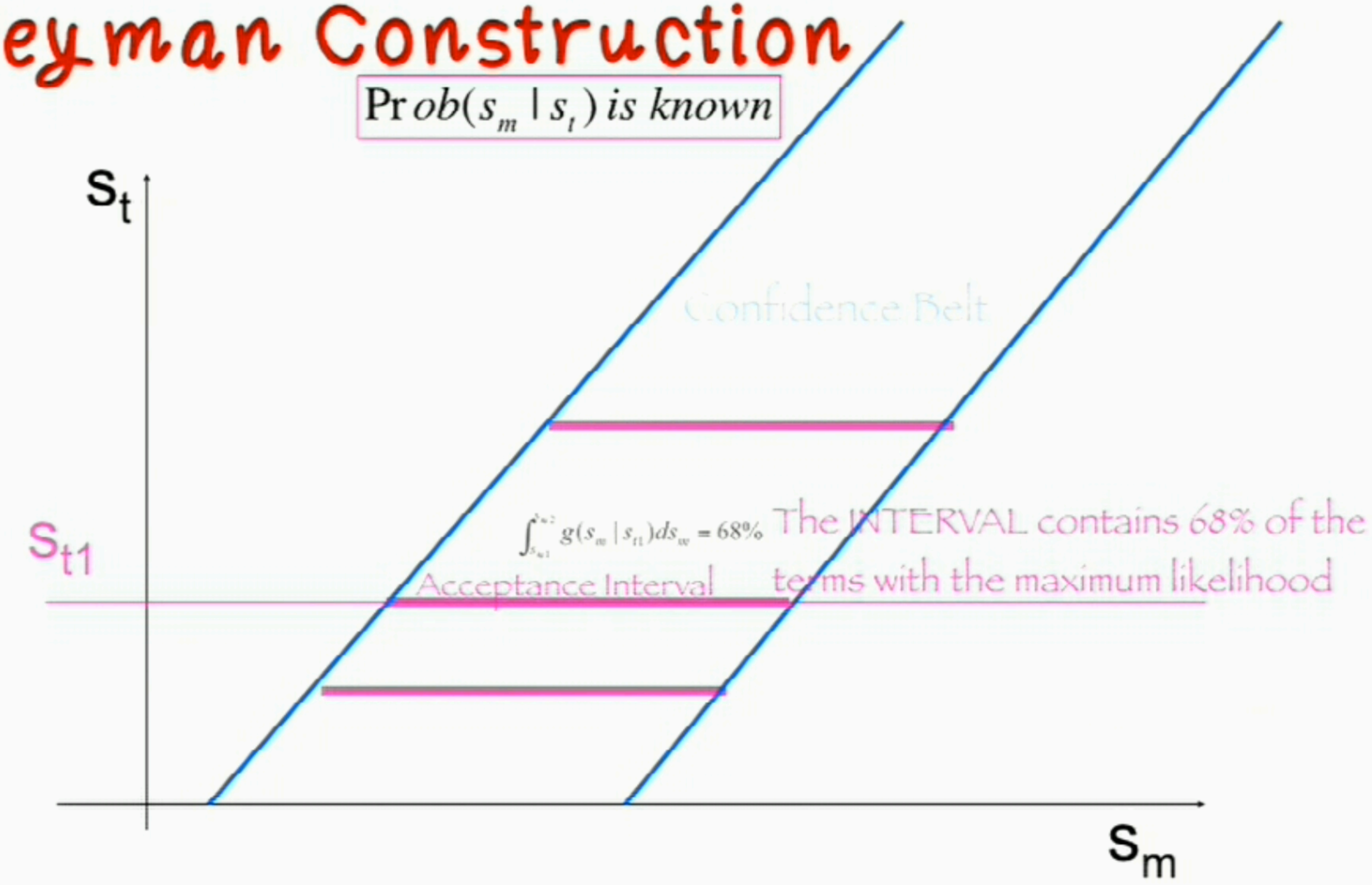


$\int_{s_{m1}}^{s_{m2}} g(s_m \mid s_{t1}) ds_m = 68\%$

The INTERVAL contains 68% of the terms with the maximum likelihood

Neyman Construction

$\Pr ob(s_m \mid s_t)\ is\ known$

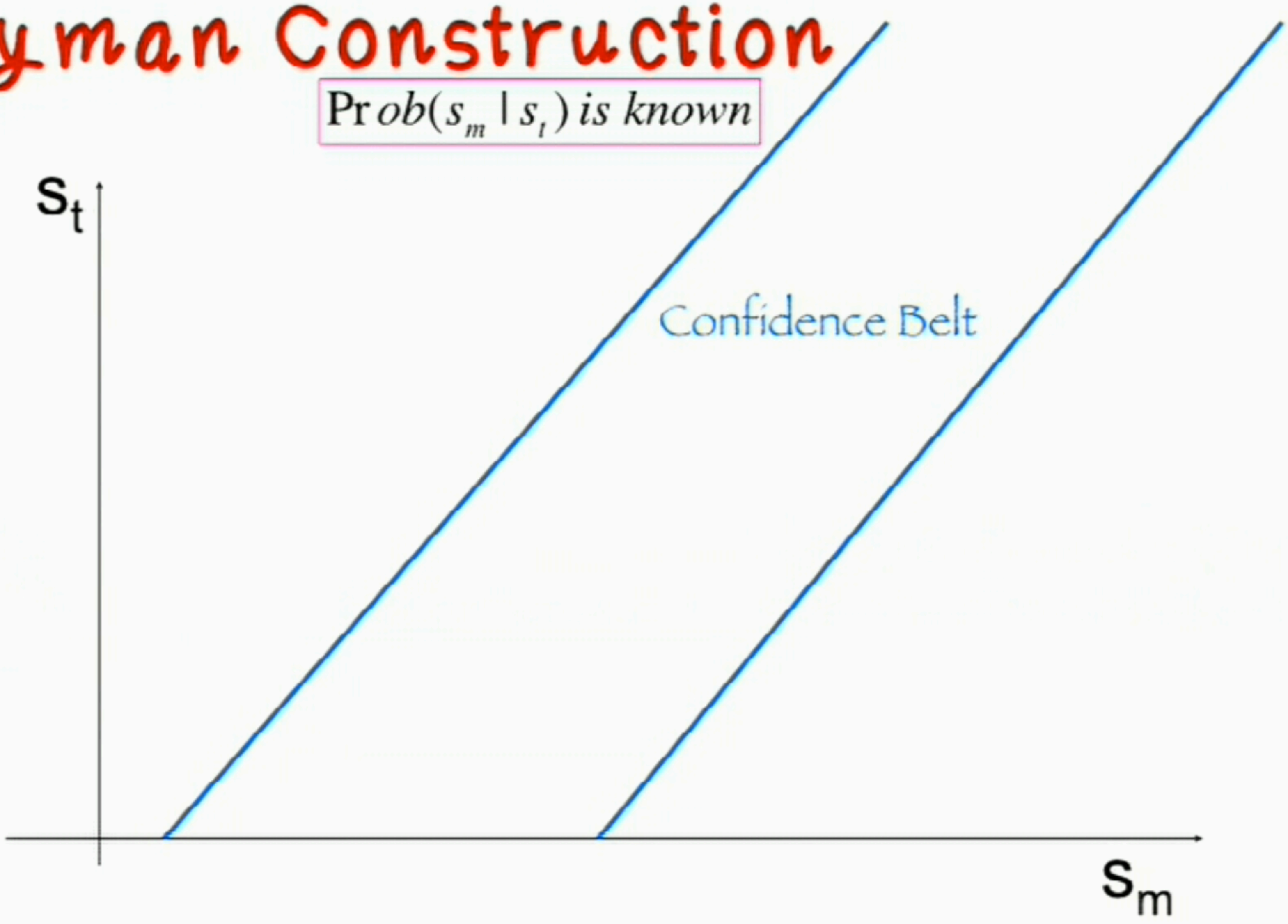


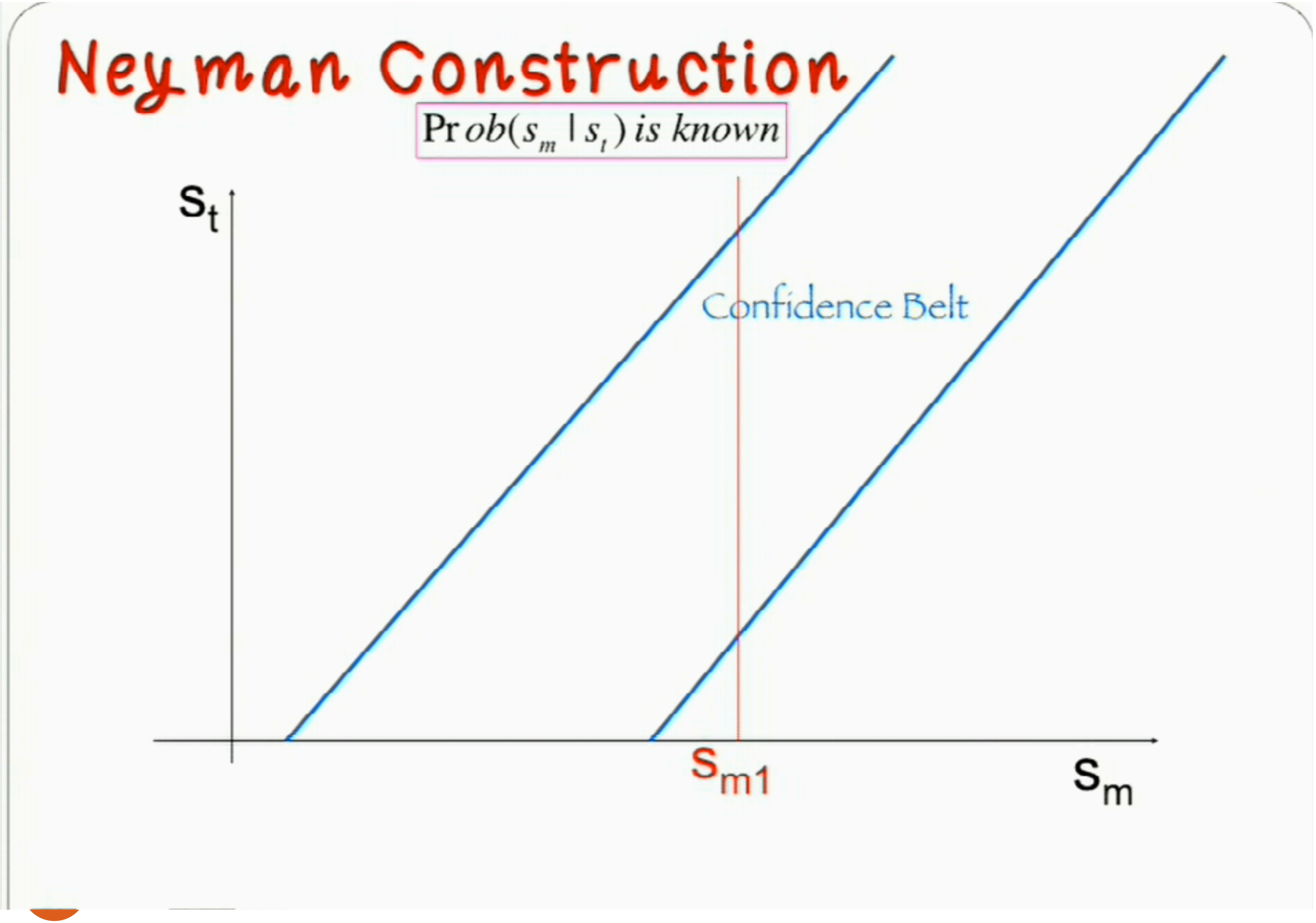
$$\int_{s_{m1}}^{s_{m2}} g(s_m \mid s_{t1})\,ds_m = 68\%$$

The INTERVAL contains 68% of the terms with the maximum likelihood

Neyman Construction

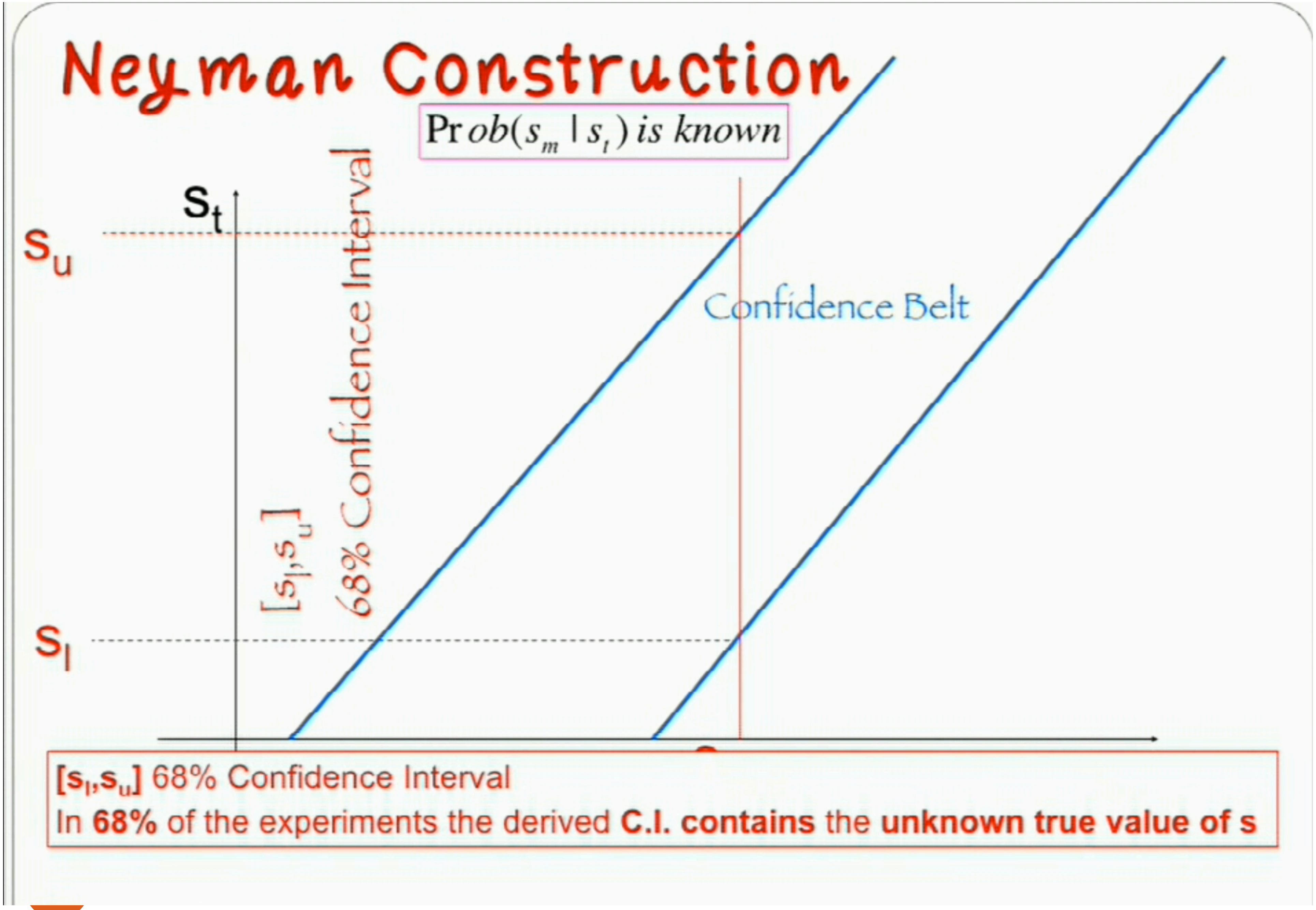
$\Pr ob(s_m \mid s_t)$ is known





Neyman Construction

$\Pr ob(s_m \mid s_t)\ is\ known$



Neyman Construction

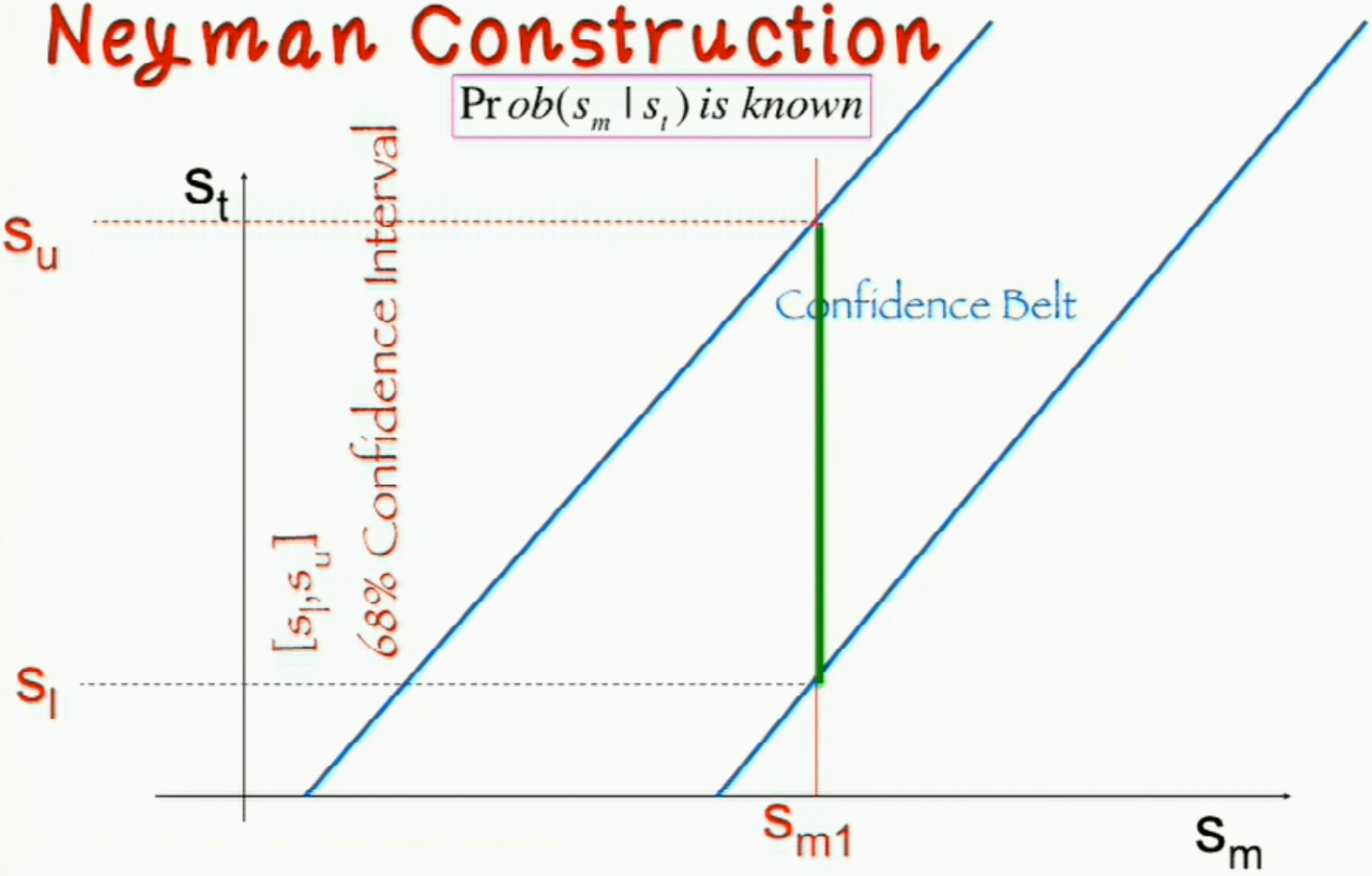
$\Pr ob(s_m \mid s_t)\ is\ known$

[s_l,s_u] 68% Confidence Interval

In **68%** of the experiments the derived **C.I. contains** the **unknown true value of s**

Neyman Construction

$\Pr ob(s_m \mid s_t)\ is\ known$



- With Neyman Construction we guarantee a coverage via construction, i.e. for any value of the unknown true s, the Construction Confidence Interval will cover s with the correct rate.

Neyman Construction

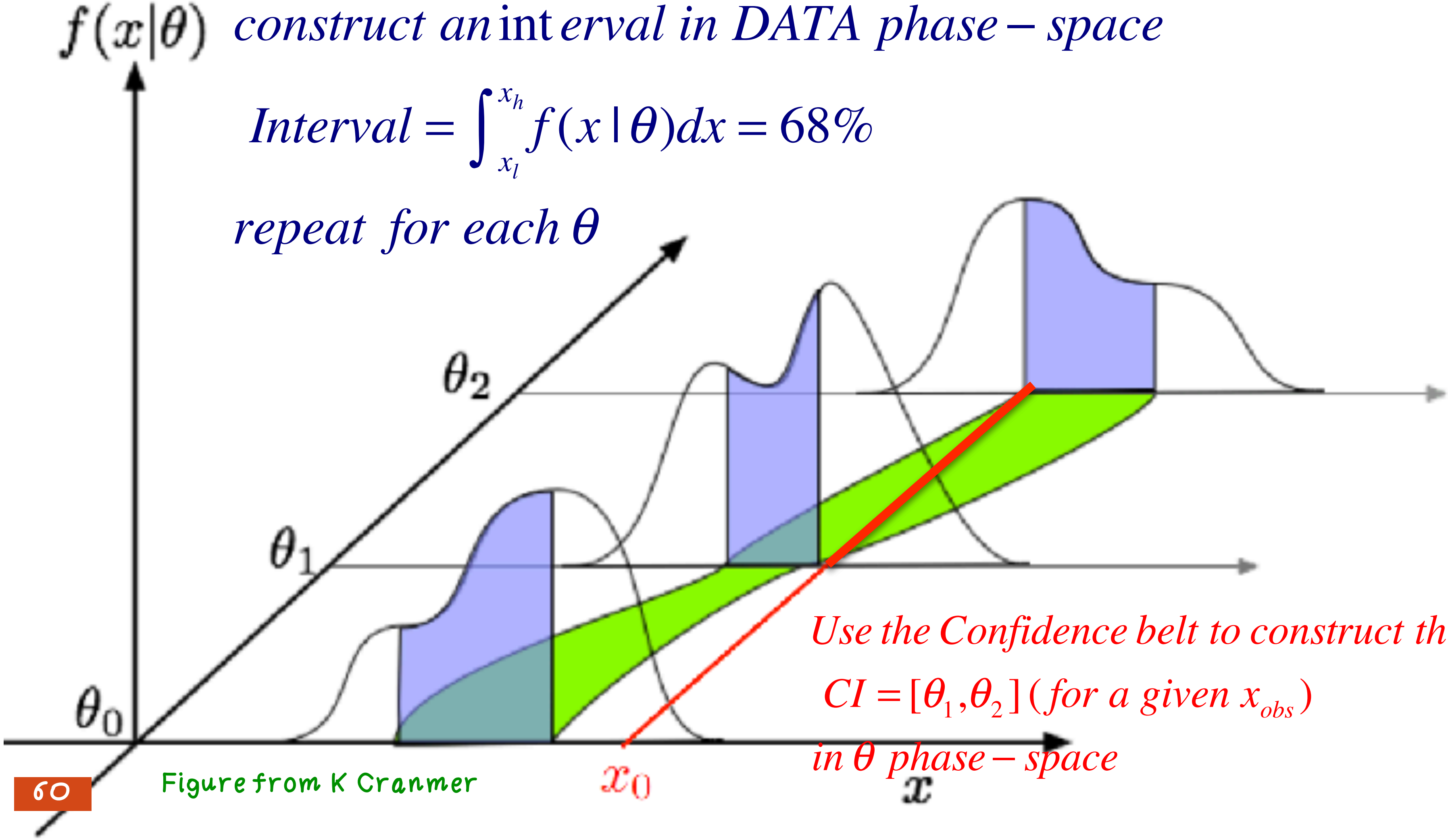
$\theta \equiv s_{true} \qquad x \equiv s_{measured} \qquad pdf\ \ f(x \mid \theta)\ is\ known$

for each prospective θ *generate* x

construct an interval in DATA phase – space

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

repeat for each θ



Use the Confidence belt to construct the

$CI = [\theta_1, \theta_2]\ (for\ a\ given\ x_{obs})$

in θ *phase – space*

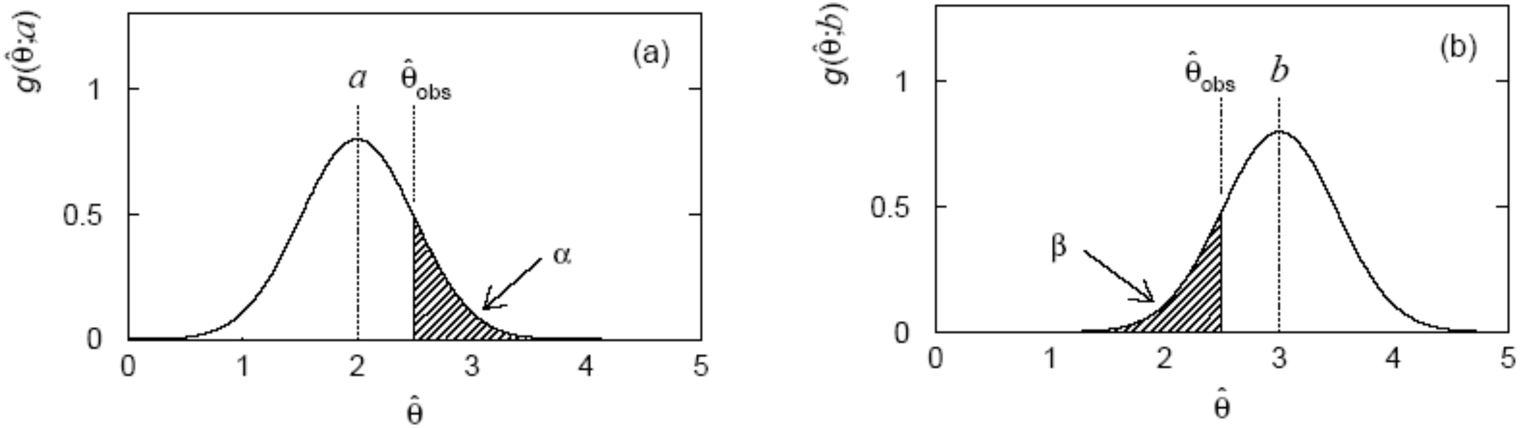
Figure from K Cranmer

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 θ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha$.

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

Meaning of a confidence interval

N.B. the interval is random, the true θ 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:

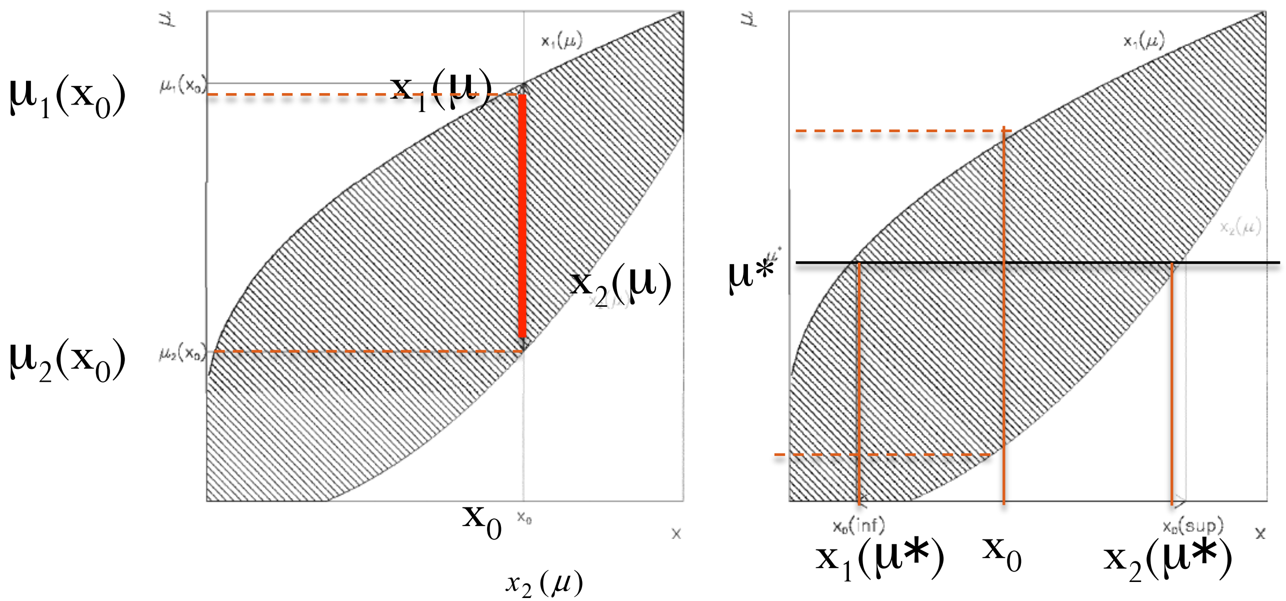
$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

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 θ.

Neyman’s construction



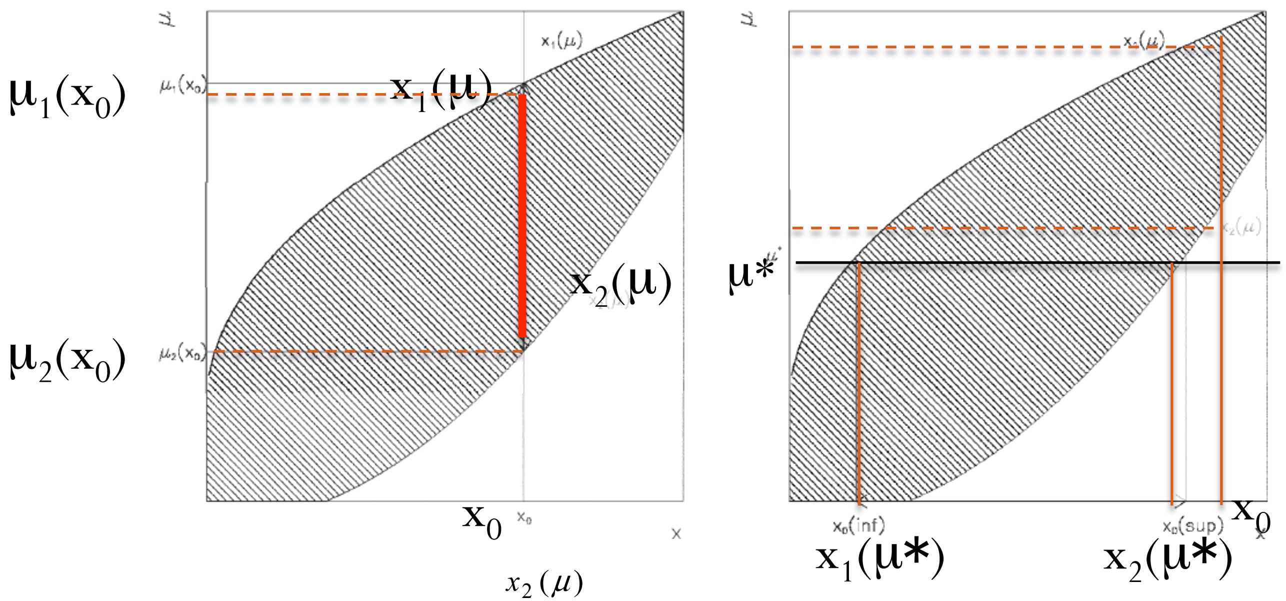
$$\int_{x_1(\mu)}^{x_2(\mu)} f(x/\mu)dx = \alpha$$

By construction the probability to measure $x_0' < x_0$ if the true value $\mu = \mu_1(x_0)$ is $(1-\alpha)/2$
$x_0' > x_0$ if the true value $\mu = \mu_2(x_0)$ is $(1-\alpha)/2$

Coverage: suppose μ^* the true value

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

Neyman's construction



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

By construction the probability to measure $x_0' < x_0$ if the true value $\mu=\mu_1(x_0)$ is $(1-\alpha)/2$
$x_0' > x_0$ if the true value $\mu=\mu_2(x_0)$ is $(1-\alpha)/2$

Coverage: suppose μ^* the true value

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

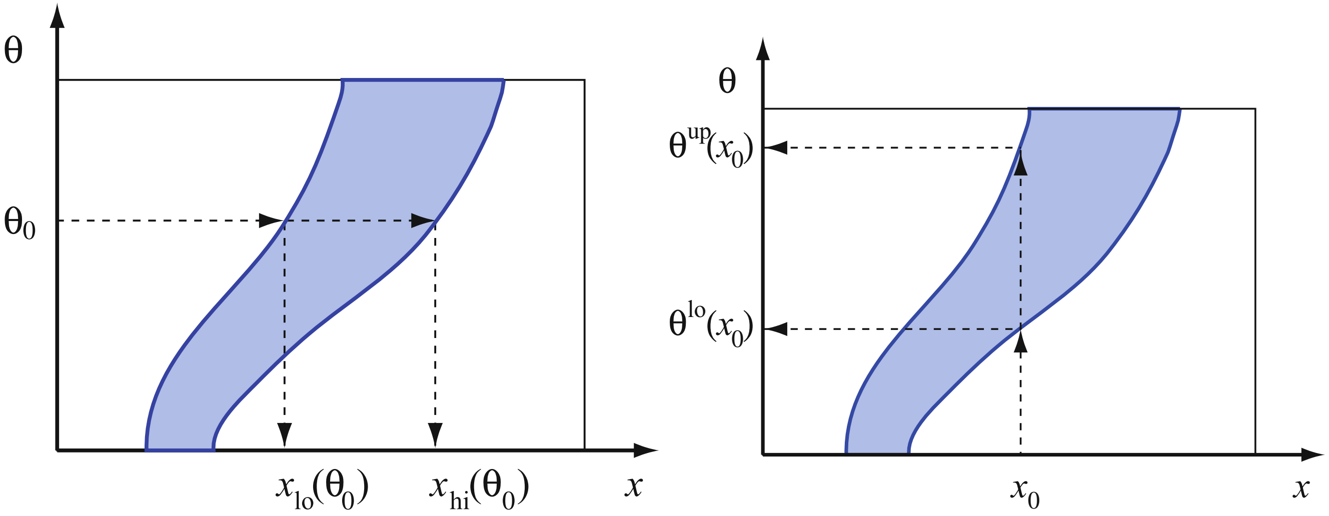


Fig. 7.1 Graphical illustration of Neyman belt construction (*left*) and inversion (*right*)

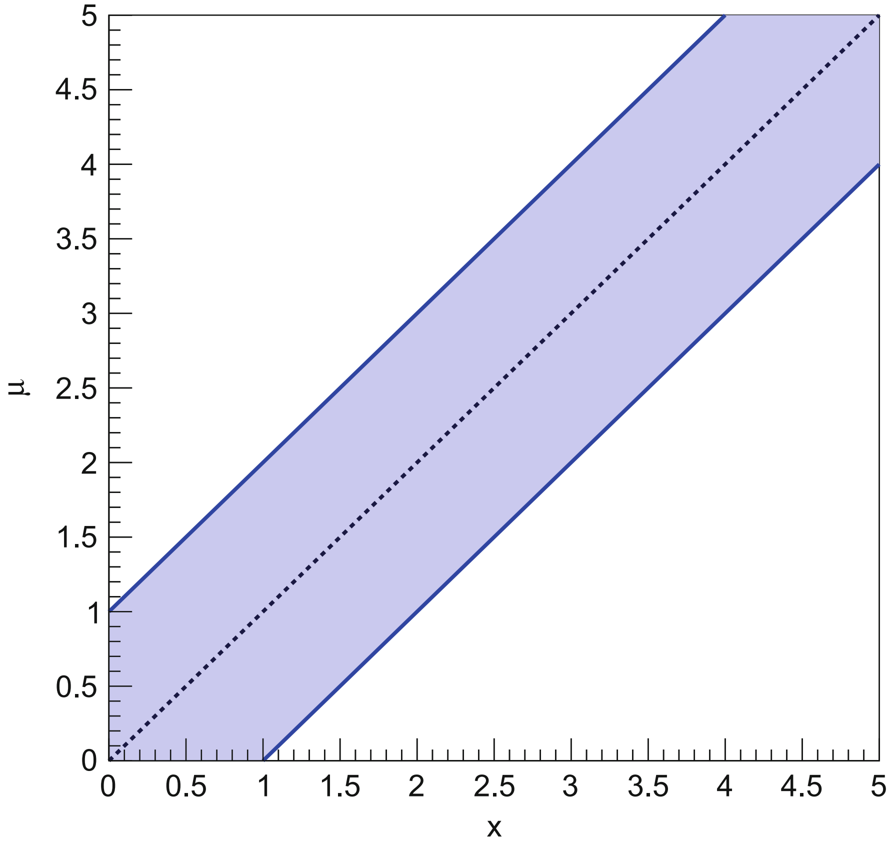


Fig. 7.3 Neyman belt for the parameter μ of a Gaussian with $\sigma = 1$ at the 68.27% confidence level

Suppose Poisson variable and n=0 is measured (no background) Upper limit (lower limit =0)
=> 0±0 (freq) or 1±1 (Bayes) ?

By construction the probability to measure $x_0' < x_0$ if the true value $\mu=\mu_1(x_0)$ is $(1-\alpha)$ (only one limit) or the probability to measure $x_0' > x_0$ if the true value $\mu=\mu_1(x_0)$ is α

$$P(n>0/\lambda)=\sum_{n=1}^{\infty}\frac{\lambda^{n}e^{-\lambda}}{n!}=1-e^{-\lambda}=\alpha$$

frequentist

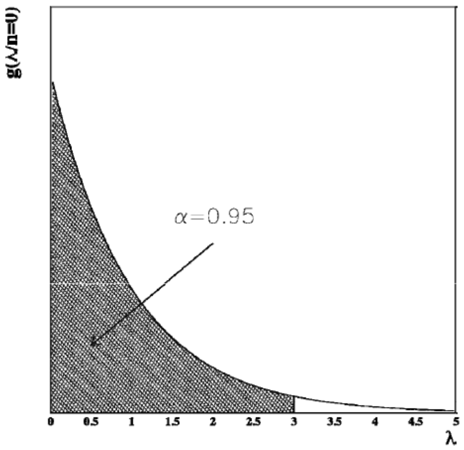
$$\overline{\lambda}=-\ln(1-\alpha)$$

$$g(\lambda/n=0)=\frac{p(n=0/\lambda)f_0(\lambda)}{\int_0^{\infty}p(n=0/\lambda)f_0(\lambda)d\lambda}=\frac{e^{-\lambda}}{\int_0^{\infty}e^{-\lambda}d\lambda}=e^{-\lambda}$$

Bayesian
(uniform prior)

$$p(\lambda<\overline{\lambda})=\int_0^{\overline{\lambda}}e^{-\lambda}d\lambda=1-e^{-\overline{\lambda}}=\alpha$$

	90%	95%	99%
$\overline{\lambda}$	2.3	3.0	4.6



Poisson

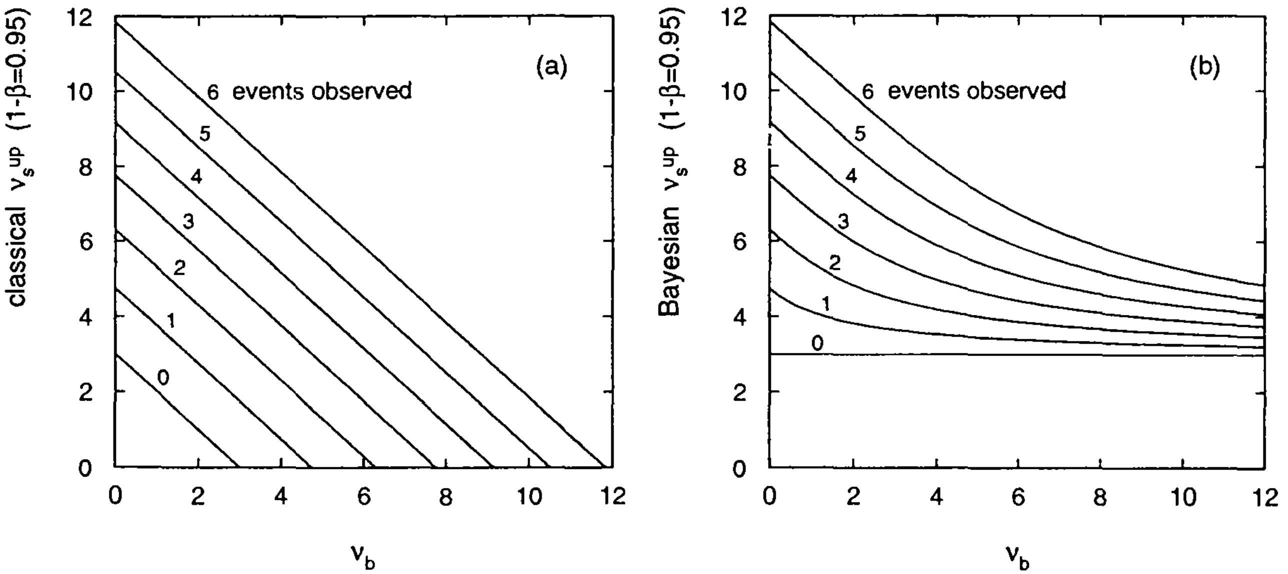


Fig. 9.9 Upper limits ν_s^{up} at a confidence level of $1 - \beta = 0.95$ for different numbers of events observed n_{obs} and as a function of the expected number of background events ν_{b}. (a) The classical limit. (b) The Bayesian limit based on a uniform prior density for ν_s.