Counting

- So we do collisions at a given $\sqrt{s}$. What do we actually measure ?
- We "count" the number of times a final state is obtained. This frequency is somehow related to the probability of that final state and so it allows to measure the cross-section/decay width/branching ratios
- Connection btw probability and frequency:
  - Population → probability
  - Sample → frequency
- Sampling fluctuations

Random Variables – Outline - I

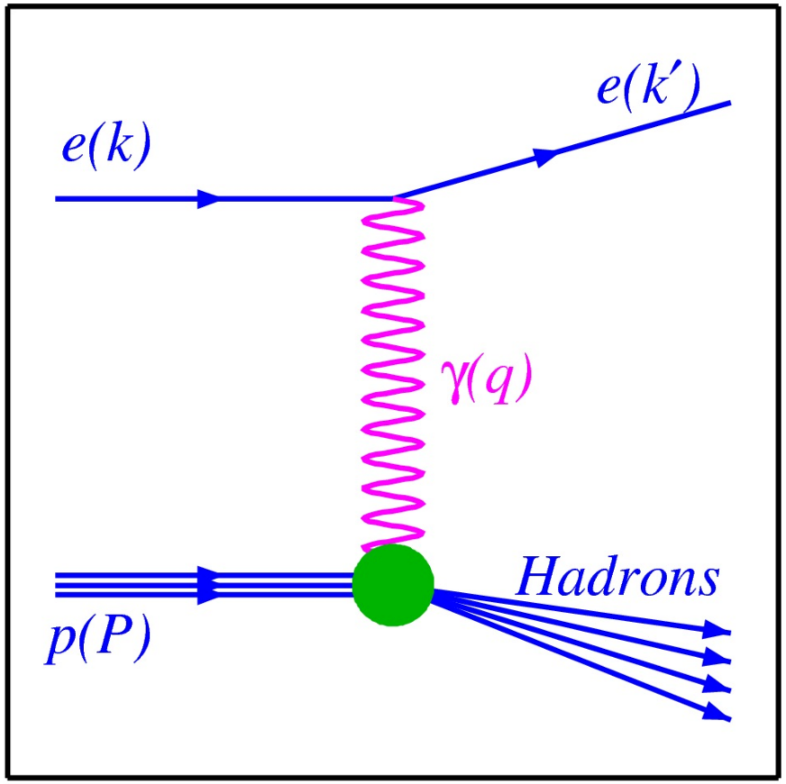
- Concept of PDF
  - Meaning and connection to actual probabilities
  - Discrete vs. real variables
  - Single vs. multiple variables: factorization
- Definitions/properties
  - Physical dimension, positivity, normalization
  - Momenta → "functional"
  - Mean, variance, standard deviation, skewness, kurtosys
  - Covariance matrix
  - Propagation

Random variables - II

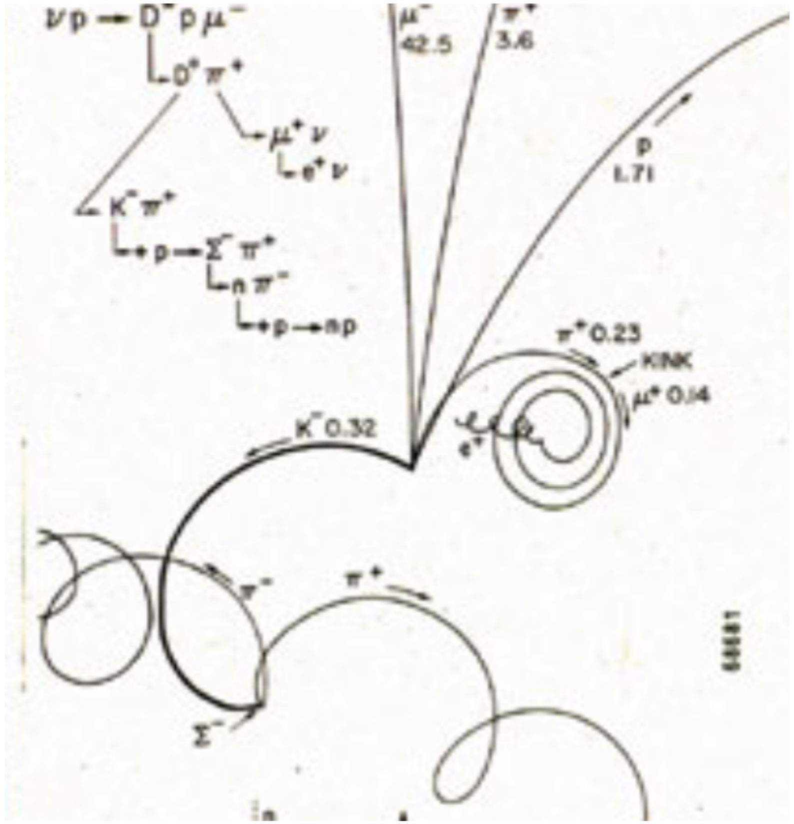
- The average and the RMS: two particular and interesting random variables, functions of random variables
- Few random variables that allow to have good statistical models of typical situations in experimental physics:
  - Binomial
  - Poissonian
  - Exponential
  - Gaussian
  - χ^2
- BUT: up to here only "populations"
- =>Statistical inference

Event: a "photo" of a collision/decay

Inclusive Event: measure the electron only



Exclusive Event: measure all particles to "close" the kinematics



"Logic" of an EPP experiment - II

- An ***ideal detector*** allows to measure the quadri-momentum of each particle involved in the reaction.
  - Direction of flight;
  - Energy $\boldsymbol{E}$ and/or momentum modulus $|\boldsymbol{p}|$;
  - Which particle is (e.g. from independent measurements of $\boldsymbol{E}$ and $|\boldsymbol{p}|$, $\boldsymbol{m}^2=\boldsymbol{E}^2-|\boldsymbol{p}|^2$) ➔ Particle ID
- BUT for a ***real detector***:
  - Not all quadri-momenta are measured: some particles are out of acceptance, or only some quantities are accessible, there are unavoidable **inefficiencies**;
  - Measurements are affected by **resolution**
  - Sometimes the particle nature is "confused"

“Logic” of an EPP experiment - III

- Selection steps:
  1. **TRIGGER SELECTION**
     - Retain only “interesting events”: from bubble chambers to electronic detectors
     - ➔ “logic-electronic” eye: decides in a short time O(μs) if the event is interesting or not.
     - In some cases (e.g. pp), it is crucial since interactions are so probable…
     - LHC: every 25 ns is a bunch crossing giving rise to interactions: can I write 40 MHz on “tape” ? A tipical event has a size of 1 MB ➔ 40 TB/s. Is it conceivable ? And how many CPU will be needed to analyze these data ? At LHC from 40 MHz to 200 Hz ! Only one bunch crossing every 200000 !
     - “pre-scale” is an option
     - e^+e^-: the situation is less severe but a trigger is in any case necessary.

"Logic" of an EPP experiment - IV

2. **EVENT RECONSTRUCTION**: Once you have the final event sample, for each trigger you need to reconstruct at your best the kinematic variables.
3. **OFFLINE SELECTION**: choice of a set of discriminating variables on which apply one of the following:
   - cut-based selection
   - discriminating variables selection
   - multivariate classifier selection
4. **PHYSICS ANALYSIS**: analysis of the sample of ***CANDIDATES***

The selection strategy is a crucial part of the experimentalist work: defined and optimized using *simulated data samples*.

“Logic” of an EPP experiment - V

- Simulated samples of events: the Montecarlo.
  - “Physics” simulation: final state with correct kinematic distributions; also dynamics in some cases is relevant.
  - “Detector” simulation: the particles are traced through the detector, interactions, decays, are simulated.
  - “Digitization”: based on the particle interactions with the detector, signals are simulated with the same features of the data.
- ➔ For every interesting final state MC samples with the same format of a data sample are built. These samples can be analyzed with the same program. In principle one could run on a sample without knowing if it is data or MC.
- To design a “selection” strategy for a given searched signal one needs: ***signal MC samples*** and ***background MC samples***.

“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.

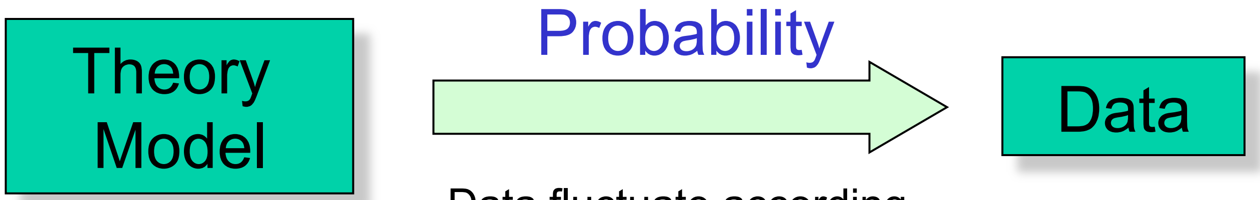
$$\varepsilon N_X = N_{cand} - N_b$$

- where:
  - ε = efficiency ($0<\varepsilon<1$); $\boldsymbol{\varepsilon} = \boldsymbol{A} \times \boldsymbol{\varepsilon_d}$
  - N_b = number of background events
- Estimate ε 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.

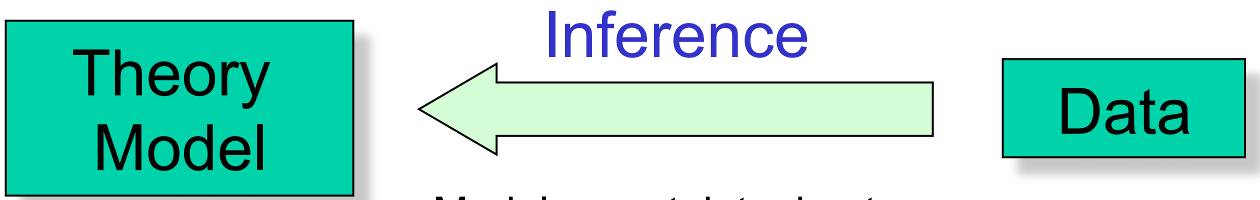
Meaning of parameter estimate

- 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

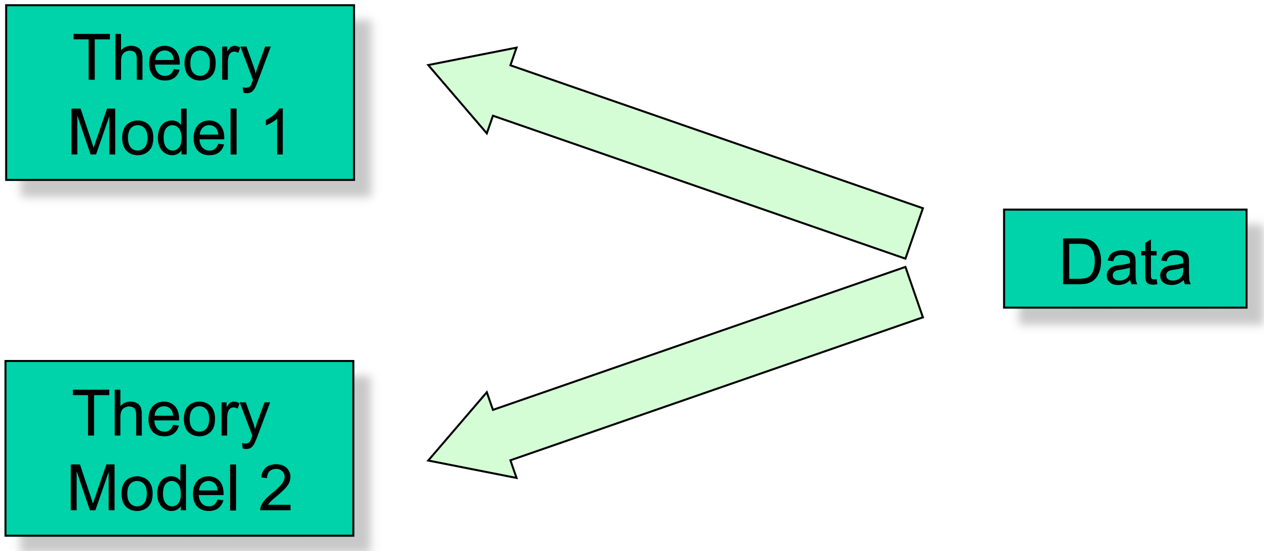


Data fluctuate according
to process randomness



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.

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.

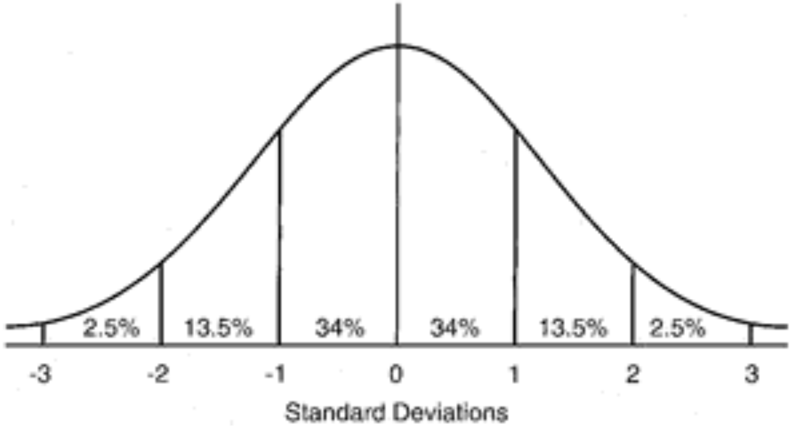
Confidence Interval & Coverage

- 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

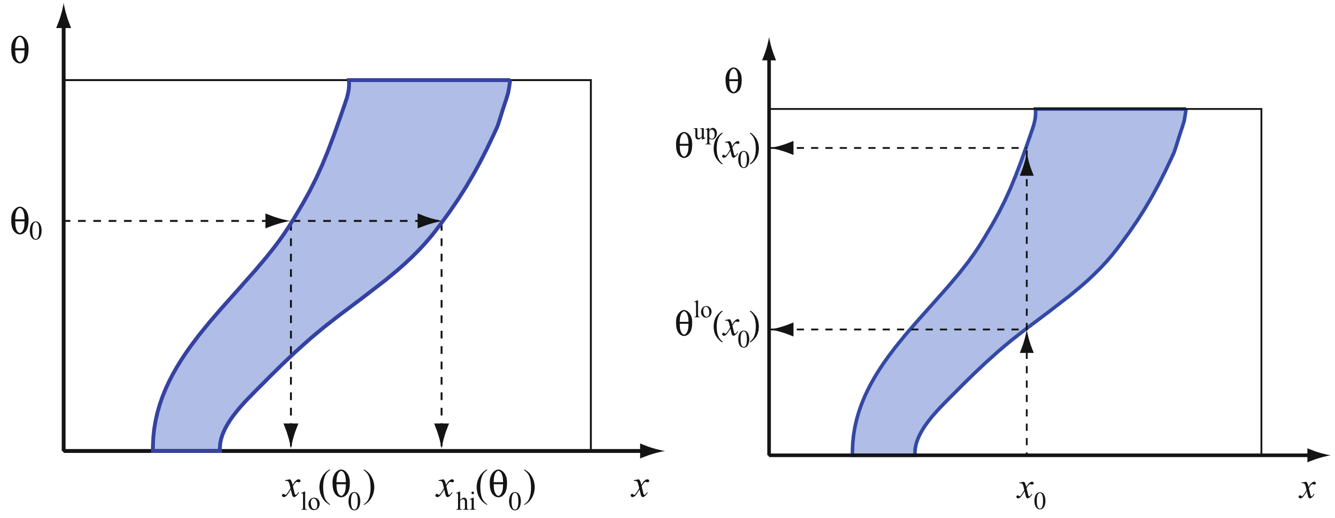
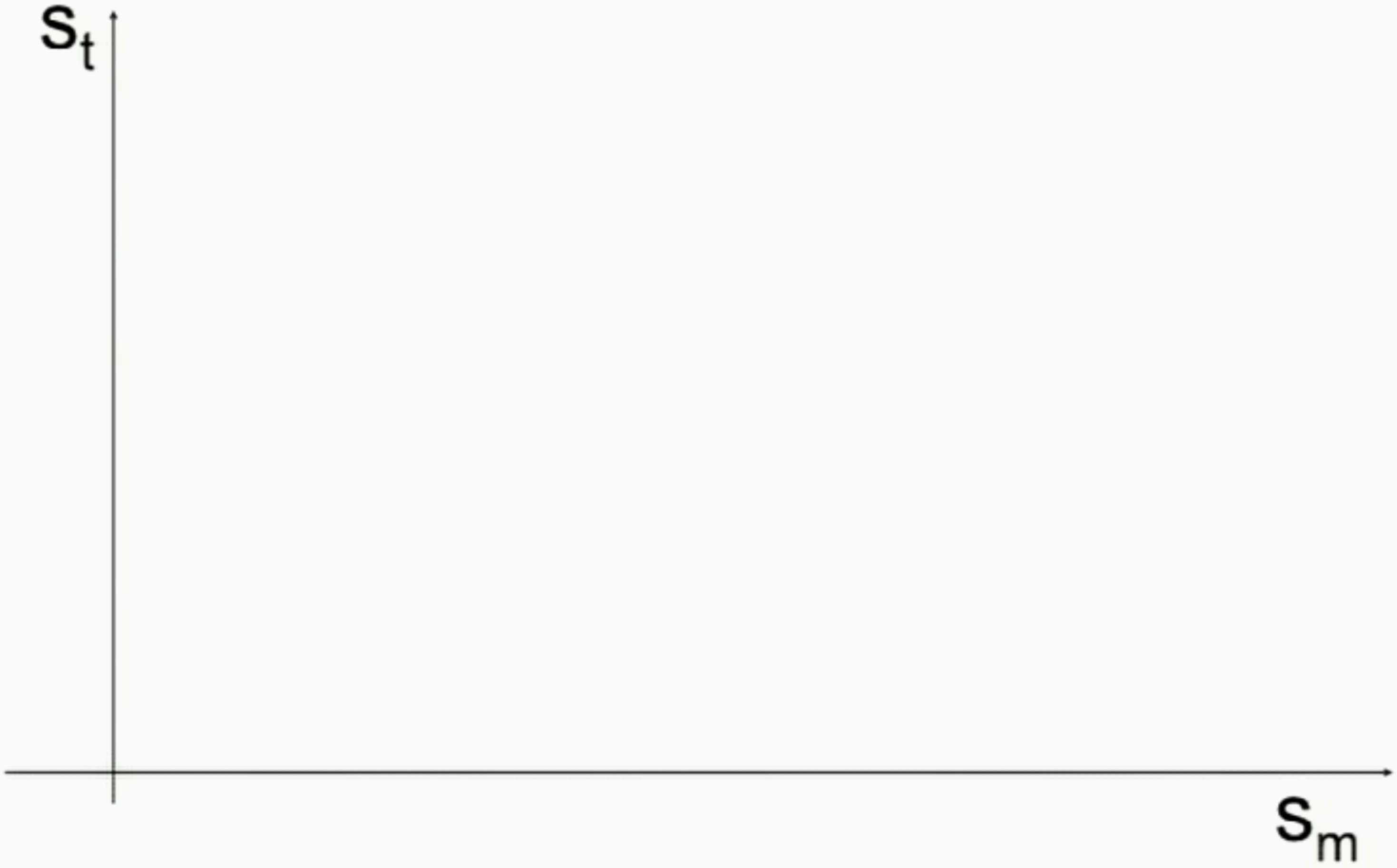


Fig. 7.1 Graphical illustration of Neyman belt construction (*left*) and inversion (*right*)

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

Neyman Construction

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

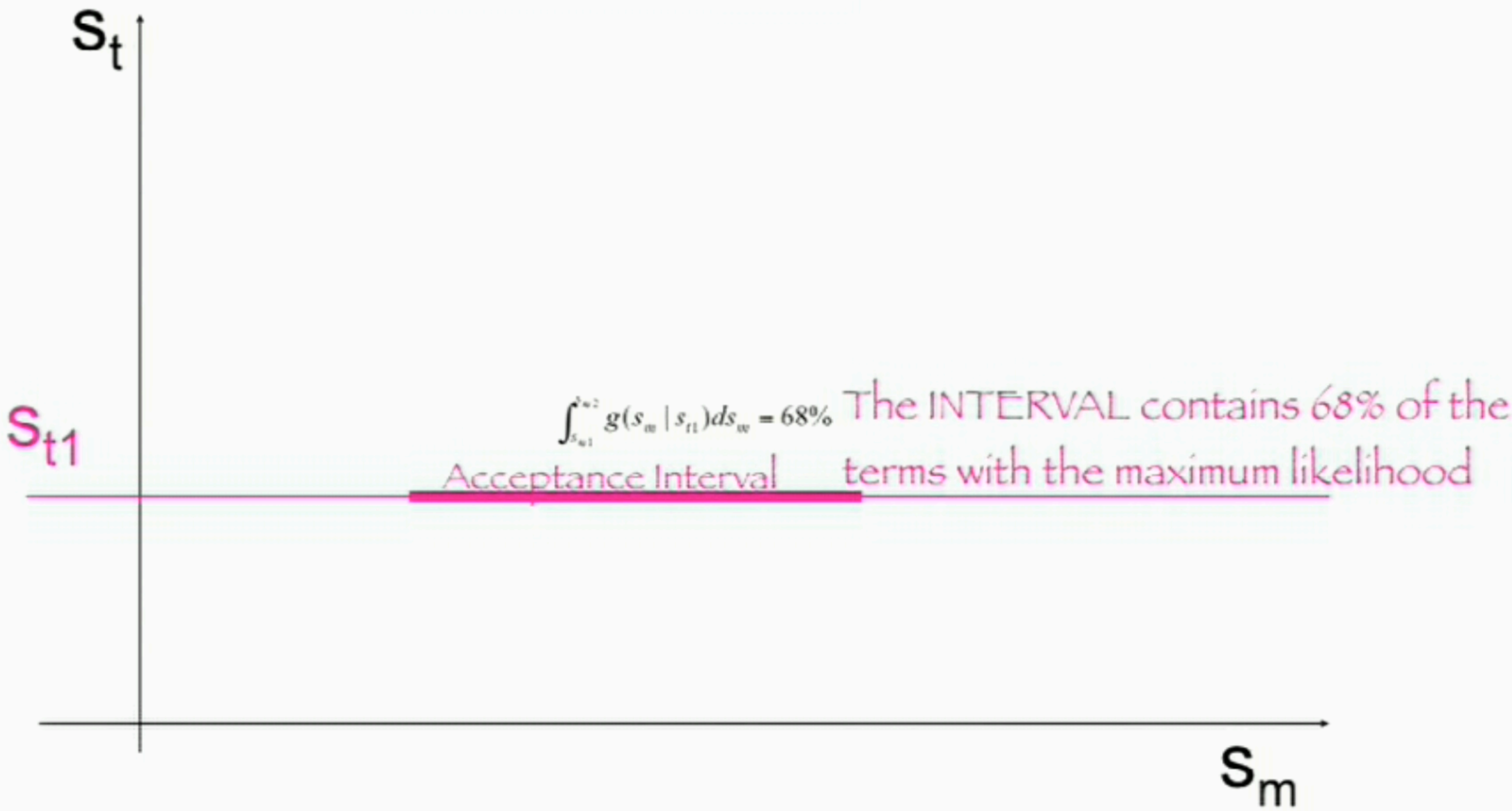


Neyman Construction

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

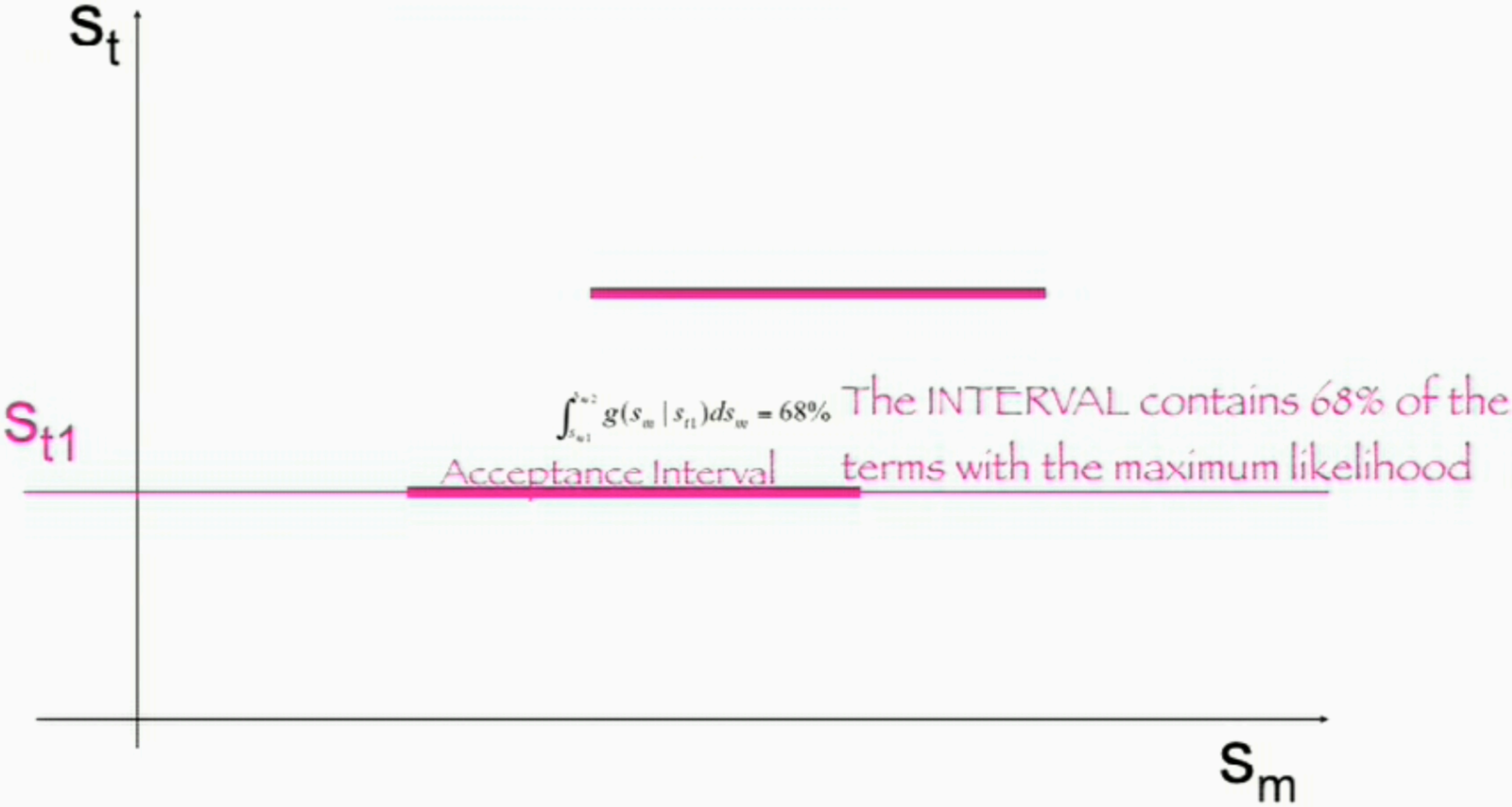
Neyman Construction

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



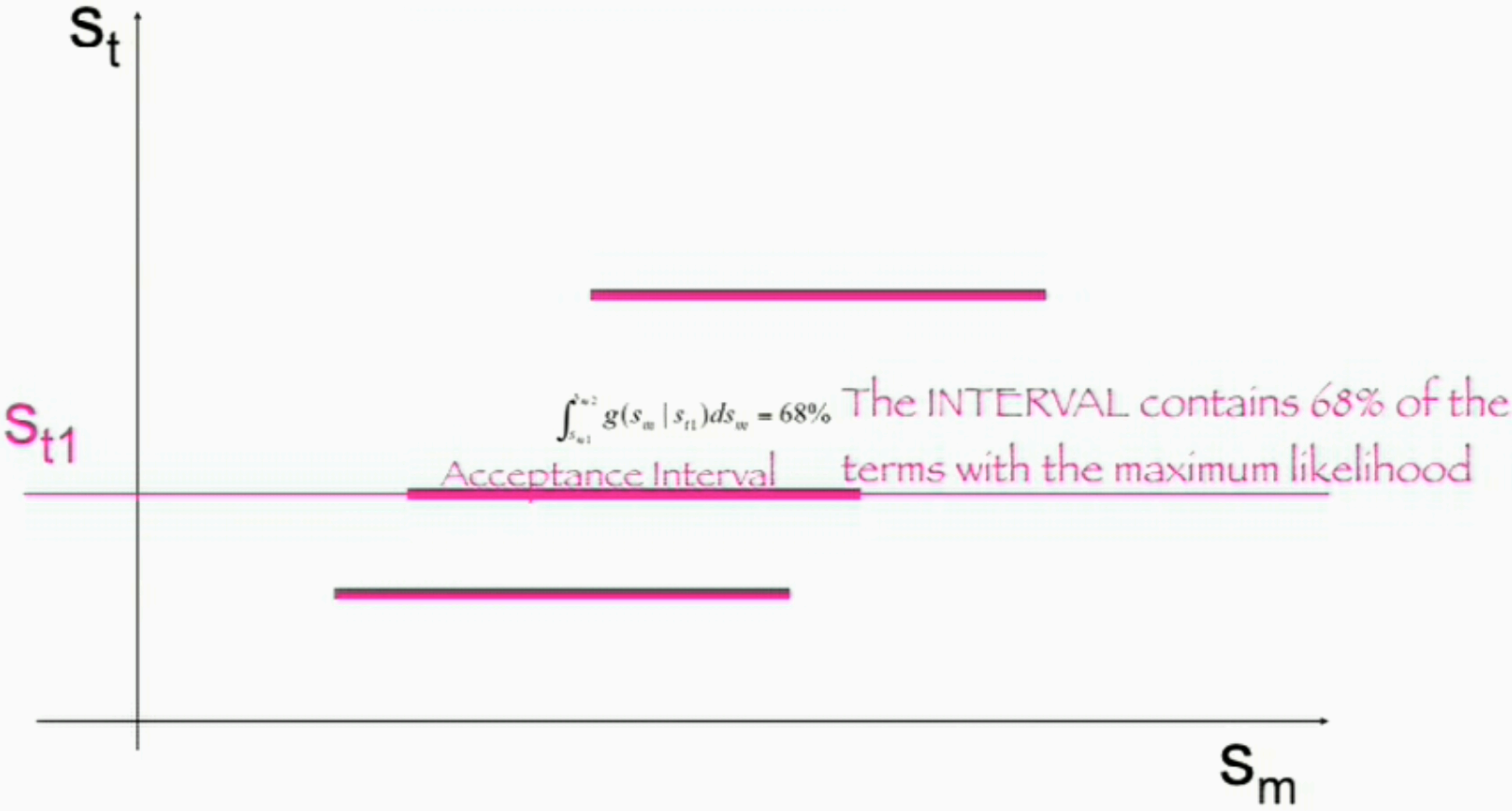
Neyman Construction

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

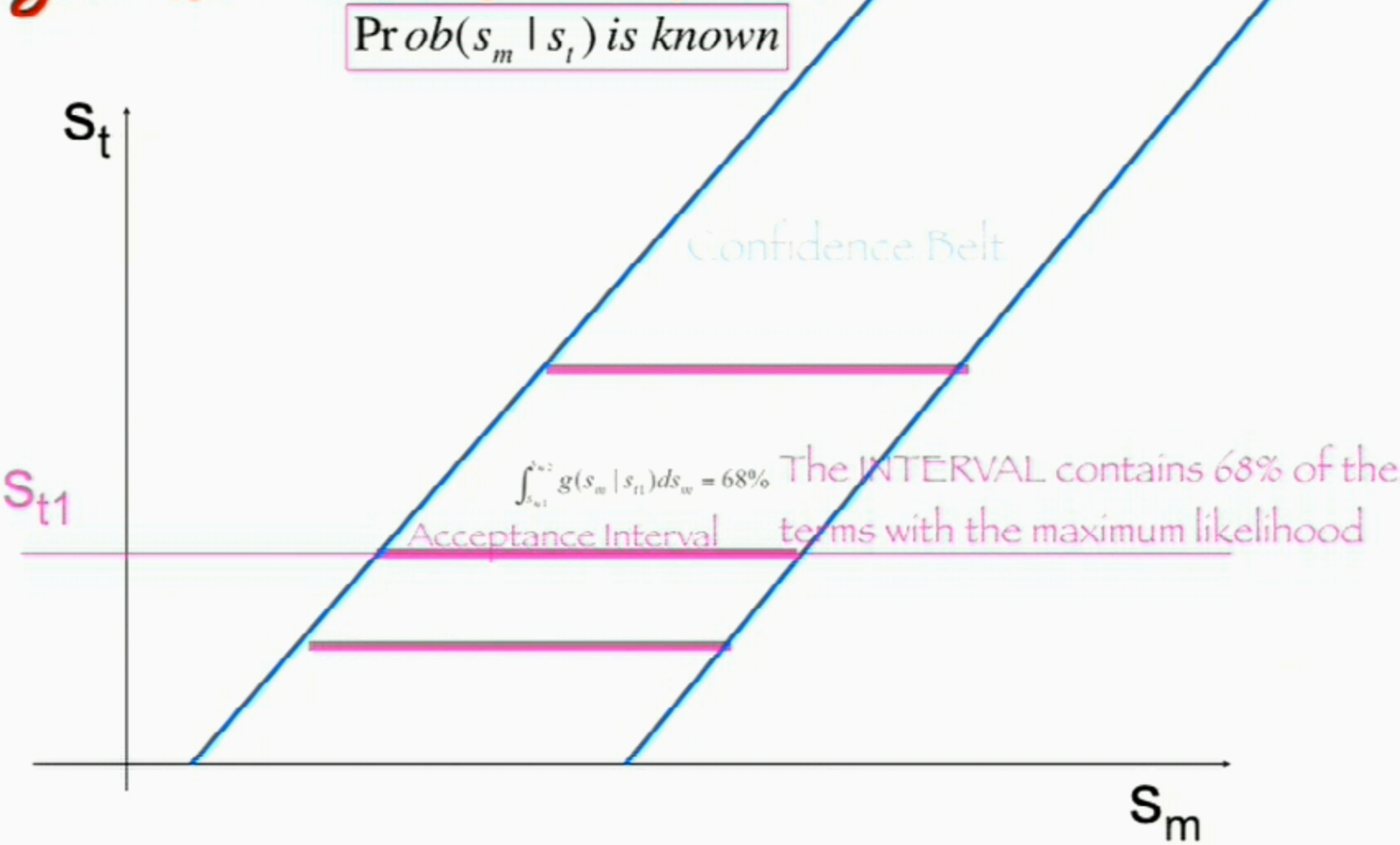


Neyman Construction

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

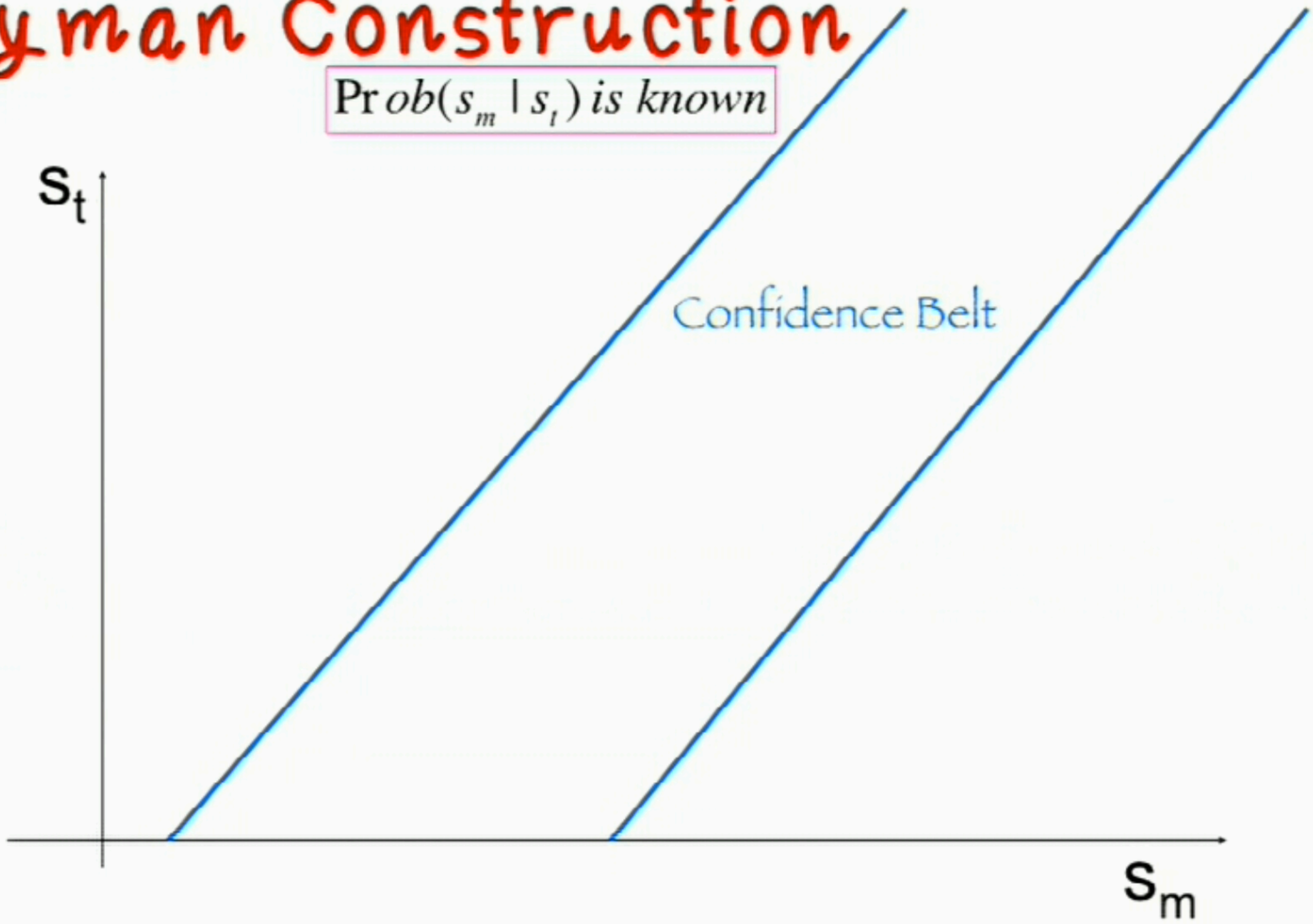


Neyman Construction



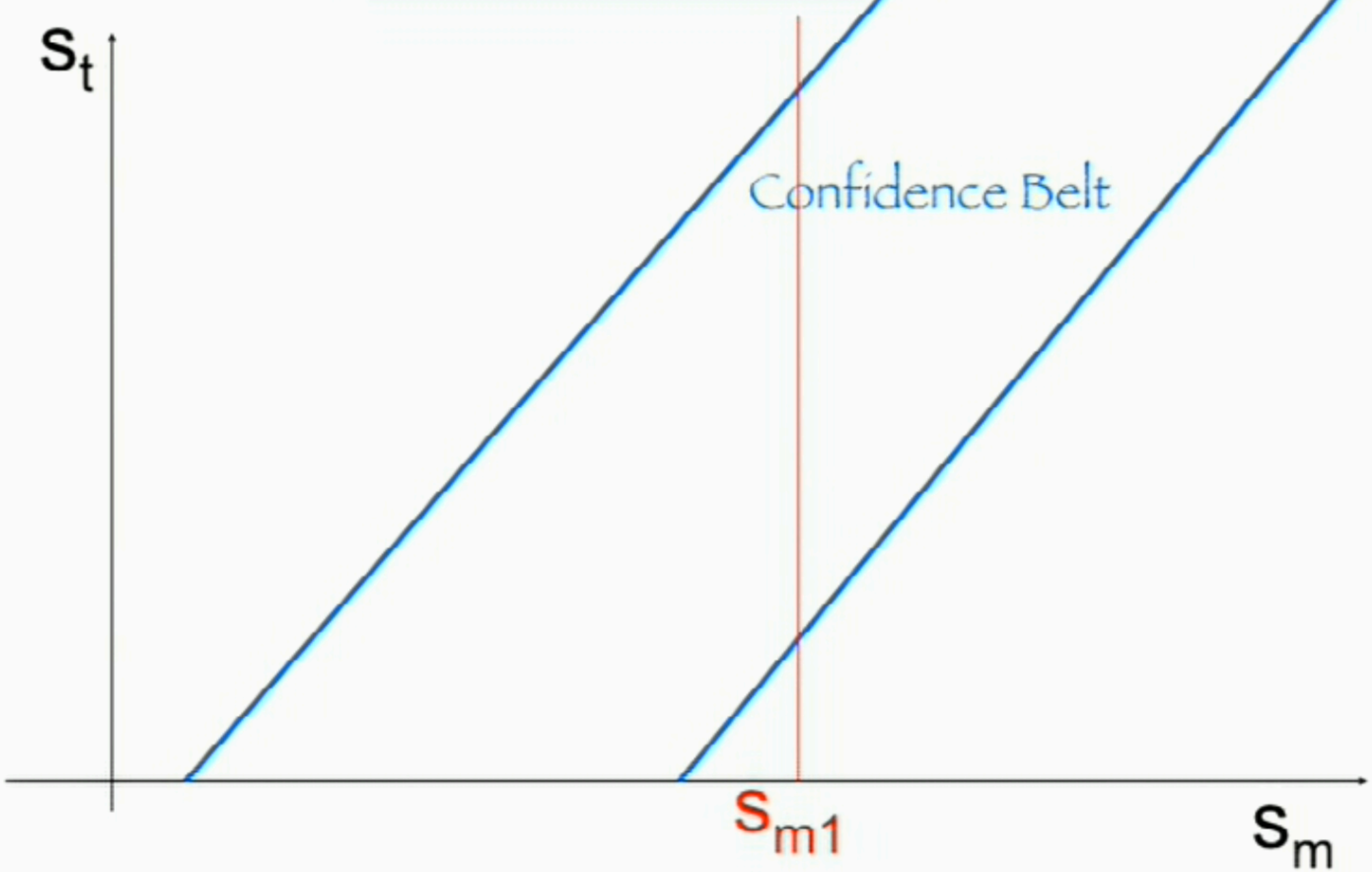
Neyman Construction

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



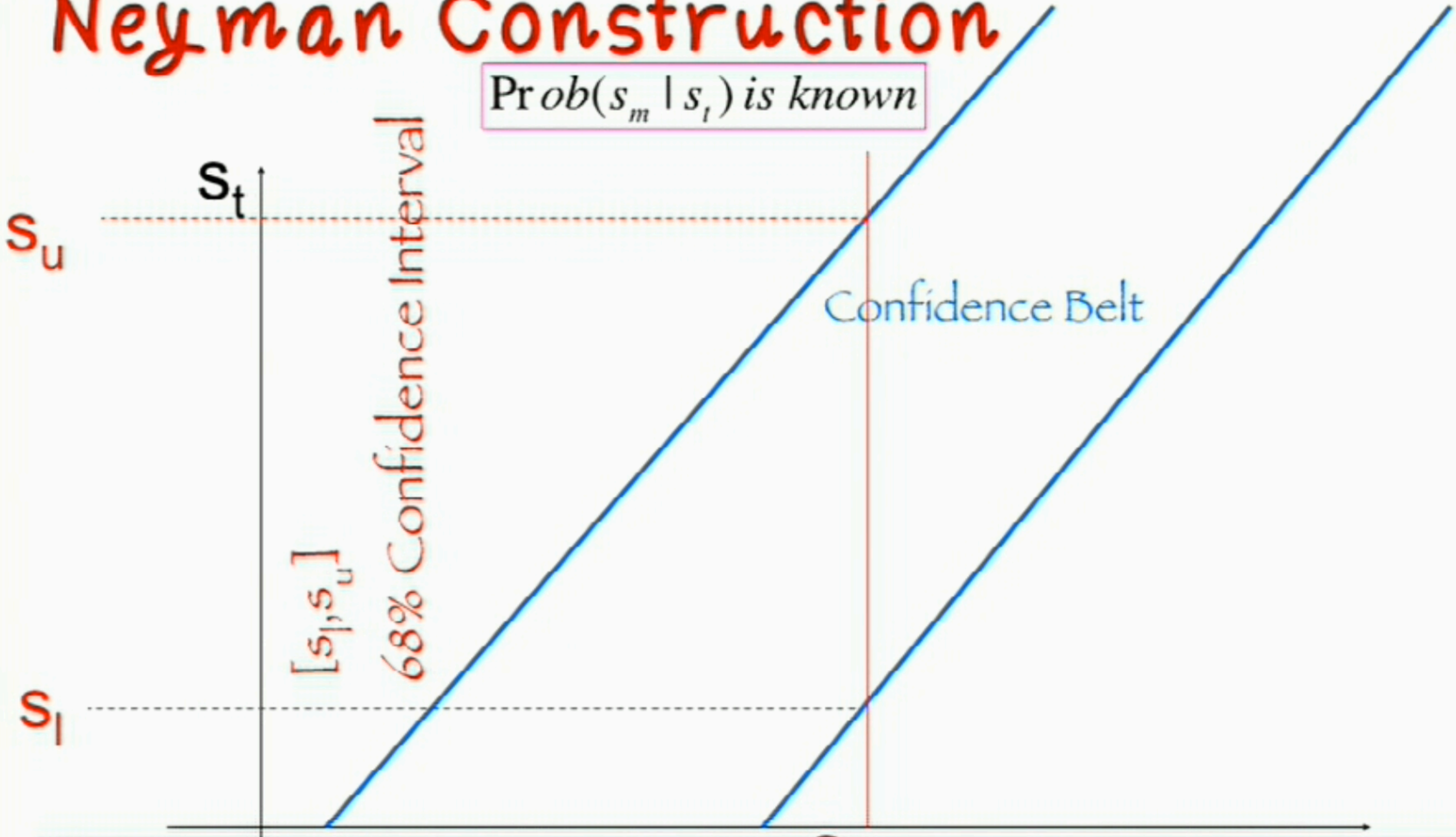
Neyman Construction

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



Neyman Construction

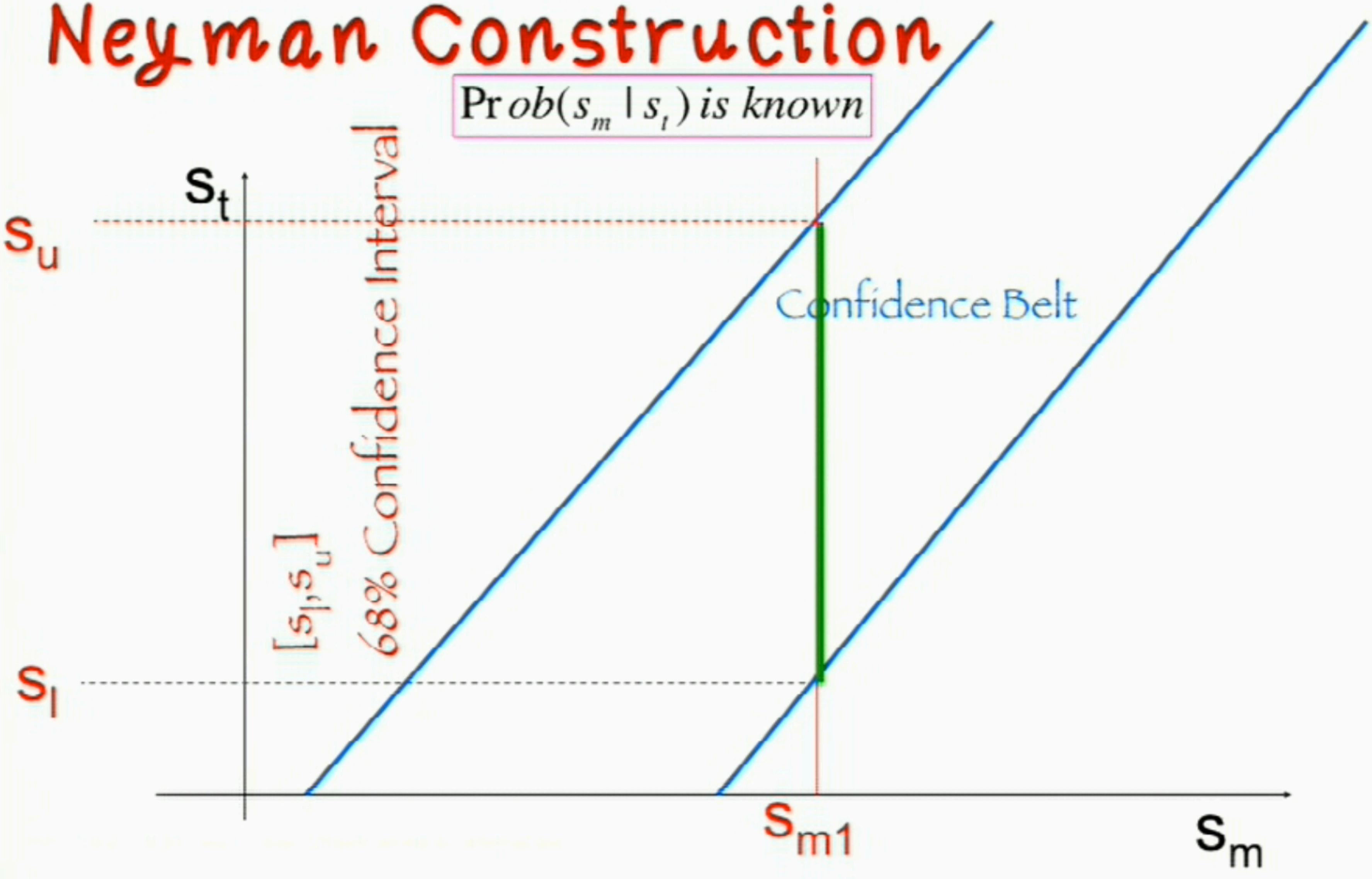
$\Pr ob(s_m \mid s_t)$ is known



$[s_l,s_u]$ 68% Confidence Interval

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

Neyman Construction



- 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

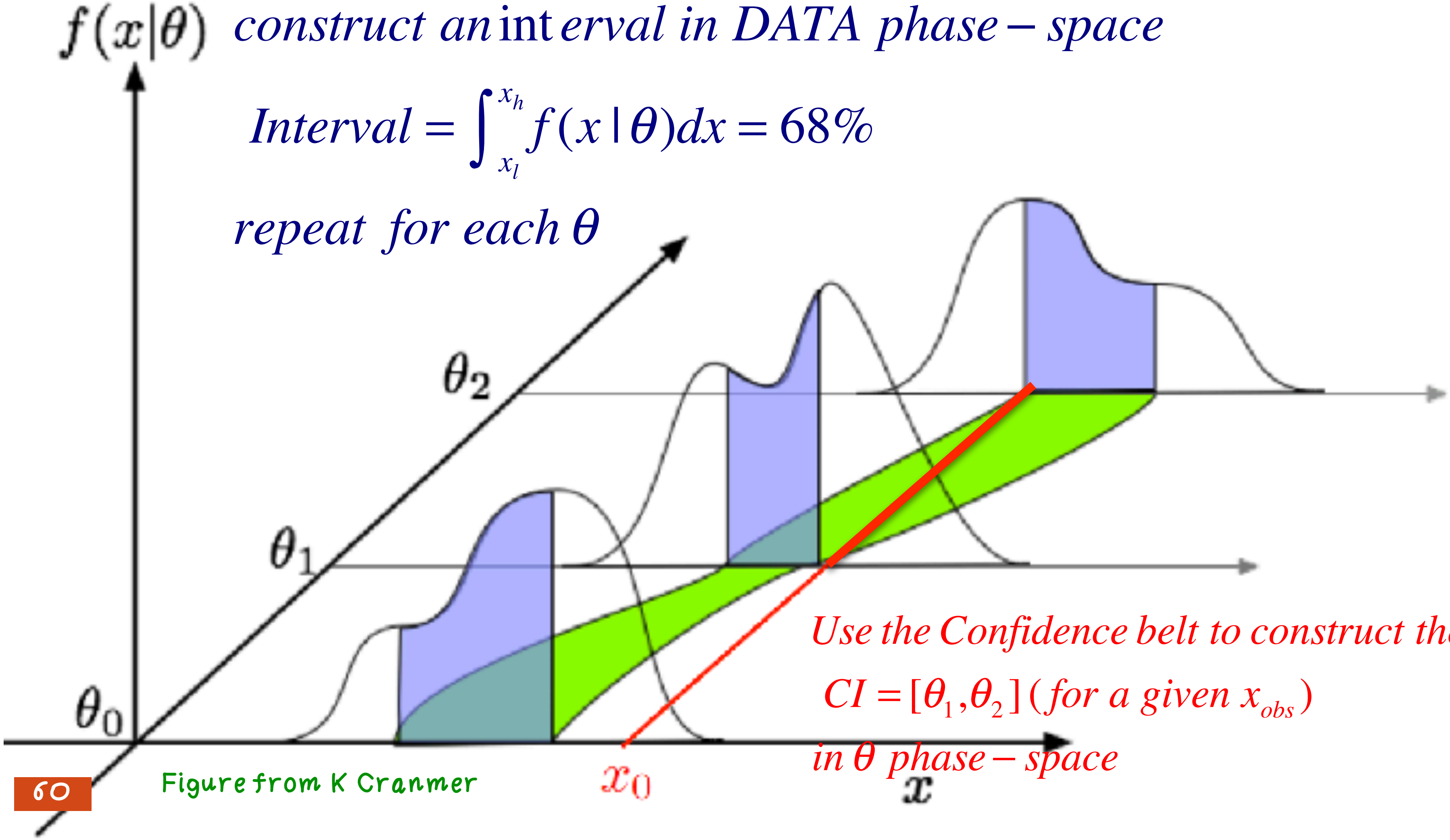
$\theta \equiv s_{true} \qquad x \equiv s_{measured} \qquad pdf \;\; f(x \mid \theta)$ *is known*

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