



Statistical analysis methods

Guillaume MENTION
CEA Paris-Saclay / Irfu

A word of caution



*Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in **basic probability** and mathematical **statistics**.*

Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid.

Larry Wasserman, Professor of Statistics and Data Science, Carnegie Mellon University,



Today's journey in "Stat-land"

1

Statistical
inference

2

Probability
reminders

3

Parameter estimation
LS & MLE

4

Goodness of fit
*How well model
follows data*

5

Hypothesis testing
*" H_0 or not H_0 " that is
the question*

6

Interval estimation

Why these methods

3 key questions in astroparticle physics

Question	Statistical tool
Discovery - Is there an excess above background?	Test statistic q_0 , p -value, significance Z
Measurement - What is the parameter value (flux, mass...)?	MLE + confidence interval
Exclusion - If no signal, what can we exclude?	Upper limit, CLs, Feldman-Cousins

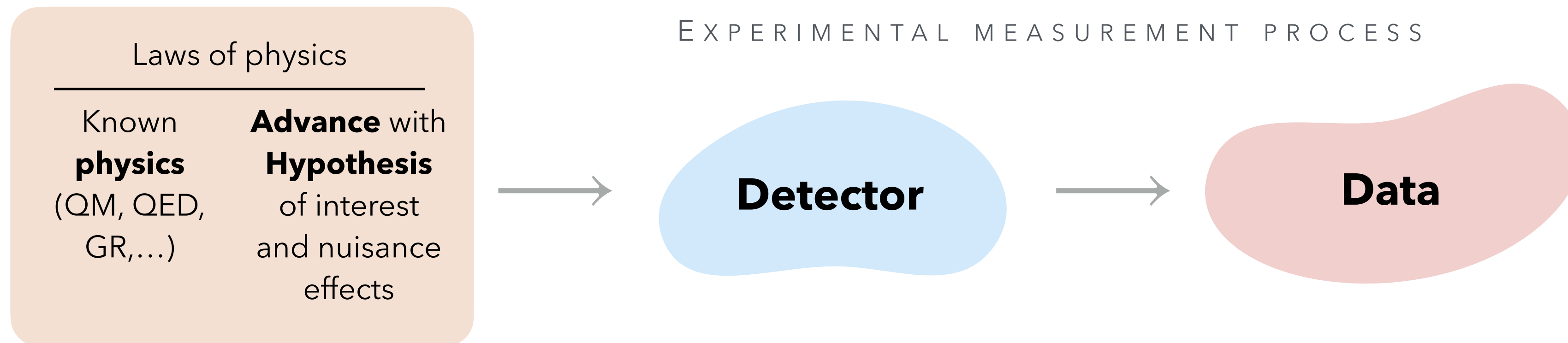
Everything in this course serves to answer these 3 fundamental questions.

Why probability and statistics in Physics

Randomness introduced in physics (quantum mechanics)
and also in detection process (random experimental errors)



Collected data are random

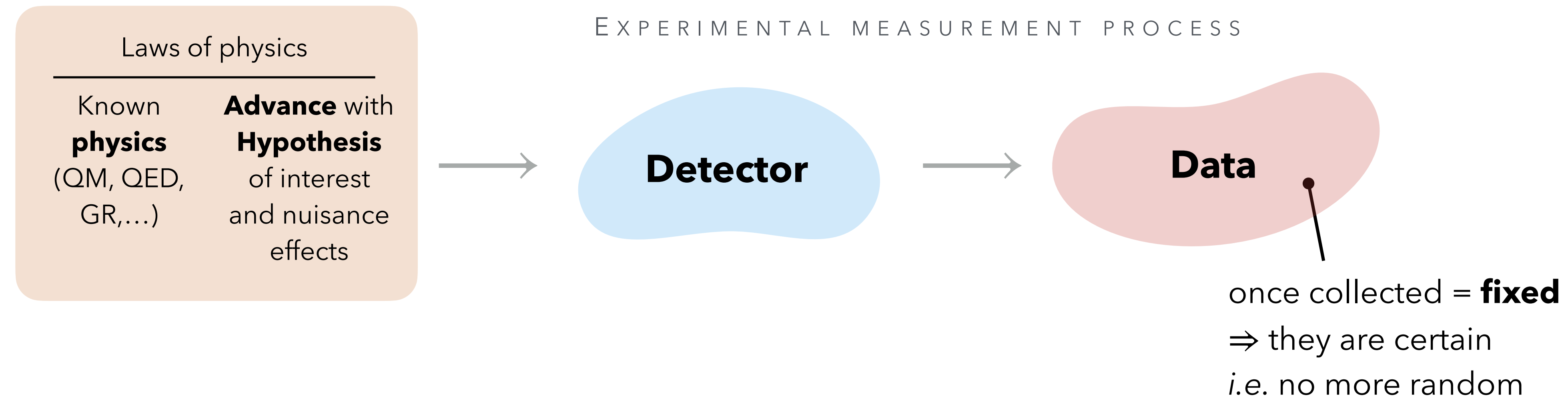


Random because: repeat the experiment in same conditions but get different outputs

Not possible to predict a given data set, as the collected experimental data (non-deterministic).

However, possible to attribute probability statements to the ensemble of possible data from given hypothesis and experimental conditions

Forward and Backward information flow



Forward process (hypothesis $\rightarrow$ data) makes possible estimation of $\mathbb{P}[\text{data} \mid \text{hypothesis}]$

Occurs in **real experiment** BUT the original conditions (hypotheses) are unknown to us...

Occurs in **simulations** WHERE the original conditions (hypotheses) are known to us, but not necessarily the ones of Nature.

Monte Carlo
simulations

Backward process (data $\rightarrow$ hypothesis) is called **Statistical Inference**

Statistical inference

There are **2** different **ways** of **inverting**
the **forward reasoning** to do statistics:

The **Bayesian** way and the **Frequentist** way

~ subjective

~ objective

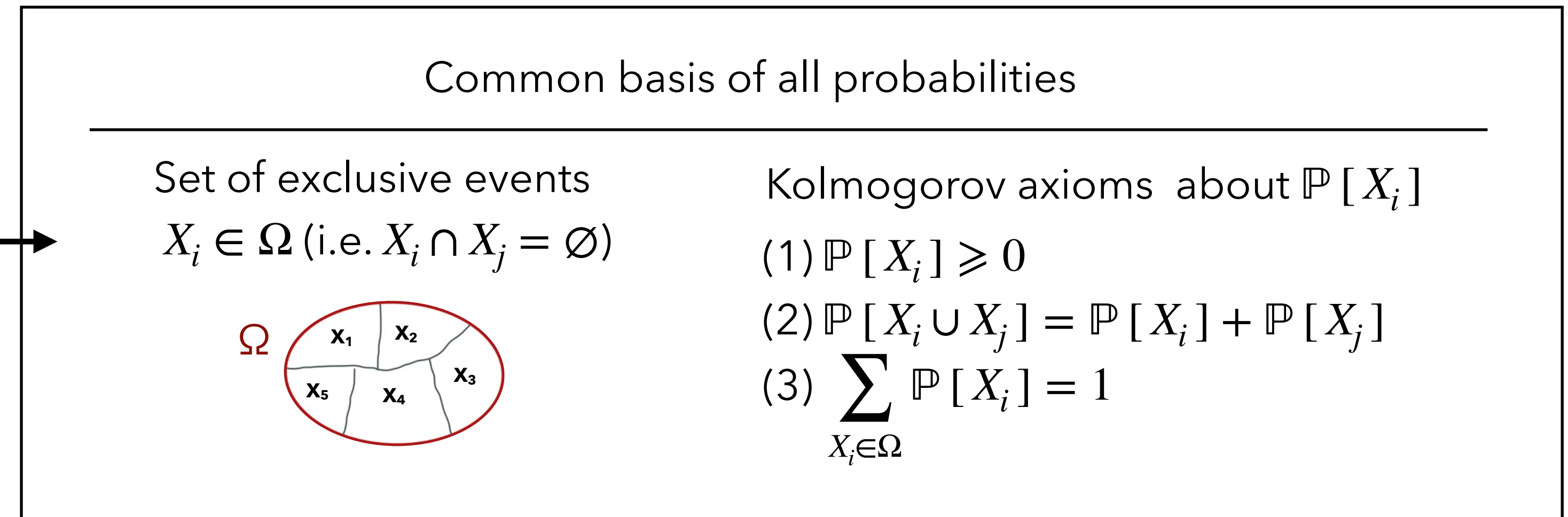
Reminders on probability

Probability

All statistical methods are based on **probability** computations.

Mathematical abstract statements

+ formal measure theory $\mu(dx)$, etc.



Two main classes of **interpretation** for experimental use:

Frequentist probability is defined as the limiting frequency of favourable outcomes in a large number of identical experiments

Needed: Reference ensemble – i.e. *what would have happened in other experiments*

Bayesian probability is defined as the degree of belief in a favourable outcome of a single experiment.

Needed: Priors – i.e. *what you believed before the experiment*

$$\mathbb{P}[A] = \lim_{N \rightarrow \infty} \frac{N(A)}{N}$$

success
 trials

$N \rightarrow \infty$ in what?

$\mathbb{P}[\text{Hyp} \mid \text{Data}]$

$\mathbb{P}[\text{Hyp}]$?

neither approach is assumption-free!

Some properties of any probability

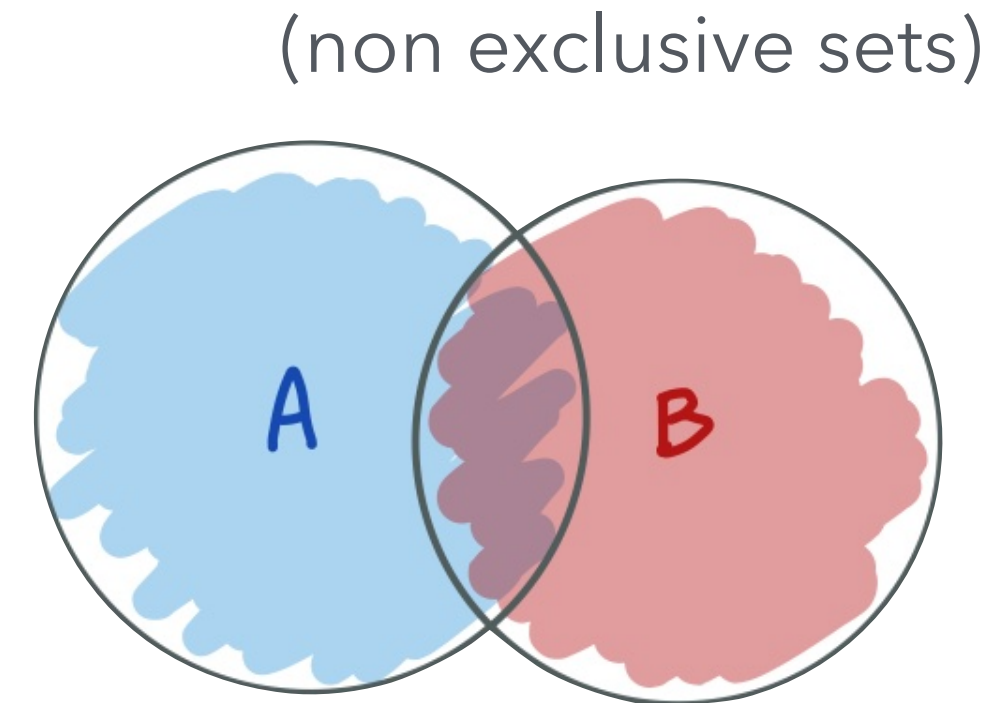
Properties derived from Kolmogorov axioms

$P(A \text{ or } B)$ means A or B or both

$P(A \text{ and } B)$ means both A and B

From the Venn diagram, we see that:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Conditional probability: $P(A|B)$ means the probability that A is true, given that B is true.

If A and B are **independent**, then $P(A|B) = P(A)$.

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[A, B]}{\mathbb{P}[B]}$$

“just renormalize probability
with the fact we observed B”

The conditioning is non symmetric: $P(A|B) \neq P(B|A)$

Example of conditional probability:

HB is a human being in a given population. Consider the 2 events A and B:

A: HB is pregnant.

B: HB is a woman.

Then: $P(A|B) \approx 1\%$ but $P(B|A) \approx 1$!

Bayes theorem

Bayes' Theorem says that the probability of both **A and B** being true simultaneously can be written:

$$\mathbb{P}[A, B] = \mathbb{P}[A | B] \mathbb{P}[B]$$

$$\mathbb{P}[A, B] = \mathbb{P}[B | A] \mathbb{P}[A]$$

here **A** and **B** are statements
i.e. either *True* or *False*

which can be written as:

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \mathbb{P}[A]}{\mathbb{P}[B]}$$

and $\mathbb{P}[B]$ can be expanded as: $\mathbb{P}[B] = \mathbb{P}[B | A] \mathbb{P}[A] + \mathbb{P}[B | \text{not } A] \mathbb{P}[\text{not } A]$
(law of total probability)

NOTE: for valid statement **A, B** this can be applied for **any probability definition**

Note on use of Bayes' theorem

Frequentist can use Bayes' theorem as soon as the statements A and B appearing in the probability are sound for a frequentist interpretation.

Question: "Do a hypothesis or a parameter have a long run frequency limit?"

Answer: NO!

So the Bayes formula can't be used to revert the probability statement such as $\mathbb{P}(\text{data} | \text{hypothesis})$ into $\mathbb{P}(\text{hypothesis} | \text{data})$ within the frequentist framework.

In **frequentist** interpretation: **hypothesis define the probability but is not a random variable**. Note: often in frequentist context, the **probability of data observation** is written

$$\mathbb{P}(\text{data} ; \text{hypothesis})$$

But $\mathbb{P}(\text{data} | \text{hypothesis})$ still tolerated

to **emphasise** the **hypothesis fixing** is **not** a **probability conditioning** as in the Bayesian case.

Within **frequentist** framework, **parameters, hypotheses** are **fixed**.

The trick to revert the probability statement and infer hypotheses or parameters from data is to use some "metric" to compare between different parameter/hypotheses. This trick is called the **Likelihood**.

Probability distributions

Probability laws for a random variable X

X discrete set of values: $\{x_1, \dots, x_N\}$
could be countable infinite set

Example:

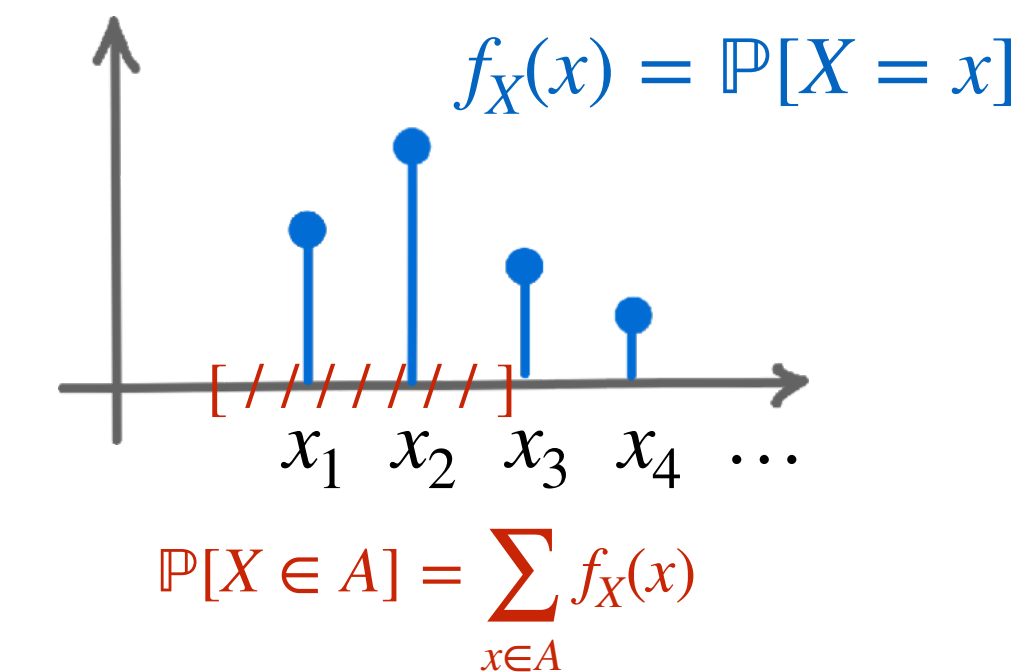
Bernoulli, Binomial, Poisson,...

X continuous range of values: $[x_{\min}; x_{\max}]$
could be infinite e.g. $\mathbb{R}$,...

Example:

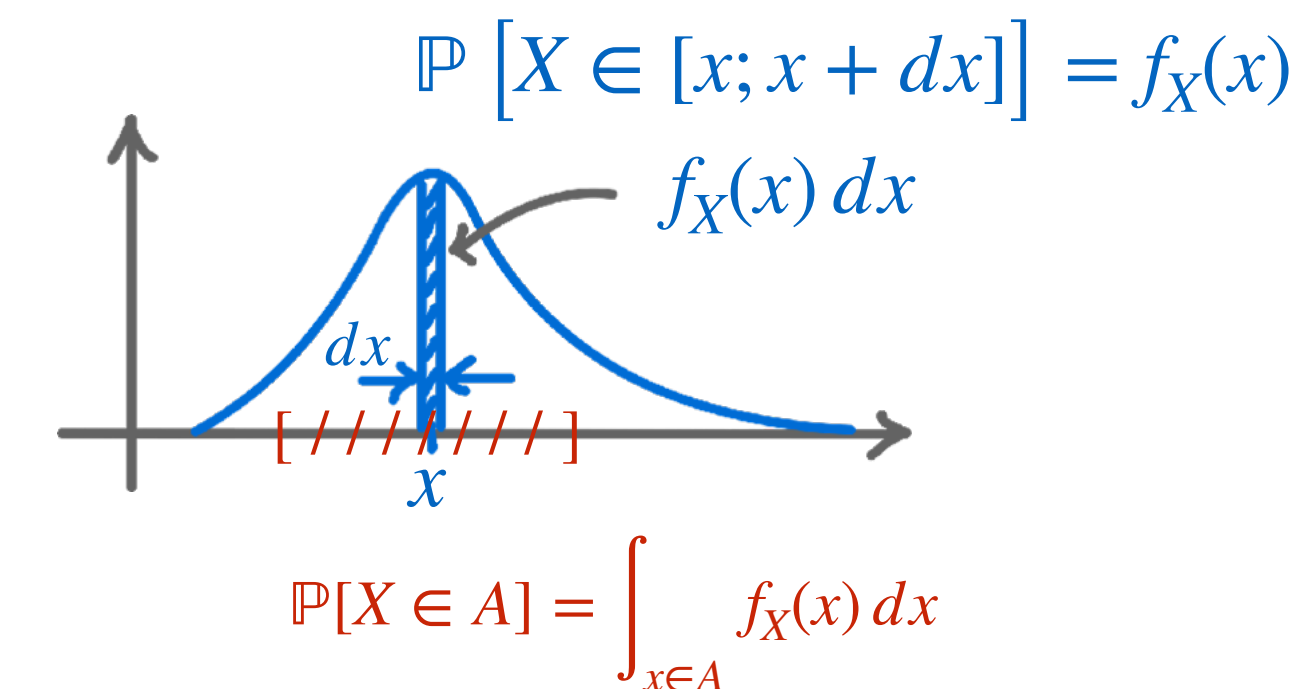
Normal ($\mathcal{N}$), Chi square (χ^2), Student (t),
Exponential, Gamma (Γ),...

Probability **Mass** Function (PMF)



Probability **Density** Function (PDF)

$\mathbb{P}[X = x] = 0 \dots !$



Properties of sampling distributions

Mean

"centrality"

$$\mathbb{E}[X] = \langle X \rangle = \bar{X} = \mu_X = \int x f_X(x) dx$$

Linearity property: $\mathbb{E}[X_1 + \dots + X_n] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]$ and $\mathbb{E}[cX] = c\mathbb{E}[X]$

If $X_1, \dots, X_n$ are independent, then $\mathbb{E}[X_1 \dots X_n] = \mathbb{E}[X_1] \dots \mathbb{E}[X_n]$, e.g. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Variance

"dispersion"

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma_X^2 = \int (x - \mu_X)^2 f_X(x) dx$$

If $X_1, \dots, X_n$ are independent $\mathbb{V}[X_1 + \dots + X_n] = \mathbb{V}[X_1] + \dots + \mathbb{V}[X_n]$

Covariance/Correlation

"relation, association"

covariance $\text{Cov}[X, Y] = \mathbb{V}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

correlation $\text{Cor}[X, Y] = \mathbb{C}[X, Y] = \rho_{X,Y} = \frac{\mathbb{V}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} \in [-1; 1]$

Covariance
matrix

(symmetric)

$$V = \begin{pmatrix} \mathbb{V}[X] & \mathbb{V}[X_1, X_2] & \dots & \mathbb{V}[X_1, X_n] \\ \mathbb{V}[X_2, X_1] & \mathbb{V}[X_2] & \dots & \mathbb{V}[X_2, X_n] \\ \vdots & \vdots & \dots & \vdots \\ \mathbb{V}[X_n, X_1] & \mathbb{V}[X_n, X_2] & \dots & \mathbb{V}[X_n] \end{pmatrix} \longrightarrow \text{in this case: } \mathbb{V}[X_1 + \dots + X_n] = \sum_{i,j=1}^n \mathbb{V}[X_i, X_j]$$

Other useful properties of sampling distributions

Median

$$x_{1/2} = Q(1/2) = F_X^{-1}(1/2)$$

Higher moments

$$\mu_n = \mathbb{E} [(X - \mathbb{E} [X])^n]$$

central moment

$$m_n = \mathbb{E} [X^n]$$

raw moment

Moment generating function

$$M_X(t) = \mathbb{E} [e^{tX}] = \sum_{k=0}^{\infty} \mathbb{E} [X^k] \frac{t^k}{k!}$$

$$\left. \frac{\partial^n M_{X-\mu_X}(t)}{\partial t^n} \right|_{t=0} = \mathbb{E} [(X - \mu_X)^n]$$

$$M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t)$$

Cumulant generating function

$$K_X(t) = \ln M_X(t)$$

$$K_{X_1+X_2}(t) = K_{X_1}(t) + K_{X_2}(t)$$

Characteristic function

closely related to Fourier
transform of $f_X(x)$

$$\varphi_X(t) = \mathbb{E} [e^{itX}] = \int_{\mathbb{R}} e^{itx} dF_X(x) = \int e^{itx} f_X(x) dx = \int_0^1 e^{it Q_X(p)} dp$$

$$M_X(t) = \varphi_X(-it)$$

used a lot in statistical analyses

mostly used in mathematical
probability computations

Each one of this quantity fully specify the distribution

f_X

F_X

Q_X

M_X

K_X

φ_X

Histograms

Histograms

- > Representation of the frequencies of the numerical outcome of a random phenomenon

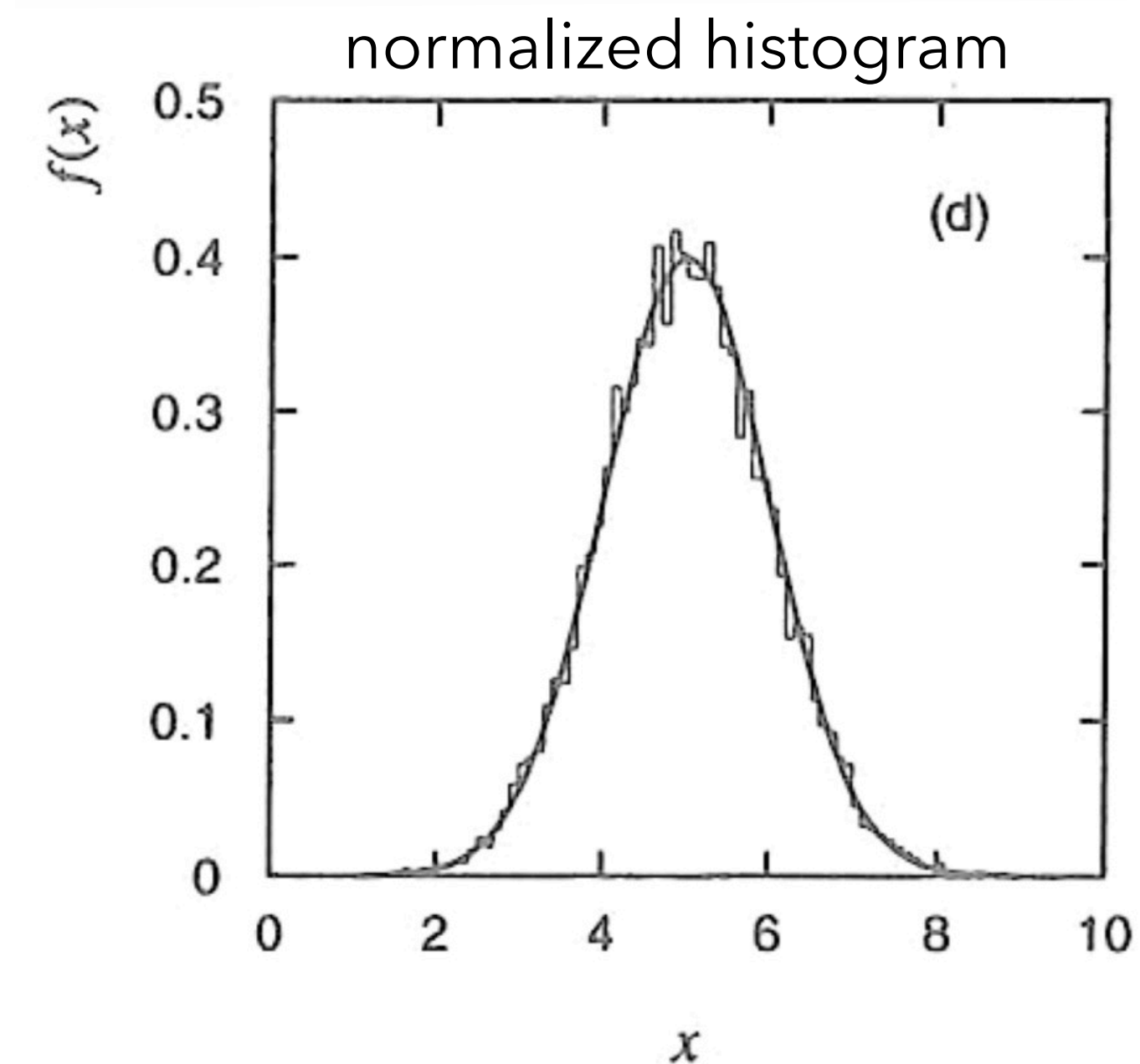
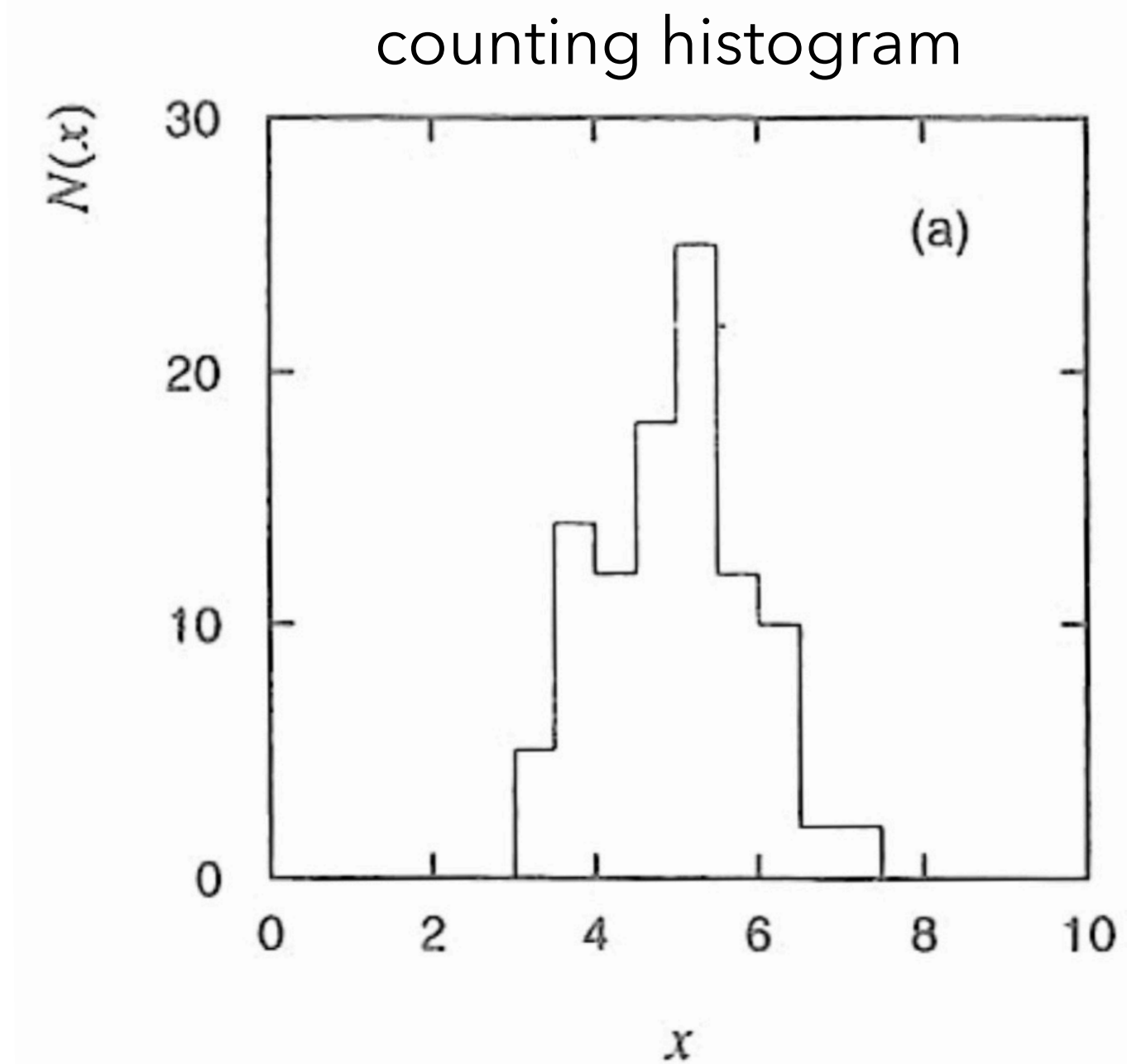
PDF = histogram for

- > infinite data sample
- > zero bin width
- > normalized to unit area

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

n = total number of entries
in the histogram

Δx = bin width



Sampling from an histogram or a PDF

Inverse transform sampling

$$\mathbb{P}_X[X \leq x] = F_X(x)$$

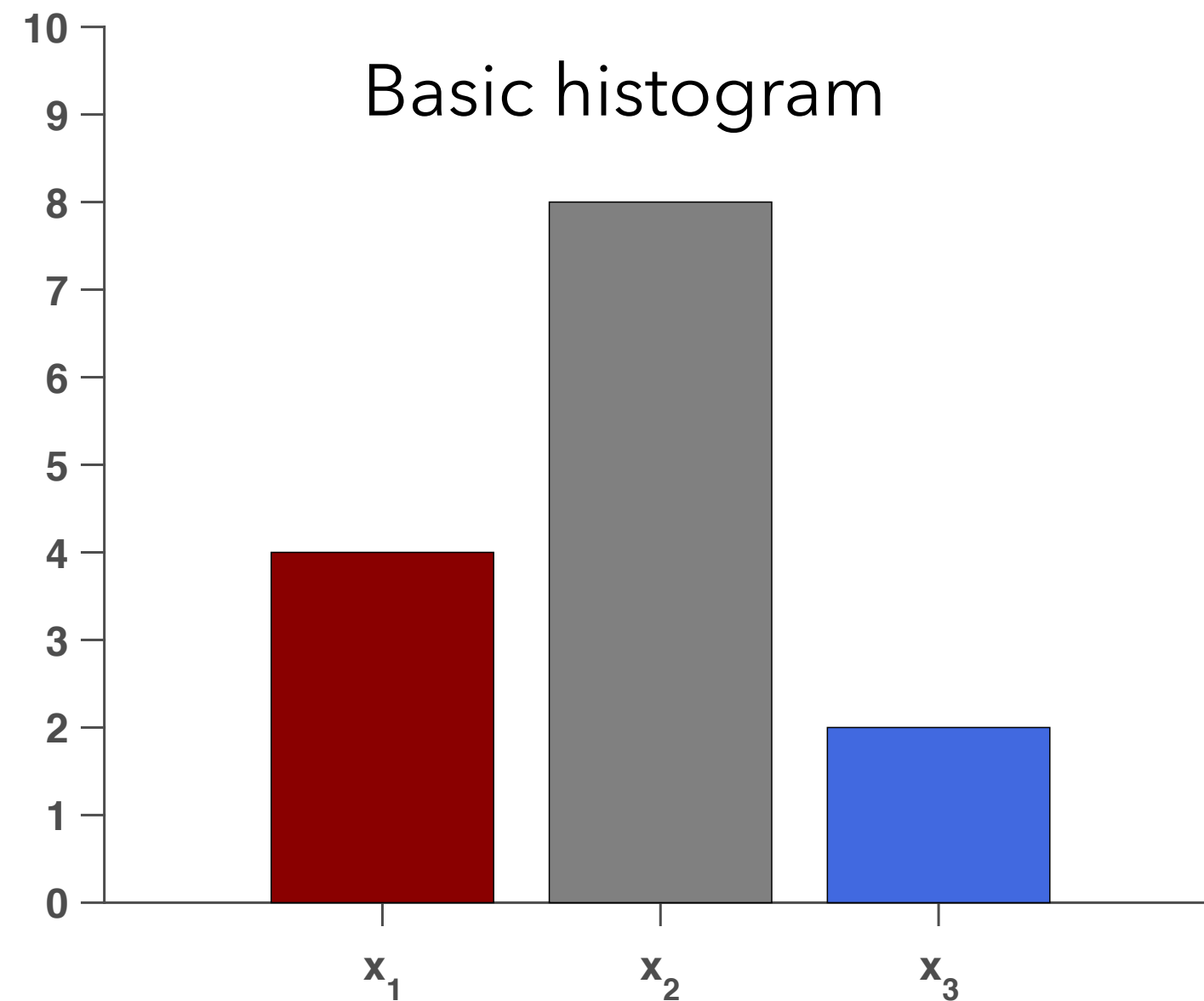
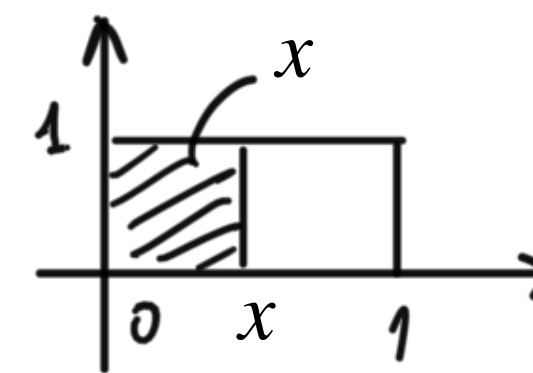
$$\mathbb{P}_X[F_X(X) \leq x] = \mathbb{P}_X[X \leq F_X^{-1}(x)] = F_X(F_X^{-1}(x)) = x$$

$\Rightarrow F_X(X)$ is a uniform random variable U on $[0; 1]$

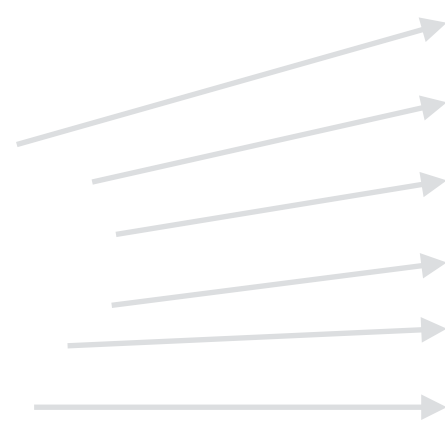
$$F_X(X) \sim U \Rightarrow X \sim F_X^{-1}(U)$$

$\Rightarrow F_X^{-1}(U)$ follows the probability distribution of X

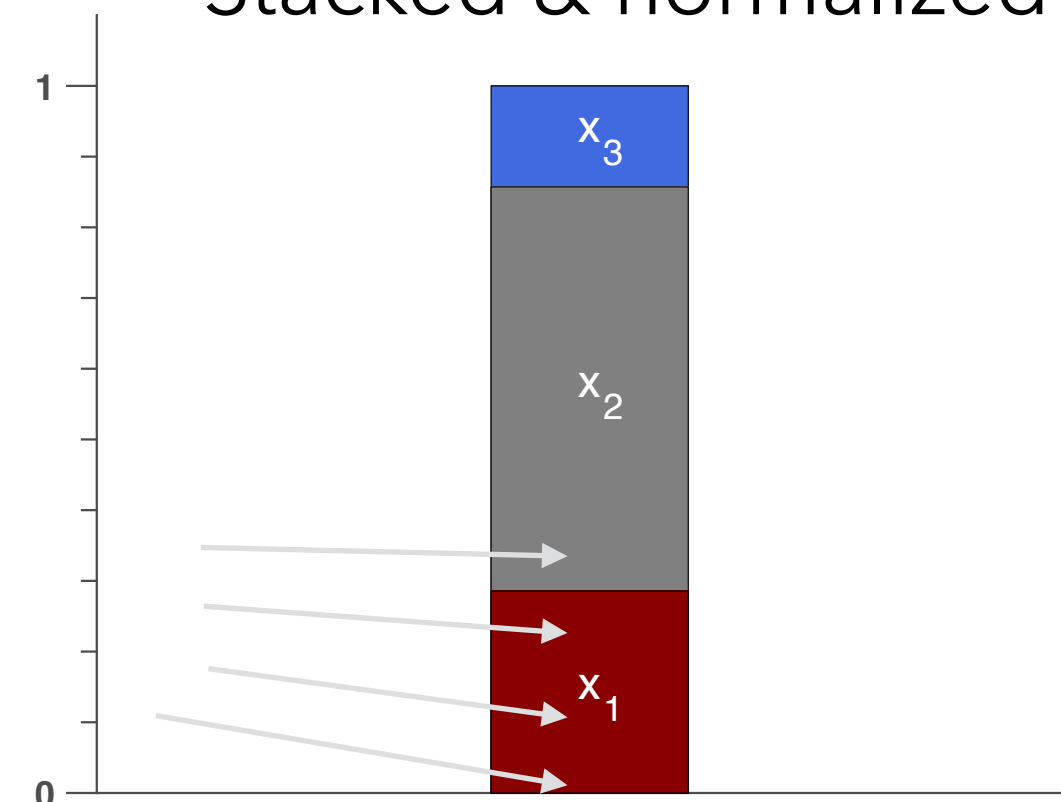
$$F_U(x) = x$$



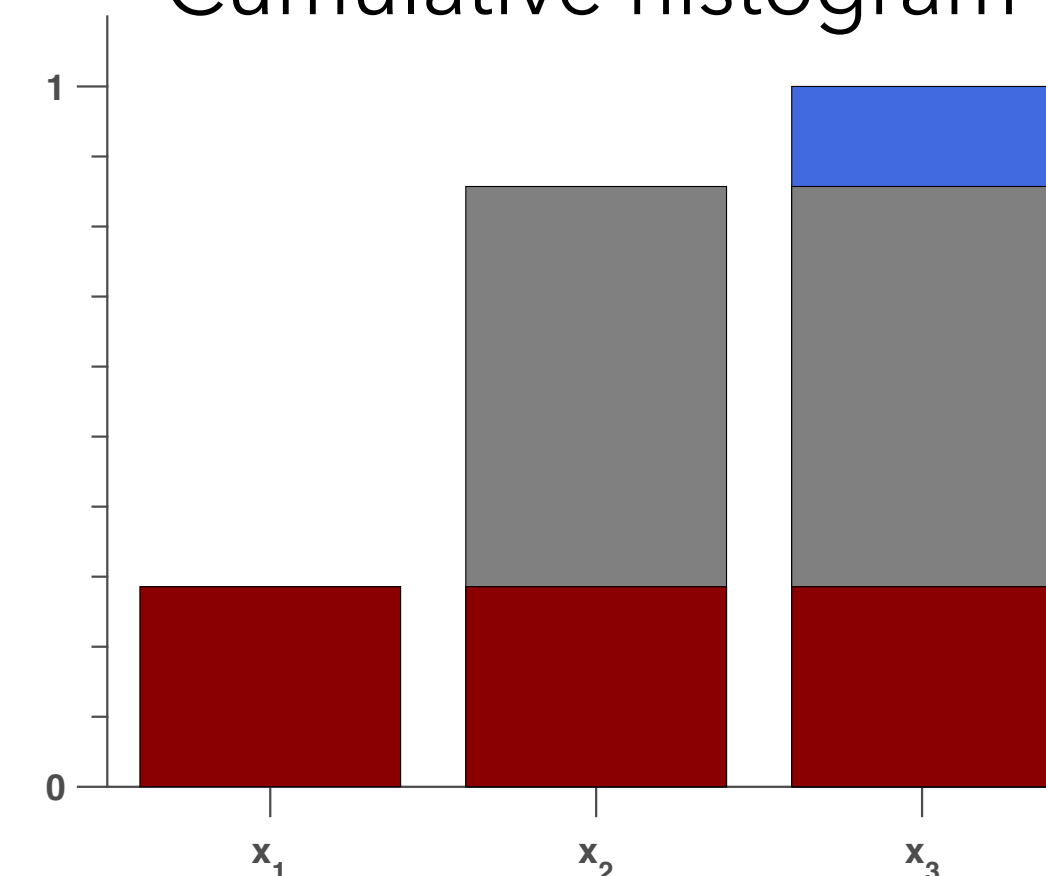
Random number generator
between 0 and 1



Stacked & normalized



Cumulative histogram

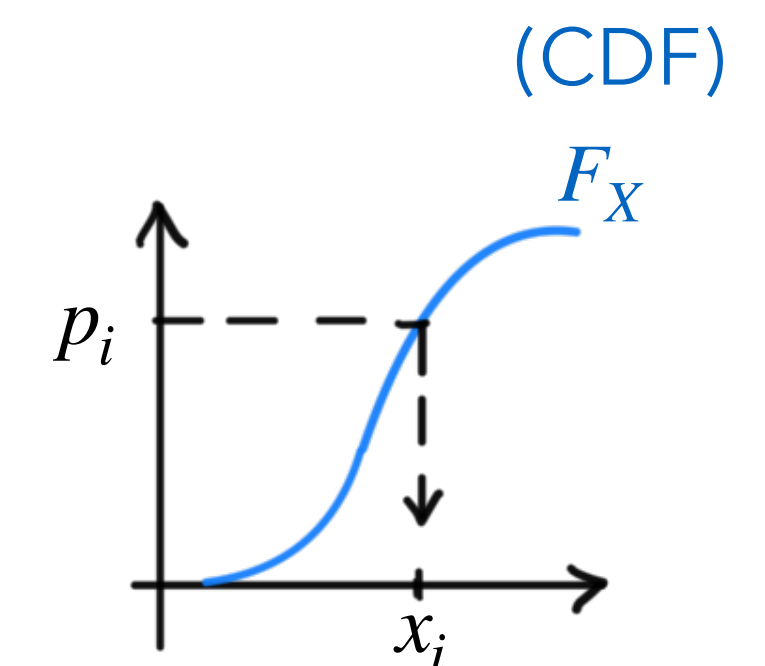


Continuous PDF case:

Inverse transform sampling

Take randomly $p \in [0; 1]$

$F_X^{-1}(p) = \text{interpolate } (F_X(x_i), x_i) \text{ at } p$



Conditional probabilities

"Conditioning is the soul of statistics" – J. Blitzstein (Harvard prof. of stats.)

Joint probability $\mathbb{P}[X, Y]$

Joint density $f_{X,Y}(x, y)$

Marginal probability

$$\mathbb{P}[X] = \sum_y \mathbb{P}[X, Y = y]$$

$$f_X[X] = \int_y f_{X,Y}(X, Y = y) dy$$

Conditional random variables

discrete

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(X = x, Y = y)}{P_X(x)}$$

continuous

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Conditional expectation

$\mathbb{E}[X | Y = y]$ function notation " $f(y)$ "

$\mathbb{E}[X | Y]$ random variable " $f(Y)$ "

$$\mathbb{E}[X | Y] = \sum_x x \mathbb{P}[X = x | Y]$$

$$\mathbb{E}[X | Y] = \int_x x f_X(X = x | Y) dx$$

Conditional variance

$$\mathbb{V}[X | Y] = \mathbb{E} \left[(X - \mathbb{E}[X | Y])^2 | Y \right]$$

Conditional expectation and variance

Law of total expectation $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

Law of total variance $\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$

mnemonic trick: "**Eve's** law or **EVVE's** law"

Typical example $S_N = \sum_{i=1}^N X_i$

N is **fixed**: Sum of a fixed number of X_i iid. random variables $\mathbb{E}[S_N] = N \mathbb{E}[X]$ and $\mathbb{V}[S_N] = N \mathbb{V}[X]$

N is **random**: Sum of a random number of X_i iid. random variables

$$S_N = \sum_{i=1}^N X_i \quad \mathbb{E}[S_N] = \mathbb{E}\left[\sum_{i=1}^N X_i\right] \neq \sum_{i=1}^N \mathbb{E}[X_i] \quad \text{NO!!!} \quad N \text{ is random}$$

$$\mathbb{E}[S_N] = \mathbb{E}\left[\sum_{i=1}^N X_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^N X_i \mid N\right]\right] = \mathbb{E}[N \mathbb{E}[X|N]] = \mathbb{E}[N \mathbb{E}[X]] = \mathbb{E}[N] \cdot \mathbb{E}[X]$$

$$\begin{aligned} \mathbb{V}[S_N] &= \mathbb{E}[N \mathbb{V}[X]] + \mathbb{V}[N \mathbb{E}[X]] \\ &= \mathbb{E}[N] \times \mathbb{V}[X] + \mathbb{V}[N] \times (\mathbb{E}[X])^2 \end{aligned}$$

⇒ Every time random variables are compounded like with S_N example, conditioning is very handy tool

Another example of conditional variance

Energy resolution $\frac{\sigma_E}{E} = \sqrt{a^2 + \frac{b^2}{E} + \frac{c^2}{E^2}}$ (for instance PMTs Energy resolution)

Reconstructed Energy $E_{\text{rec}} = g_0 \cdot \alpha \cdot N(E) + \varepsilon$ Calibration model = linear energy scale + detection noise

Conversion factor g_0 is a calibrated constant (e.g. "gain" for PMTs)
average number of ADC counts (or charge) produced per photoelectron

Energy scale α is random with $\mathbb{E}[\alpha] = 1$ and $\mathbb{V}[\alpha] = a^2$
 α for dynode multiplication fluctuations, PMT gain non-uniformity, quantum efficiency variations

Detector noise β is random with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{V}[\varepsilon] = c^2$
 ε for electronic noise, dark current, baseline

α , ε and $N(E)$ are assumed independent

Stochastic counting $N(E)$ is a random variable following a Poisson distribution with parameter $\mu(E)$
 $N(E)$ for photoelectron statistics
 with $\mathbb{E}[N] = \mu$ and $\mathbb{V}[N] = \mu$ for all energies E

$$\mathbb{V}[E_{\text{rec}}] = \mathbb{V}[\alpha \cdot g_0 \cdot N + \varepsilon]$$

$$= \mathbb{E}[\mathbb{V}[\alpha \cdot g_0 \cdot N | \alpha] + c^2] + \mathbb{V}[\mathbb{E}[\alpha \cdot g_0 \cdot N | \alpha]]$$

$$= (1 + a^2) \cdot g_0^2 \cdot \mu + c^2 + a^2 \cdot g_0^2 \cdot \mu^2$$

$$\text{" } \frac{\sigma_E}{E} \text{"} = \frac{\sqrt{\mathbb{V}[E_{\text{rec}}]}}{\mathbb{E}[E_{\text{rec}}]}$$

assuming $a^2 \ll 1$

$$= \sqrt{\frac{a^2 \cdot \langle E_{\text{rec}} \rangle^2 + (1 + a^2)g_0 \cdot \langle E_{\text{rec}} \rangle + c^2}{\langle E_{\text{rec}} \rangle^2}} \sim \sqrt{a^2 + \frac{b^2}{E} + \frac{c^2}{E^2}}$$

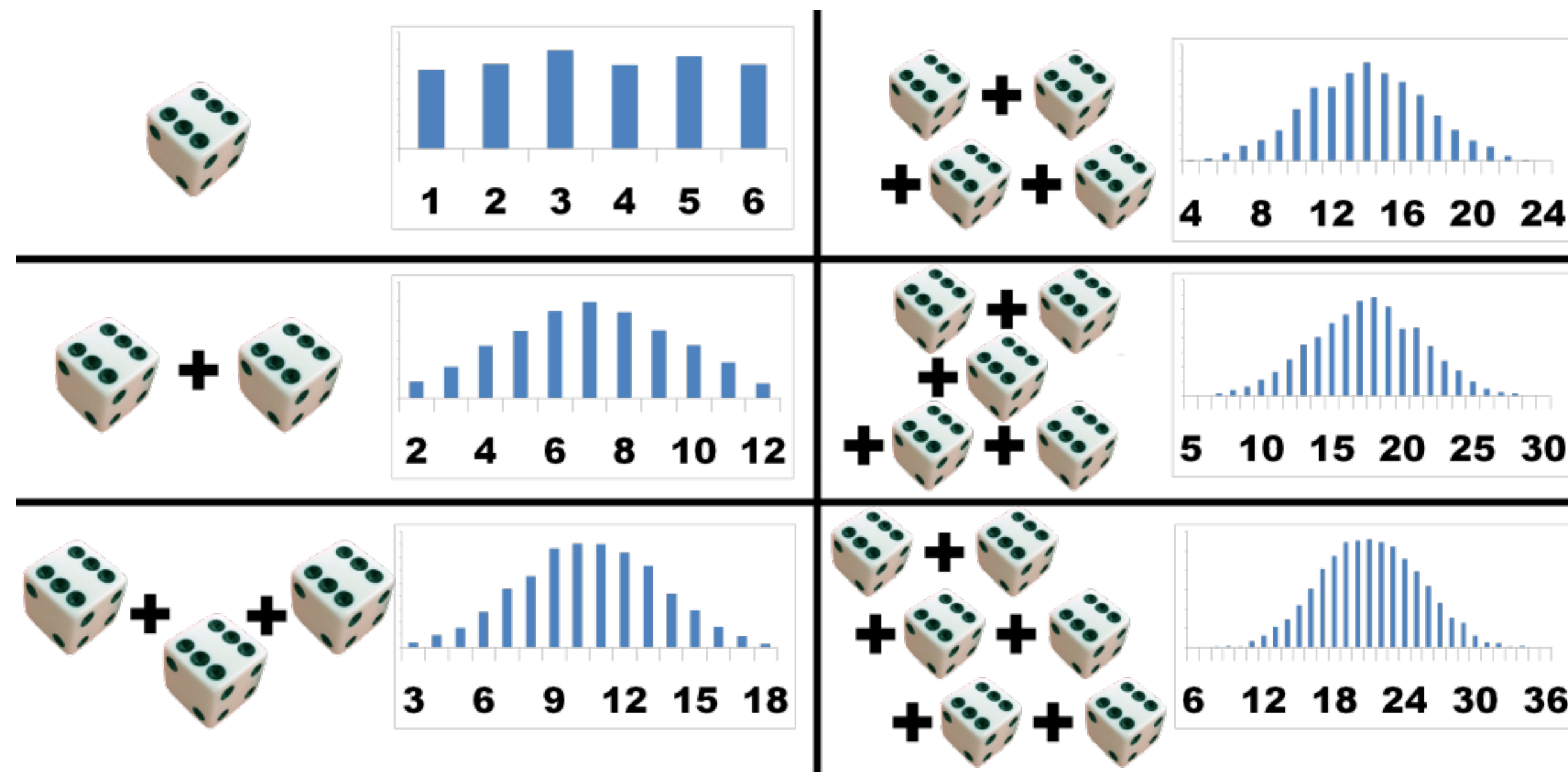
$$\mathbb{E}[E_{\text{rec}}] = \mathbb{E}[\alpha \cdot g_0 \cdot N + \varepsilon] = \mathbb{E}[\mathbb{E}[\alpha \cdot g_0 \cdot N | \alpha]] = g_0 \cdot \mu$$

Central role of Normal distribution

The Central Limit Theorem (CLT)

Suppose $X_1, X_2, \dots$ are iid random variables such that $E(X_n) = \mu, \text{Var}(X_n) = \sigma^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i$, and let N be a normally distributed random variable, with $E(N) = 0, \text{Var}(N) = 1$. Then,

$$\frac{\frac{S_n}{n} - E\left(\frac{S_n}{n}\right)}{\sqrt{\text{Var}\left(\frac{S_n}{n}\right)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \Rightarrow N$$



Note 1: the mathematical proof uses the characteristic function function $\varphi_X(t) = \mathbb{E}[e^{itX}]$ $\varphi_{S_N}(t) = \prod_{i=1}^N \varphi_X(t)$

Note 2: Finite variance = important! Otherwise if variance not finite => Look at Lévy stable distributions (heavily used in Finance)

Normal distribution

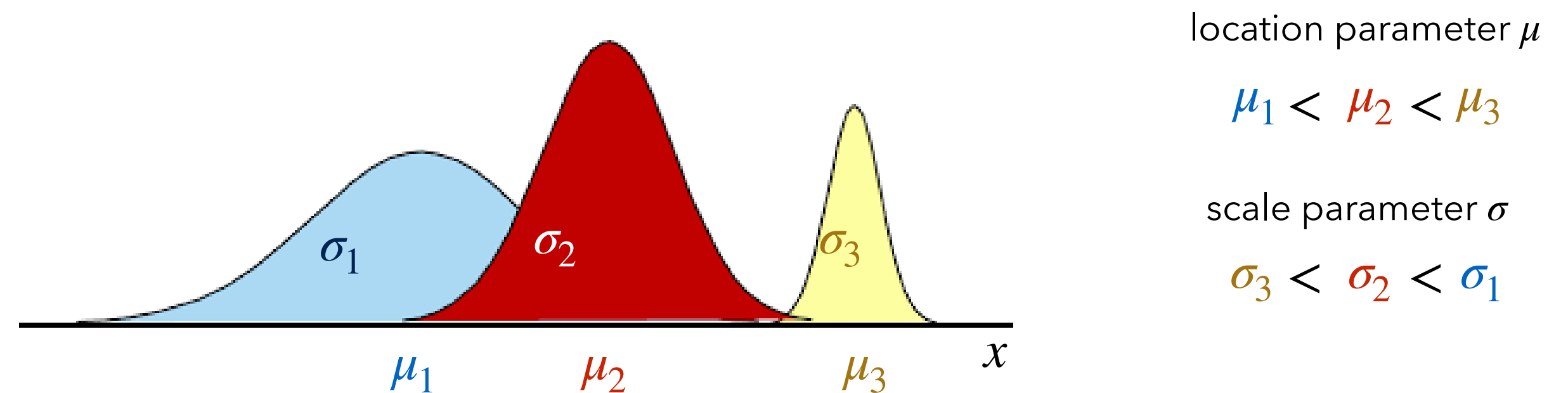
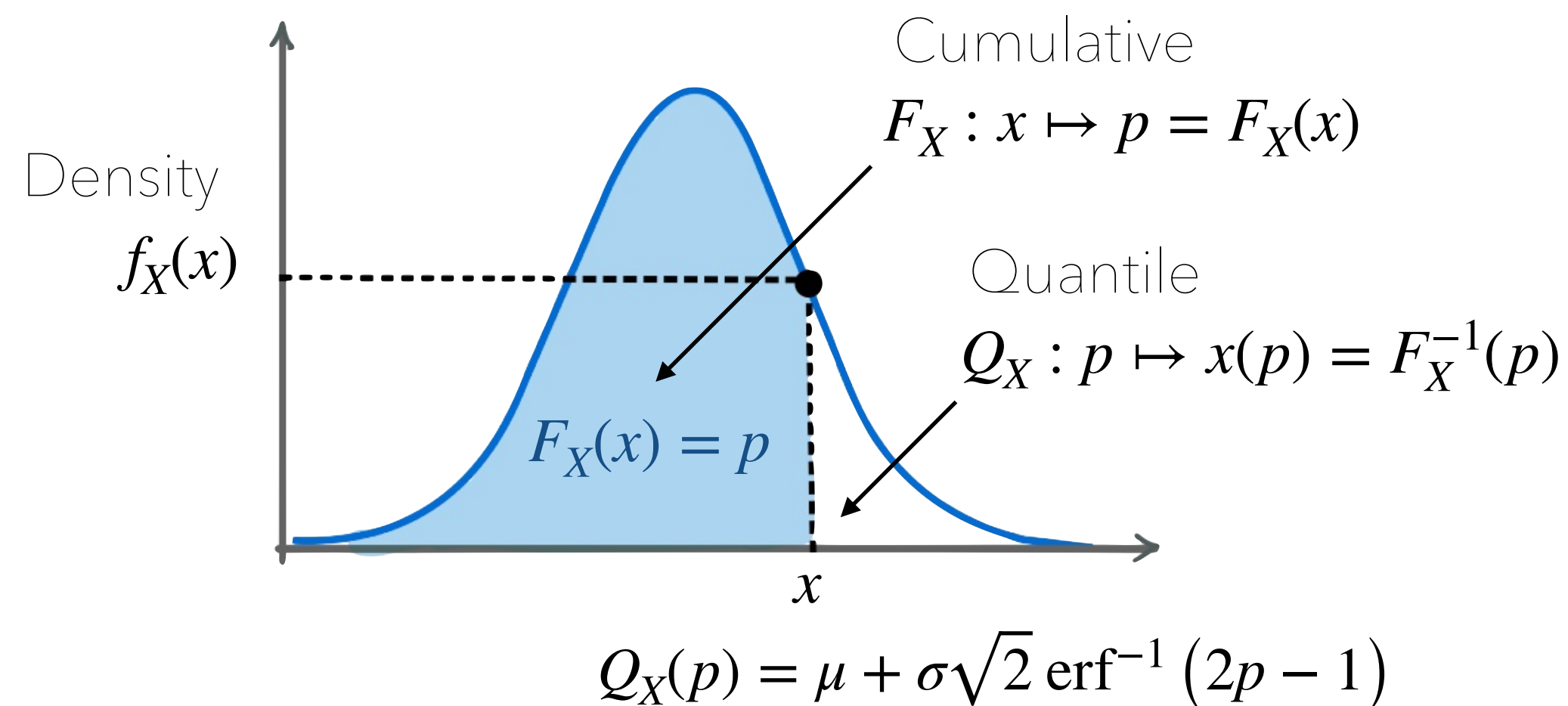
Probability Density Function

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Standard normal distribution $\phi(x) = f_X(x; \mu = 0, \sigma = 1)$

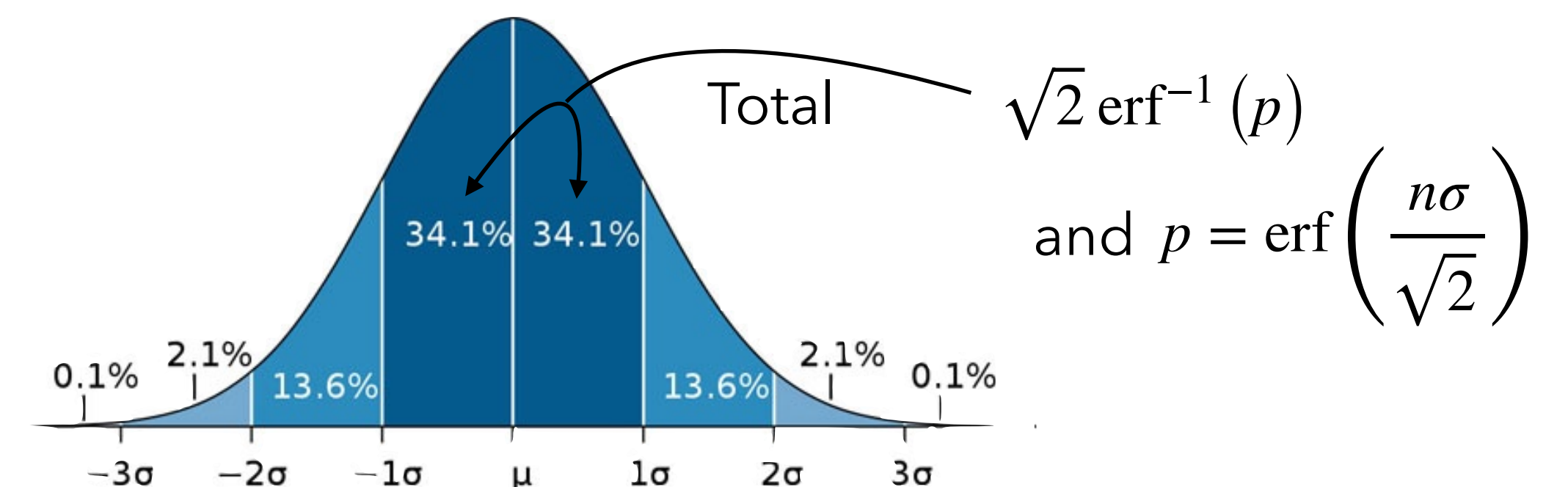
Cumulative Distribution Function

$$F_X(x; \mu, \sigma) = \mathbb{P}[X \leq x] = \int_{-\infty}^x f_X(u) du = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) \right)$$



Standard normal cumulative distribution $\Phi(x) = F_X(x; \mu = 0, \sigma = 1)$

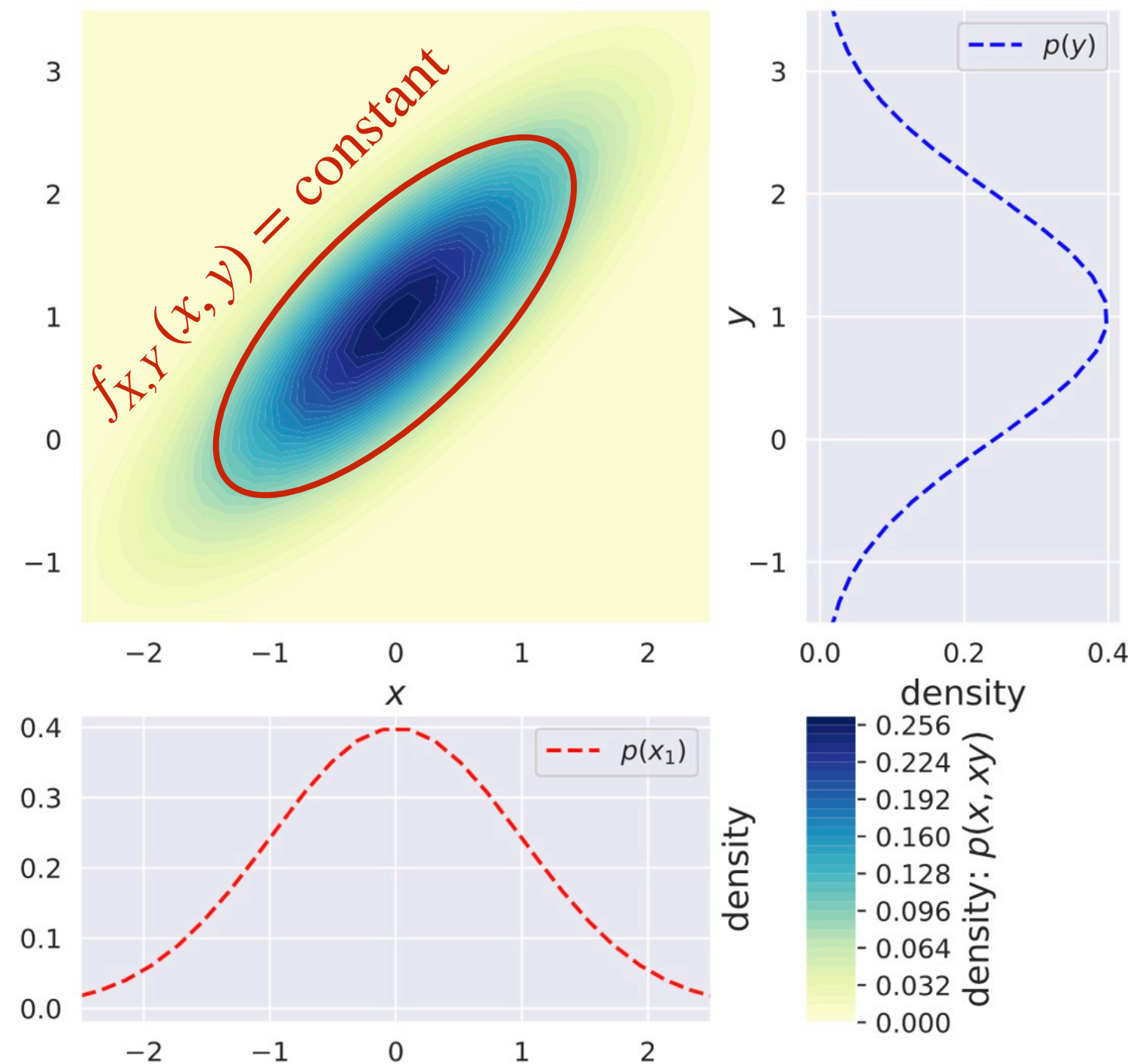
Standard normal interval: $1\sigma, 2\sigma, 3\sigma$
 $\mu = 0$ and $\sigma = 1$



Multivariate normal distribution

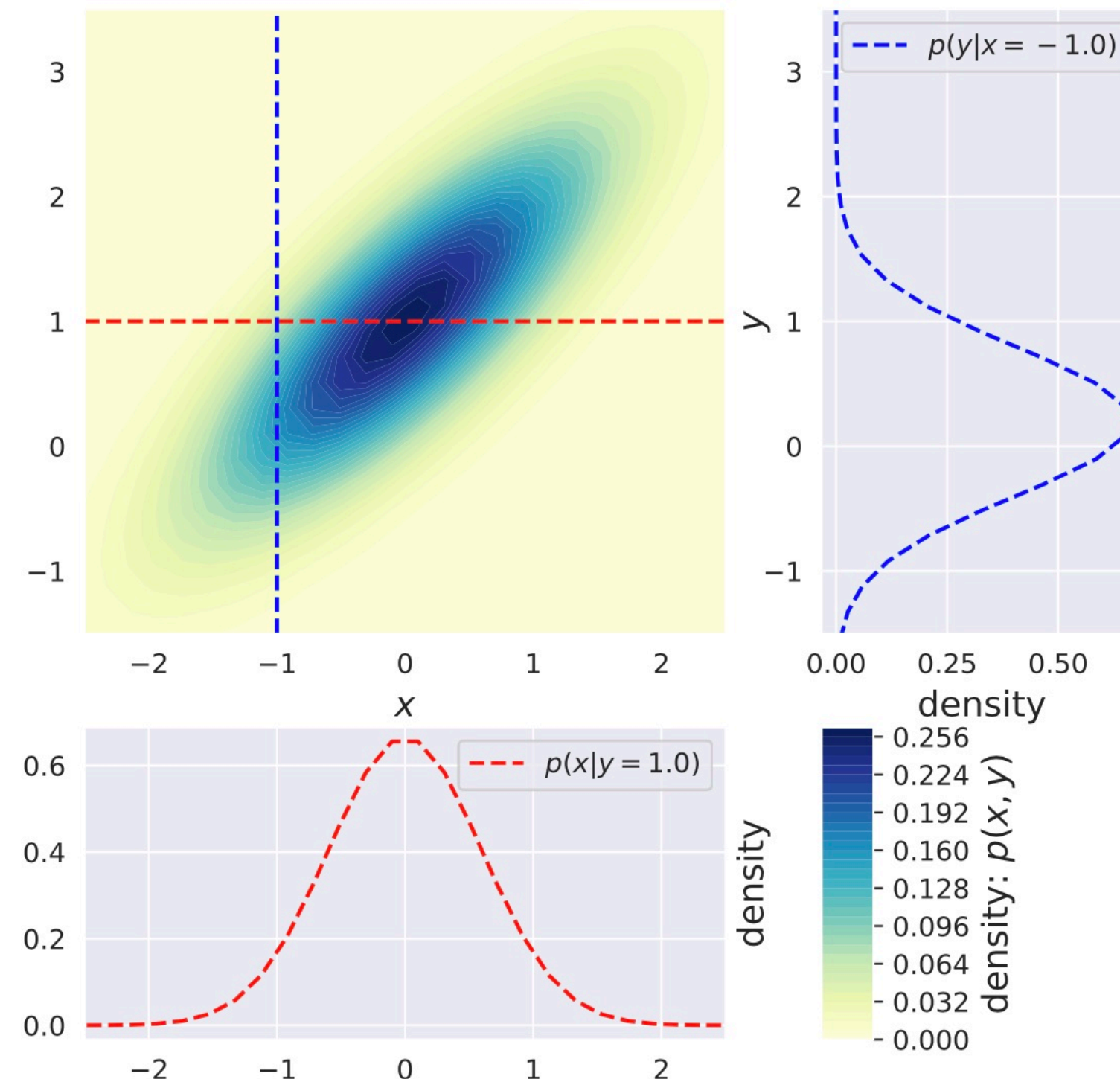
$$f_{\mathbf{X}}(x_1, \dots, x_k) = |2\pi V|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T V^{-1}(\mathbf{x}-\boldsymbol{\mu})} = \frac{1}{\sqrt{(2\pi)^n \det(V)}} \exp \left(-\frac{1}{2} \sum_{i,j=1}^n V_{i,j}^{-1} (x_i - \mu_i)(x_j - \mu_j) \right)$$

Marginal distributions



$$f_X(x) = \frac{1}{\sqrt{2\pi V_{X,X}}} \exp \left(-\frac{(x - \mu_X)^2}{2V_{X,X}} \right)$$

Conditional distributions



Important properties:

- Linear combinations of multivariate normals are normally distributed
- Marginal 1D distribution are also normally distributed
- Conditional distributions are normally distributed

in 2 dimensions

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad V = \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}$$

$$X | Y = y \sim \mathcal{N} \left(\mu_X + \frac{\sigma_X}{\sigma_Y} \rho (y - \mu_Y), (1 - \rho^2) \sigma_X^2 \right)$$

in n dimensions

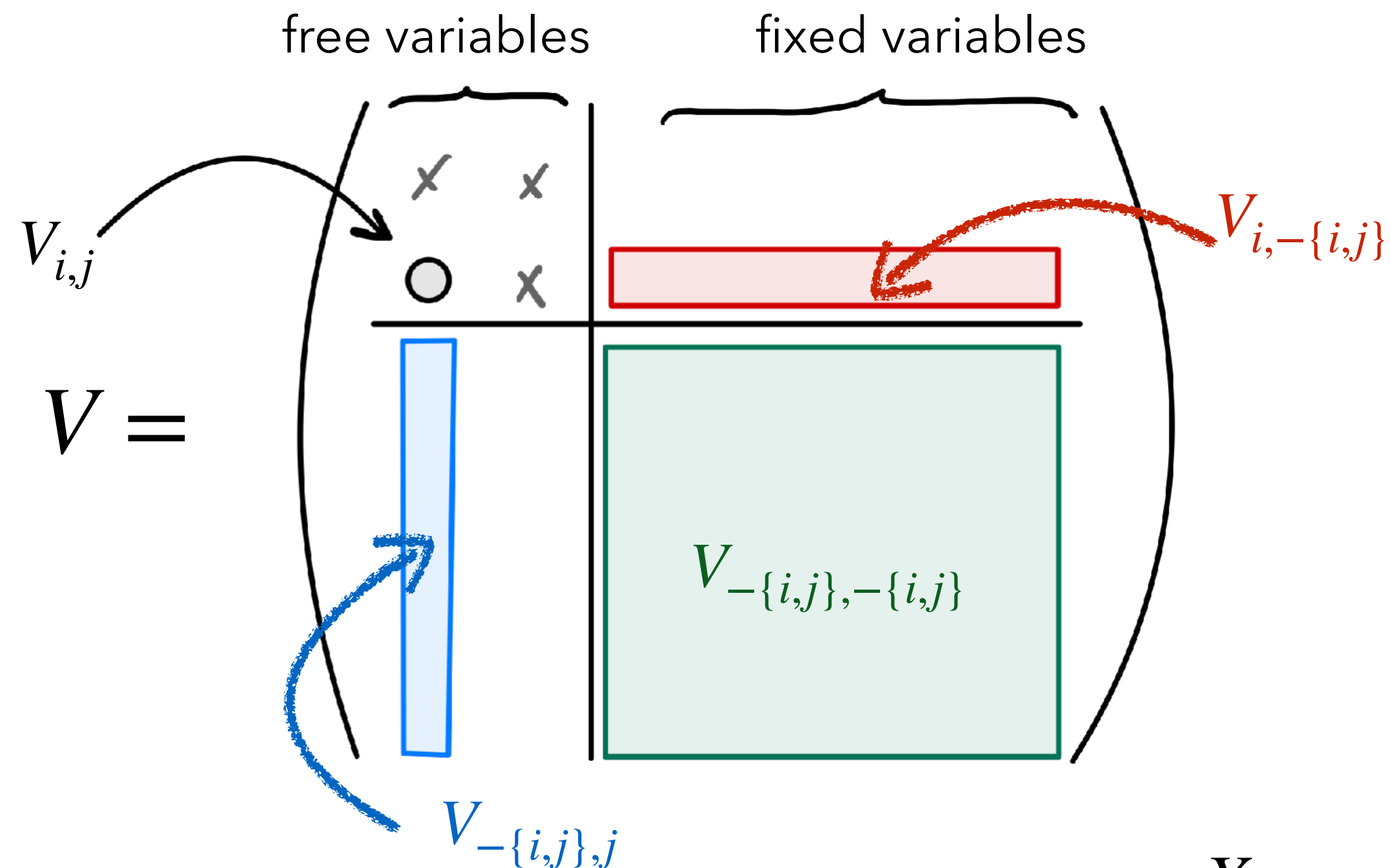
$$\mu_{X|Y=y} = \mu_X + V_{X,Y} V_Y^{-1} (y - \mu_Y)$$

$$V_{X|Y} = V_X - V_{X,Y} V_Y^{-1} V_{Y,X}$$

$$X | Y = y \sim \mathcal{N} (\mu_{X|Y}, V_{X|Y})$$

Useful trick: conditional covariance

Each element $V_{i,j}$ of covariance matrix V contains the TOTAL covariance, i.e. the direct covariance between i and j but also the ones indirectly induced by all the other variables... !



For instance, i and j could be directly uncorrelated but still have a non-zero associated $V_{i,j}$ element.

To understand from where the covariance come from: look at **conditional covariance**

2 dimensional conditional covariances:

$$V_{i,j | -\{i,j\}} = V_{i,j} - V_{i,-\{i,j\}} \left(V_{-\{i,j\},-\{i,j\}} \right)^{-1} V_{-\{i,j\},j}$$

also works when $j = i$

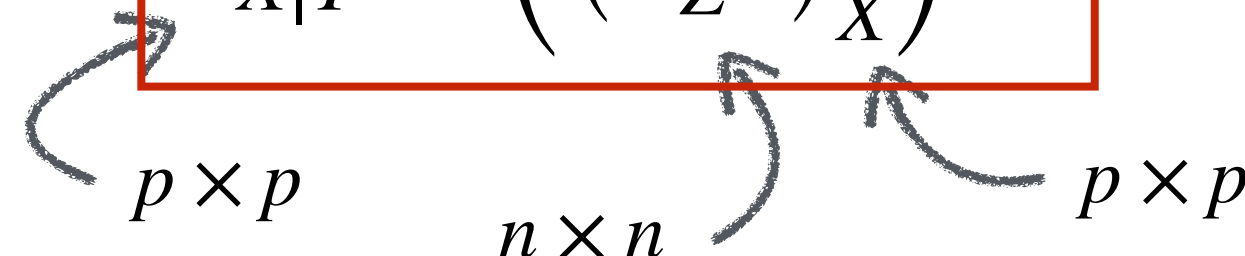
$$V_{i,i | -\{i\}} = V_{i,i} - V_{i,-\{i\}} \left(V_{-\{i\},-\{i\}} \right)^{-1} V_{-\{i\},i}$$

Generic case in n-dim for free variables

$$Z = \overbrace{(X_1, \dots, X_p)}^X, \overbrace{(Y_1, \dots, Y_{n-p})}^Y$$

The double inversion method

$$V_{X|Y} = \left((V_Z^{-1})_X \right)^{-1}$$

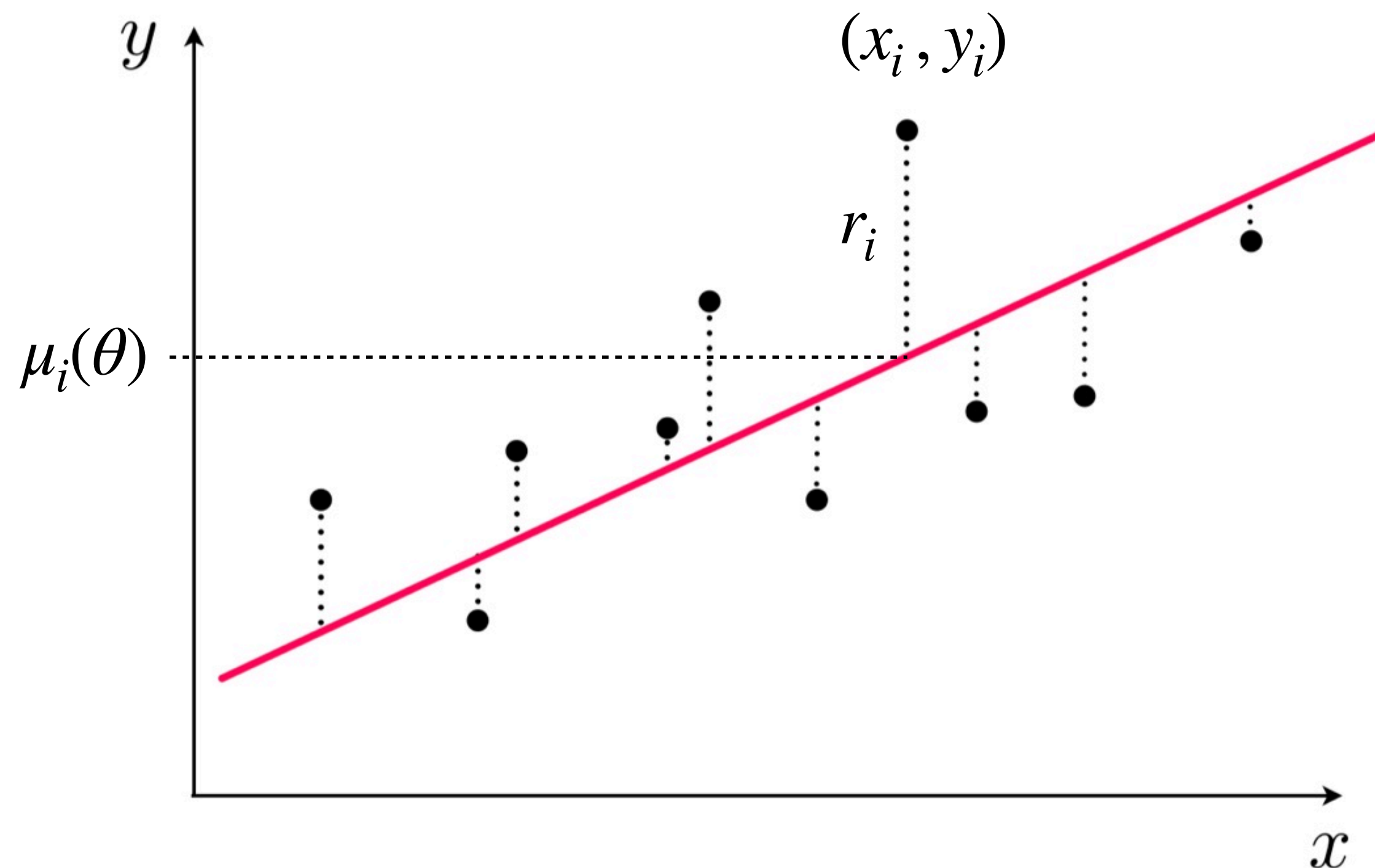


*The proof involves Schur complement and Sherman-Morrison-Woodbury identity
For instance check Wikipedia for further details...*

Estimation of parameters

Note: In the remaining of this lecture, we focus on the **frequentist interpretation of probability**

Ordinary least squares estimation (OLS)



Simply compute the sum of the **squared distances** between data points and model:

$$\sum_{i=1}^n (y_i - \mu_i(\theta))^2 = \sum_{i=1}^n r_i^2$$

Here the model is a line where **generic parameters "θ"** are slope ***a*** and intercept ***b***

$$\mu_i(\theta) = a x_i + b$$

the quantity $r_i = y_i - \mu_i(\theta)$ is called the i^{th} **residual**

The **best model** is the one which **is the closest on average** to the **data**, thus which **minimizes this sum of the squared residuals**

Variants used in practice: Weighted LS (WLS) when measurements have unequal uncertainties σ_i ($w_i = 1/\sigma_i^2$),
Generalized LS (GLS) when measurements are correlated (full covariance matrix V)

Generalisation of least squares

Weighted least squares (WLS)

Weights on data entries: from frequencies, uncertainties, priors on data (robust fitting) etc.

$$\sum_{i=1}^n w_i \cdot (y_i - \mu_i(\theta))^2 \quad \text{e.g. least squares with uncertainties on } y_i \text{ values } \sigma_i \text{ then use } w_i = 1/\sigma_i^2$$

Generalised least squares (GLS) Full covariance $V_{i,j} = \mathbb{V}[Y_i, Y_j]$
$$\sum_{i=1}^n V_{i,j}^{-1} (y_i - \mu_i(\theta)) (y_j - \mu_j(\theta)) = (y - \mu(\theta))^T V^{-1} (y - \mu(\theta))$$

Caution: in all the previous cases, **weights and covariance do not depend on θ**

Otherwise, use the following:

Iteratively reweighted least squares (IRLS) for handling the case when the **covariance matrix depends on parameters**

$$(y - \mu(\theta))^T V(\theta)^{-1} (y - \mu(\theta)) \longrightarrow \text{decouple mean and variance} \quad (y - \mu(\theta^{(n+1)}))^T V(\theta^{(n)})^{-1} (y - \mu(\theta^{(n+1)}))$$

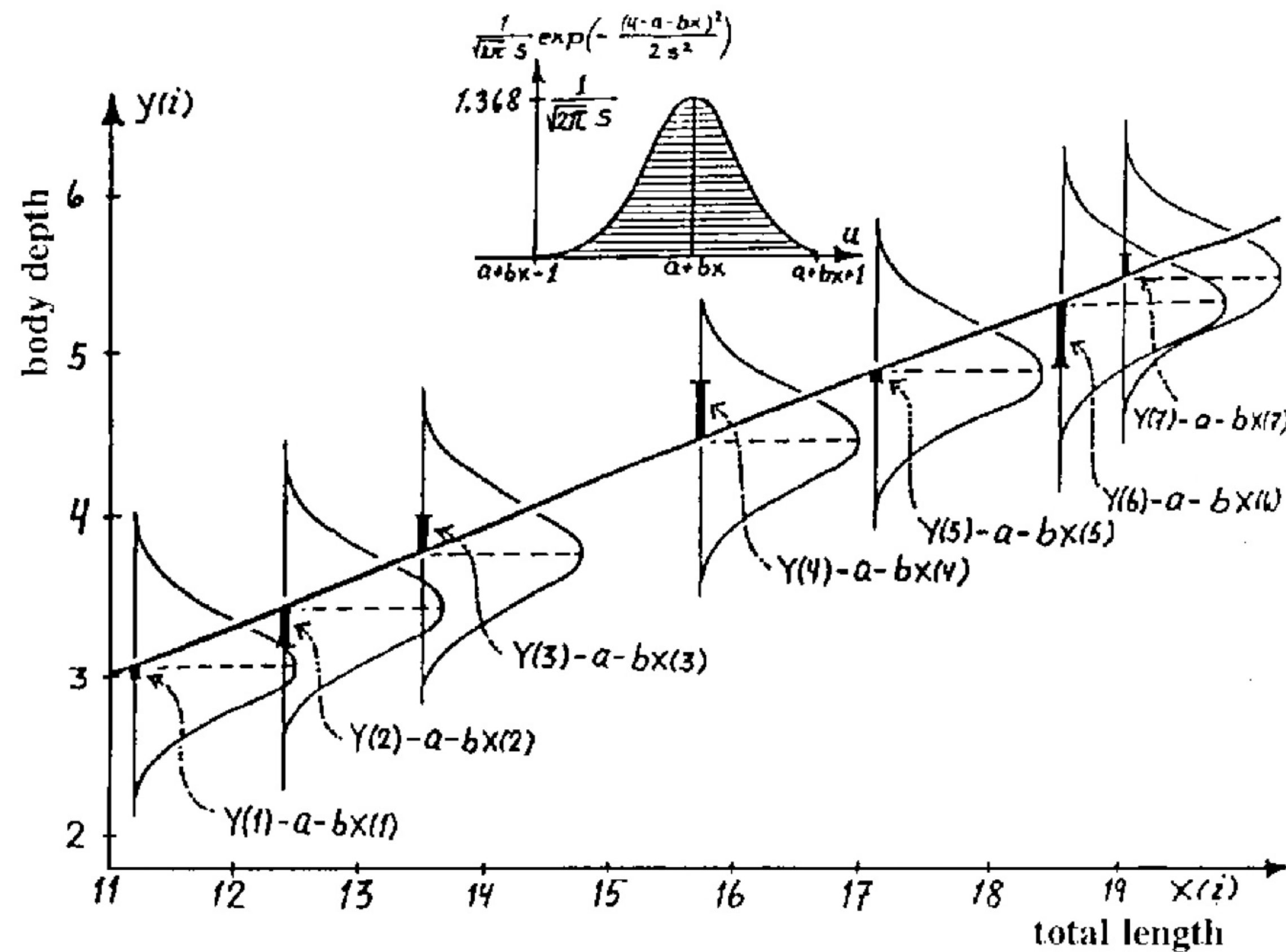
IRLS Algorithm:

- 1/ **Fix** $\theta^{(0)}$ at some a priori initial value. **Fit** $\theta^{(1)}$ in $\mu(\theta^{(1)})$
- 2/ Replace $\theta^{(0)}$ by $\theta^{(1)}$ in $V(\theta)$ and fix it. Fit $\theta^{(2)}$ in $\mu(\theta^{(2)})$,...
- ... Proceed iteratively with fitting $\theta^{(n+1)}$ with $V(\theta)$ fixed at $\theta = \theta^{(n)}$ obtained from previous step until reaching convergence i.e. $\|\theta^{(n+1)} - \theta^{(n)}\| < \text{tolerance}$.

Probabilities with least squares

χ^2 and the statistical interpretation of the Least squares

(χ^2 pronounced "khi" square, and often written chi square)



Hypothesis

Assume normal distribution of errors

assume $y_i = \mu_i(\theta) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$

Then the quantity $\sum_{i=1}^n \left(\frac{y_i - \mu_i(\theta)}{\sigma_i} \right)^2$ is distributed as the sum of n squared standard normal

$$\sum_{i=1}^n Z_i^2 \text{ with } Z_i \sim \mathcal{N}(0, 1)$$

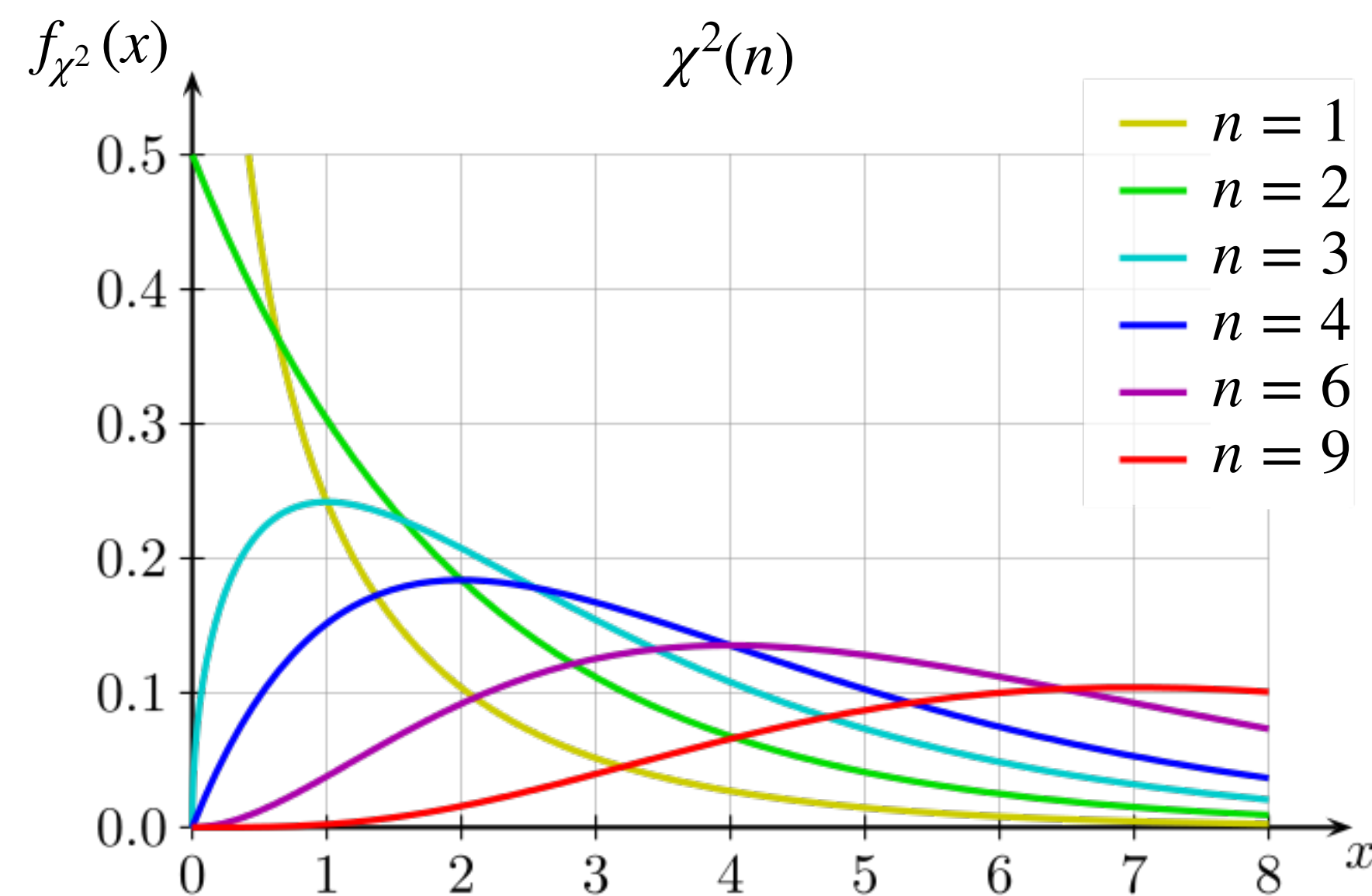
This distribution is known as a χ^2 **distribution** with parameter n , called **degrees of freedom**

Chi square (χ^2) distribution

n random variable $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ normal distribution of mean μ_i and standard deviation σ_i

Centered and reduced random variables $Z_i = \frac{X_i - \mu_i}{\sigma_i} \sim \mathcal{N}(0,1)$

Then the quantity $\chi^2 = \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$ follows a χ^2 distribution with parameter n (degrees of freedom)



Chi square PDF

$$f_{\chi^2}(x; n) = \begin{cases} \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Expectation $\mathbb{E}[\chi^2] = n$

Variance $\mathbb{V}[\chi^2] = 2n$

In Python
scipy.stats.chi2(n).

pdf
cdf
ppf
rvs

Minimum chi square $\chi^2_{\min}$ and Delta chi square $\Delta\chi^2$

follows a χ^2 with n degrees of freedom (**dof**)

if $y_i \sim \mathcal{N}(\mu_i(\theta), \sigma_i^2)$

for $i = 1, \dots, n$

$\theta = (\theta_1, \dots, \theta_p)$

$$\chi^2(y; \mu(\theta)) = \sum_{i=1}^n \left(\frac{y_i - \mu_i(\theta)}{\sigma_i} \right)^2 = \Delta\chi^2 + \chi^2_{\min}$$

follows a χ^2 with $n - p$ degrees of freedom
Used to assess **agreement** between **model & data**

follows a χ^2 with p degrees of freedom
Used to extract **uncertainty on parameters**

Why $n - p$? Because we impose p **restrictions**

$$\left. \frac{\partial \chi^2(y; \theta)}{\partial \theta_k} \right|_{\theta=\hat{\theta}} = 0 \text{ with } k = 1, \dots, p$$

Why p ? Because we get $\hat{\theta}$ estimate from p **equations** $\frac{\partial \chi^2(y; \theta)}{\partial \theta_k} = 0$ with $k = 1, \dots, p$

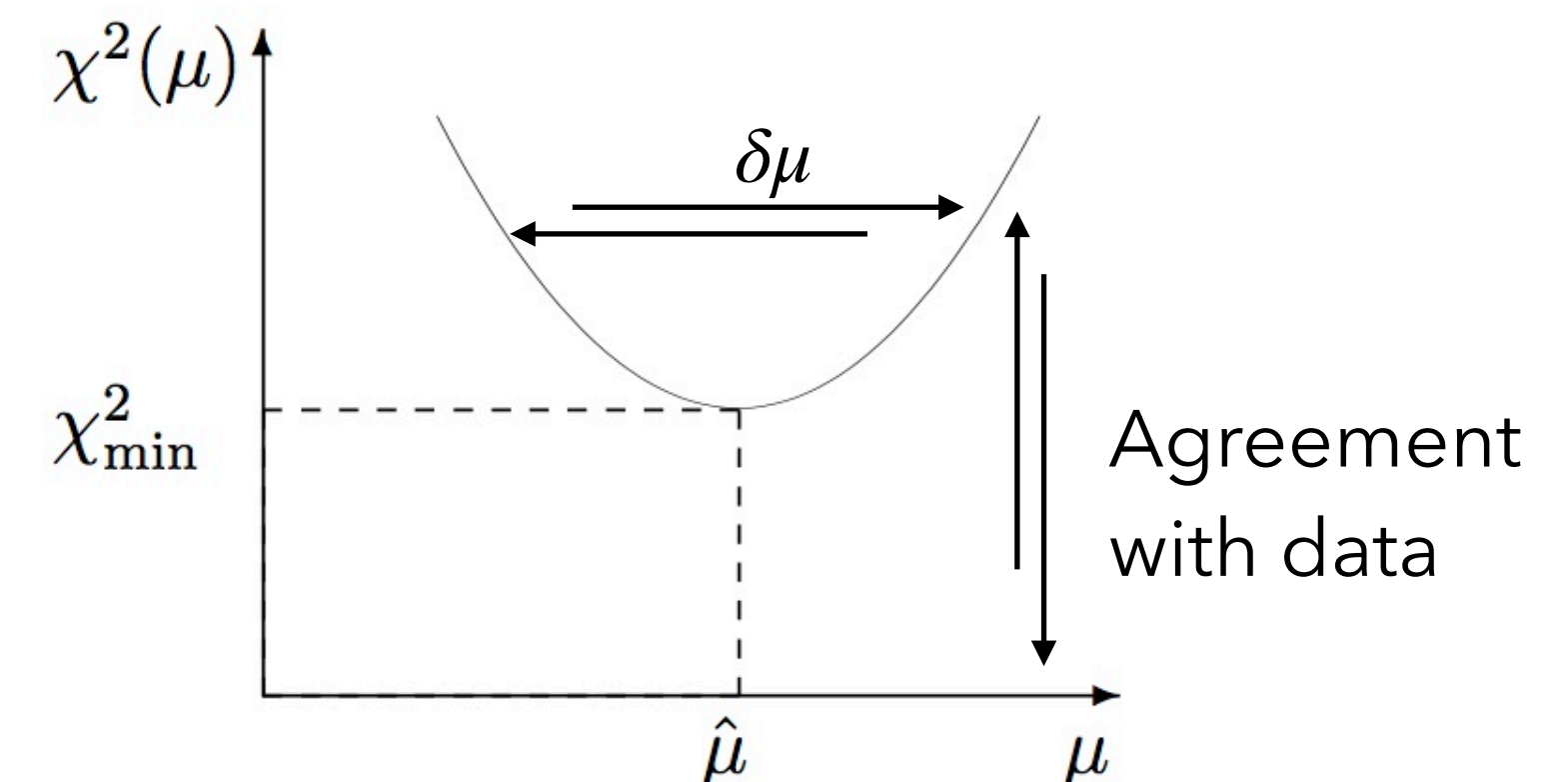
$\Delta\chi^2$ and $\chi^2_{\min}$ are **independent χ^2 random variables** (Cochran theorem)

Ex: single parameter case $\chi^2(y; \mu) = \left(\frac{\mu - \hat{\mu}}{\sigma_\mu} \right)^2 + \chi^2_{\min}$

χ^2 with 1 dof $\quad \quad \quad \chi^2$ with $n - 1$ dof

$\Rightarrow$ To find 1σ uncertainty on μ use $\Delta\chi^2 = \chi^2 - \chi^2_{\min} = +1$

Chi square profile



Goodness of fit with χ^2

The $\chi^2_{\min}$ PDF has an expectation value equal to the number of degrees of freedom $n - p$

So if $\chi^2_{\min} \simeq n - p$ or the reduced $\chi^2_{\text{red}} = \frac{\chi^2_{\min}}{n - p} \simeq 1 \longrightarrow$ the fit is "good" (i.e. as expected!)

More precisely,

- $\chi^2_{\text{red}} = \frac{\chi^2_{\min}}{n - p} \simeq 1 \pm \sigma_{\text{red}}$ with $\sigma_{\text{red}} = \sqrt{\frac{2}{n - p}} \longrightarrow$ all is as expected
- $\chi^2_{\text{red}} = \frac{\chi^2_{\min}}{n - p} \ll 1 \pm \sigma_{\text{red}} \longrightarrow$ the fit is better than expected given the measurement uncertainties.

This is not bad in the sense of providing evidence against the model, but it is usually better to check if the uncertainties σ_i have not been overestimated or are not correlated...

- $\chi^2_{\text{red}} = \frac{\chi^2_{\min}}{n - p} \gg 1 \pm \sigma_{\text{red}} \longrightarrow$ then there is some reason to doubt the model in use...

Note that each statement can be quantitatively assessed using the χ^2 CDF

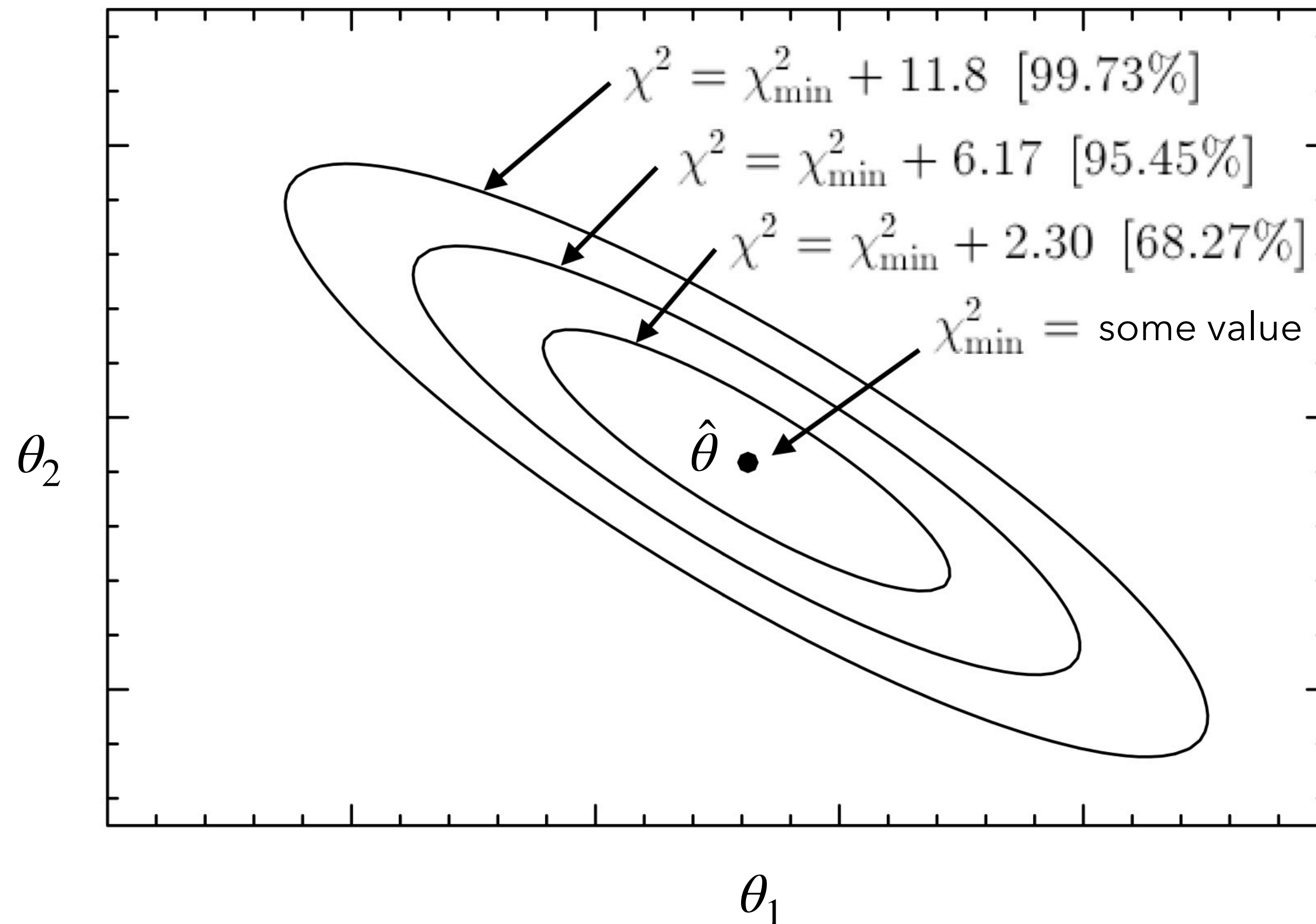
The p-value is defined as $\int_{\chi^2_{\min}}^{+\infty} f_{\chi^2}(t; n - p) dt = 1 - F_{\chi^2}(\chi^2_{\min}; n - p)$ i.e. the probability to find a $\chi^2_{\min}$ at least as extreme as the observed one.

Some warning
Variance $\mathbb{V}[\chi^2] = 2n$
so $\mathbb{V}[\chi^2/n] = 2/n$
i.e. "close to 1" is with
respect to $\sqrt{\mathbb{V}[\chi^2/n]}$

Example of χ^2 contours in 2D

To find 1σ uncertainty on μ in **1D** use $\Delta\chi^2 = \chi^2 - \chi^2_{\min} = +1$

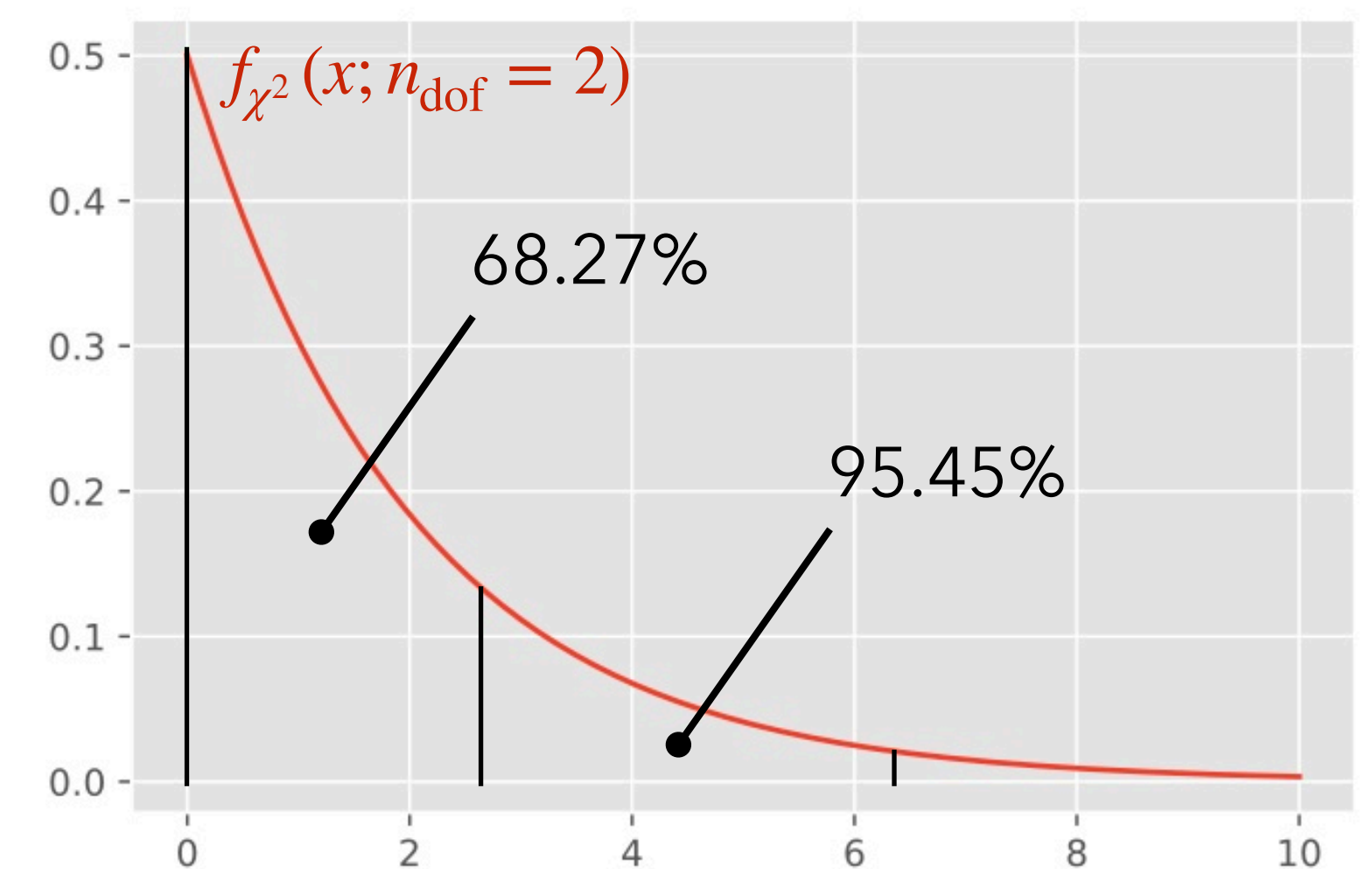
=> But, in **2D**?



Example of contours with **2 degrees** of freedom (i.e. 2 parameters) with different **probability content** corresponding to what we call **1, 2 and 3 σ**

The **offset levels** correspond to the **inverse CDF** of the χ^2 (also known as the **quantile distribution function** of the χ^2) for **2 degrees of freedom**

In Python, χ^2 quantiles: `scipy.stats.chi2(n).ppf`



Profiling the χ^2

Consider the case of a χ^2 with 2 parameters: μ is the **parameter of interest** (POI)
 θ is a **nuisance parameter** (NP)

θ can be for instance a systematic effect parameter.

A systematic error is, in any statistical inference procedure, the error due to the incomplete knowledge of the probability distribution of the observables. It could be a fixed effect (like an energy scale), or a random effect ^(★) (like a response function PSF,...).

$$\chi^2(y; \mu, \theta) = \sum_{i=1}^n \left(\frac{y_i - g_i(\mu, \theta)}{\sigma_i} \right)^2$$

The generic receipt in frequentist context to get rid of a fixed effect parameter θ but to still take into account its effect is to **minimize the χ^2 with respect to θ for each value of μ**
 $\hat{\theta}(\mu) = \operatorname{argmin}_{\theta} \chi^2(y; \mu, \theta)$ and replace θ by $\hat{\theta}(\mu)$ in $\chi^2(y; \mu, \theta)$

$$\chi_{\text{prof}}^2(y; \mu) = \chi^2(y; \mu, \hat{\theta}_{\mu}) = \sum_{i=1}^n \left(\frac{y_i - g_i(\mu, \hat{\theta}_{\mu})}{\sigma_i} \right)^2$$

This profiled χ^2 is now only a function of μ which keep asymptotic χ^2 properties

^(★) random effects are taken into account by averaging their effect (i.e. marginalization)

Be careful however: this process reduces the number of degrees of freedom $n \rightarrow n - p$ where p is $\dim(\{\theta's\})$

Profiling the χ^2 - illustration close to best fit

1 parameter of interest (POI) μ and 1 nuisance parameter (NP) θ

$-\frac{1}{2} \begin{pmatrix} \mu - \hat{\mu} \\ \theta - \hat{\theta} \end{pmatrix}^T V^{-1} \begin{pmatrix} \mu - \hat{\mu} \\ \theta - \hat{\theta} \end{pmatrix}$ defines ellipses in (μ, θ) parameter space

Generic ellipse definition

$$F_{\mu\mu}(\mu - \hat{\mu})^2 + 2F_{\mu\theta}(\mu - \hat{\mu})(\theta - \hat{\theta}) + F_{\theta\theta}(\theta - \hat{\theta})^2$$

Uncertainty on μ :

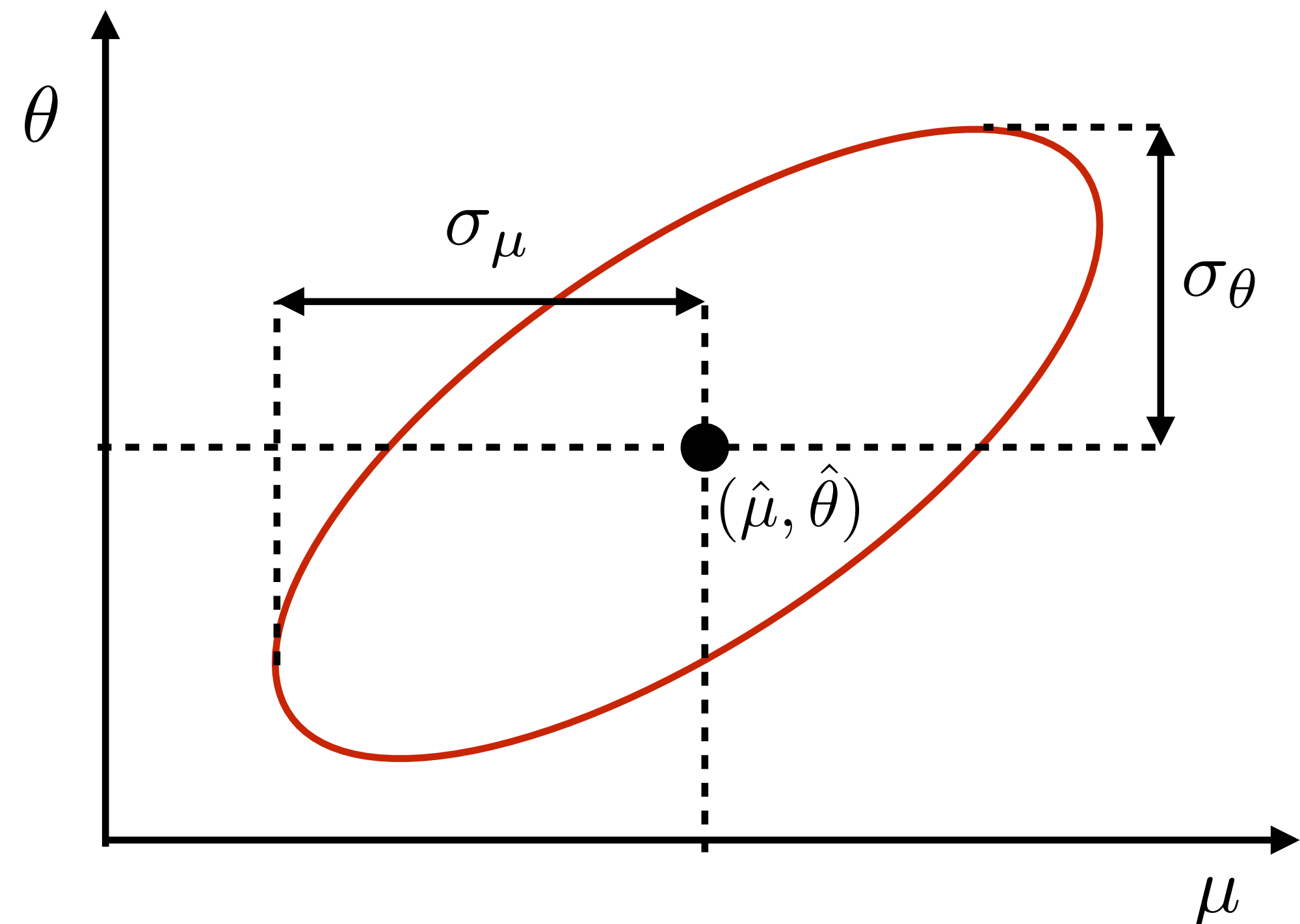
- From V , with θ included: σ_μ

Parameter covariance matrix

$$V = \begin{pmatrix} \sigma_\mu^2 & \rho \sigma_\mu \sigma_\theta \\ \rho \sigma_\mu \sigma_\theta & \sigma_\theta^2 \end{pmatrix}$$

Fisher (Information) matrix

$$F = \begin{pmatrix} F_{\mu\mu} & F_{\mu\theta} \\ F_{\mu\theta} & F_{\theta\theta} \end{pmatrix} \quad F = V^{-1}$$



Profiling the χ^2 - illustration close to best fit

$$F_{\mu\mu}(\mu - \hat{\mu})^2 + 2F_{\mu\theta}(\mu - \hat{\mu})(\theta - \hat{\theta}) + F_{\theta\theta}(\theta - \hat{\theta})^2$$

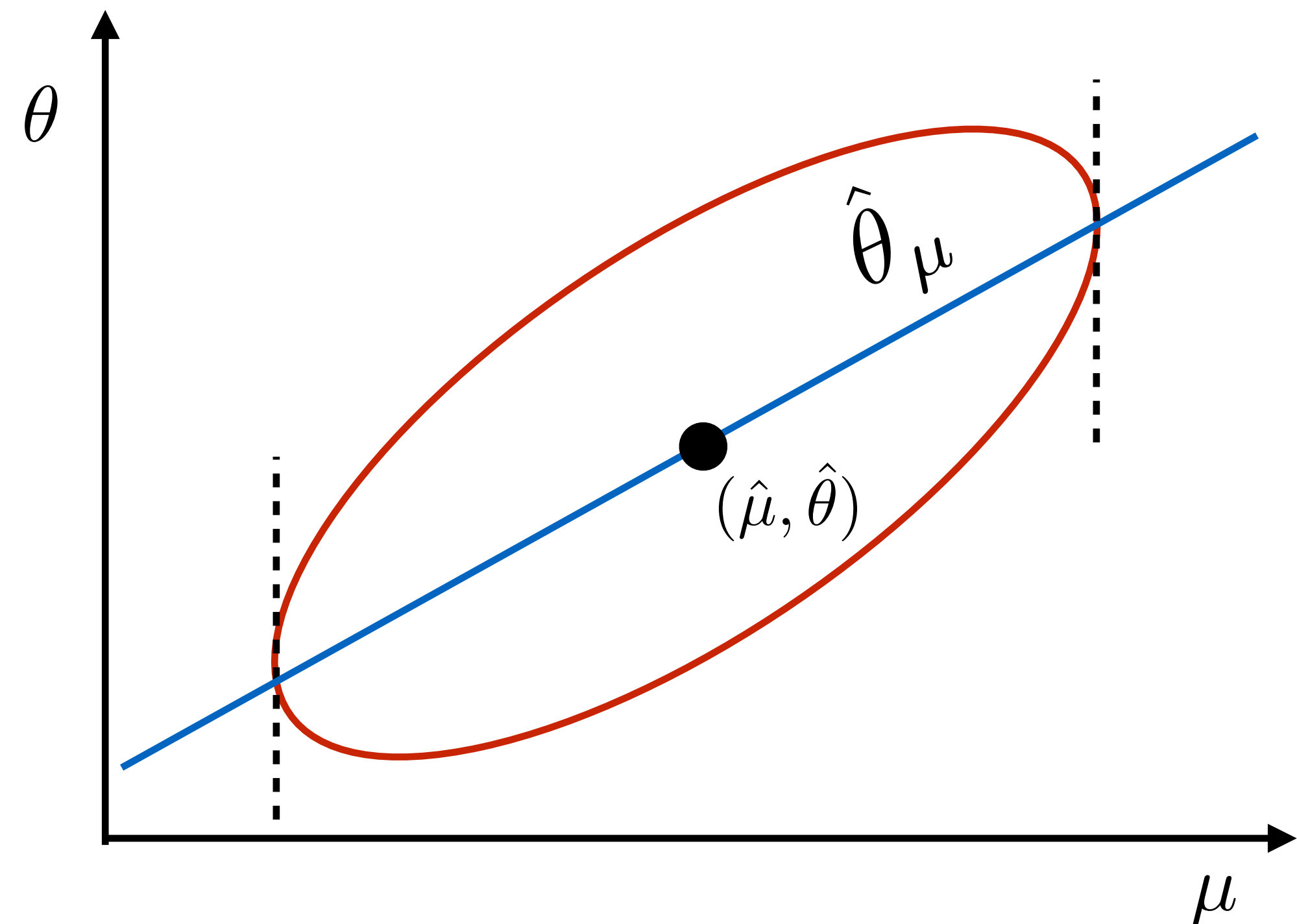
Profiled θ (minimise at fixed μ): $\hat{\theta}_\mu = \hat{\theta} - F_{\theta\theta}^{-1}F_{\theta\mu}(\mu - \hat{\mu})$

Profiled χ^2 $(F_{\mu\mu} - F_{\mu\theta}F_{\theta\theta}^{-1}F_{\theta\mu})(\mu - \hat{\mu})^2 = V_{\mu\mu}^{-1}(\mu - \hat{\mu})^2 = \left(\frac{\mu - \hat{\mu}}{\sigma_\mu}\right)^2$ $F_{\mu\mu} \neq V_{\mu\mu}^{-1}!!!$

Uncertainty on μ :

- From V , with θ included: σ_μ
- From profiled χ^2 : σ_μ

Profiled θ crosses ellipse at vertical tangents by definition
(χ^2 is higher at other points on the tangent)



Profiling the χ^2 - illustration close to best fit

$$F_{\mu\mu}(\mu - \hat{\mu})^2 + 2F_{\mu\theta}(\mu - \hat{\mu})(\theta - \hat{\theta}) + F_{\theta\theta}(\theta - \hat{\theta})^2$$

Now, for fixed $\theta = \hat{\theta}$ (i.e. conditioning),
defines another interval:

$$F_{\mu\mu}(\mu - \hat{\mu})^2 = \left(\frac{\mu - \hat{\mu}}{\sigma_\mu \sqrt{1 - \rho^2}} \right)^2$$

$$F = V^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1/\sigma_\mu^2 & -\rho/\sigma_\mu\sigma_\theta \\ -\rho/\sigma_\mu\sigma_\theta & 1/\sigma_\theta^2 \end{pmatrix}$$

Uncertainty on μ :

- From V , with θ included: σ_μ
- From profiled χ^2 : σ_μ **total uncertainty**
- From fixed $\theta = \hat{\theta}$: $\sigma_\mu \sqrt{1 - \rho^2}$ **conditional uncertainty**

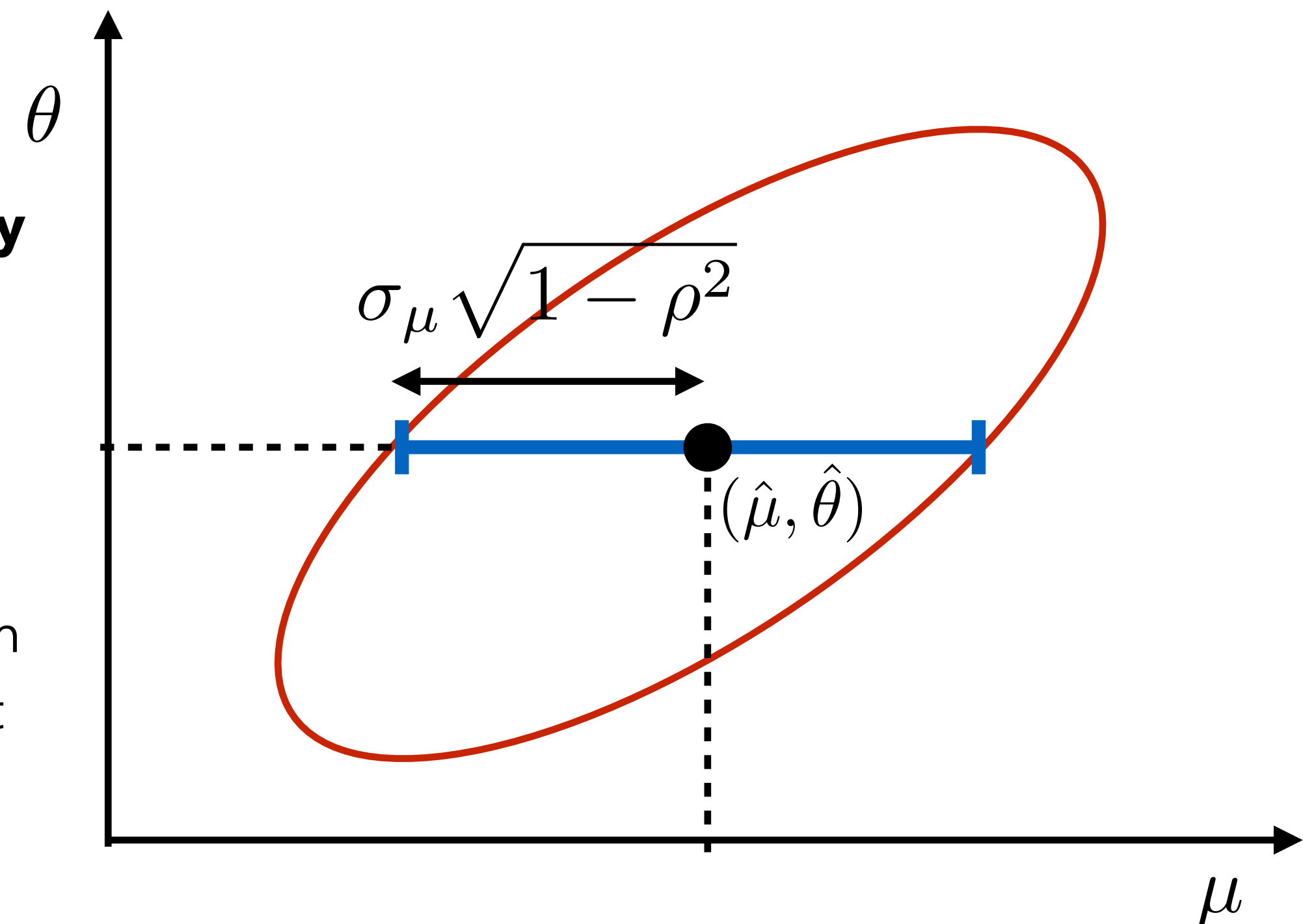
$$\sigma_\mu = \sqrt{\left(\sigma_\mu \sqrt{1 - \rho^2} \right)^2 + (\rho \sigma_\mu)^2}$$

Total uncertainty

conditional
uncertainty on $\theta = \hat{\theta}$

uncertainty from
nuisance effect

Profiling on nuisance parameters indeed **takes into account nuisance effects**



The covariance and the pull approach in χ^2

Consider for instance the case where we would like to determine the mean μ of the data y_i

But our experiment is subject to some **systematic effect** S_i

We think we have carefully removed the systematic effect but there remains **some uncertainty** σ_S

We think the **effect is fully correlated across** y_i values subject to $\pm\sigma_S \times S_i$.

We have good reason to assume these systematic uncertainty follow a normal distribution

We can model this situation with the following χ^2

$$\chi^2(y; \mu, \theta) = \sum_{i=1}^n \left(\frac{y_i - \mu - S_i \theta}{\sigma_i} \right)^2 + \left(\frac{\theta - \theta_0}{\sigma_S} \right)^2$$

here $\theta_0 = 0$

$n + 1$ degrees of freedom
2 parameters

often called "pull term"

If we minimise this χ^2 with respect to θ (profiling to get rid of nuisance parameter) we can demonstrate that the obtained χ^2 is

$$\chi^2(y; \mu, \hat{\theta}) = \sum_{i,j=1}^n (y_i - \mu) V_{i,j}^{-1} (y_j - \mu) \quad \text{with } V_{i,j} = \sigma_i^2 \delta_{i,j} + \sigma_S^2 S_i S_j$$

n degrees of freedom
1 parameter

This equivalence is exact in linear parameter case, approximate in non-linear case.

The pulls approach is often preferred as you collect information about best fit nuisance parameters in the way.

Maximum Likelihood Estimation

The Likelihood function

If in $\mathbb{P}(\text{data} \mid \text{hypothesis})$, we put in the values of the data observed in the experiment, and consider the resulting function as a function of the unknown parameter(s), it becomes

$$\mathbb{P}(\text{data} \mid \text{hypothesis}) \Big|_{\text{data obs.}} = \mathcal{L}(\text{hypothesis})$$

$\mathcal{L}$ is called the **Likelihood Function**.

R. A. Fisher, the first person to use it, knew that it was **not a probability**, so he called it a **likelihood**. The likelihood is **not** a probability distribution over parameters. It's a function that, for each parameter value, tells you how probable the observed data would have been.

Higher likelihood for a parameter value means that parameter value is better supported by the data. But – and this is essential – you cannot interpret the ratio $\mathcal{L}(\theta_1)/\mathcal{L}(\theta_2)$ directly as an odds ratio for the parameters. It's not a probability!

It will turn out to have some important properties.

Maximum Likelihood Estimation

Define the likelihood of the sample with $x = (x_1, \dots, x_N)$ independent and identically distributed (*iid*) random variables from the same PDF $f_X(x_i; \theta)$

$$\mathcal{L}(x_1, \dots, x_N; \theta) = \prod_{i=1}^N f(x_i; \theta)$$

The **Maximum Likelihood Estimator (MLE)** of the parameter $\theta^\star$ is the value $\hat{\theta}$ for which $\mathcal{L}(x_1, \dots, x_N; \theta)$ has its maximum given the sample $(x_1, \dots, x_N)$

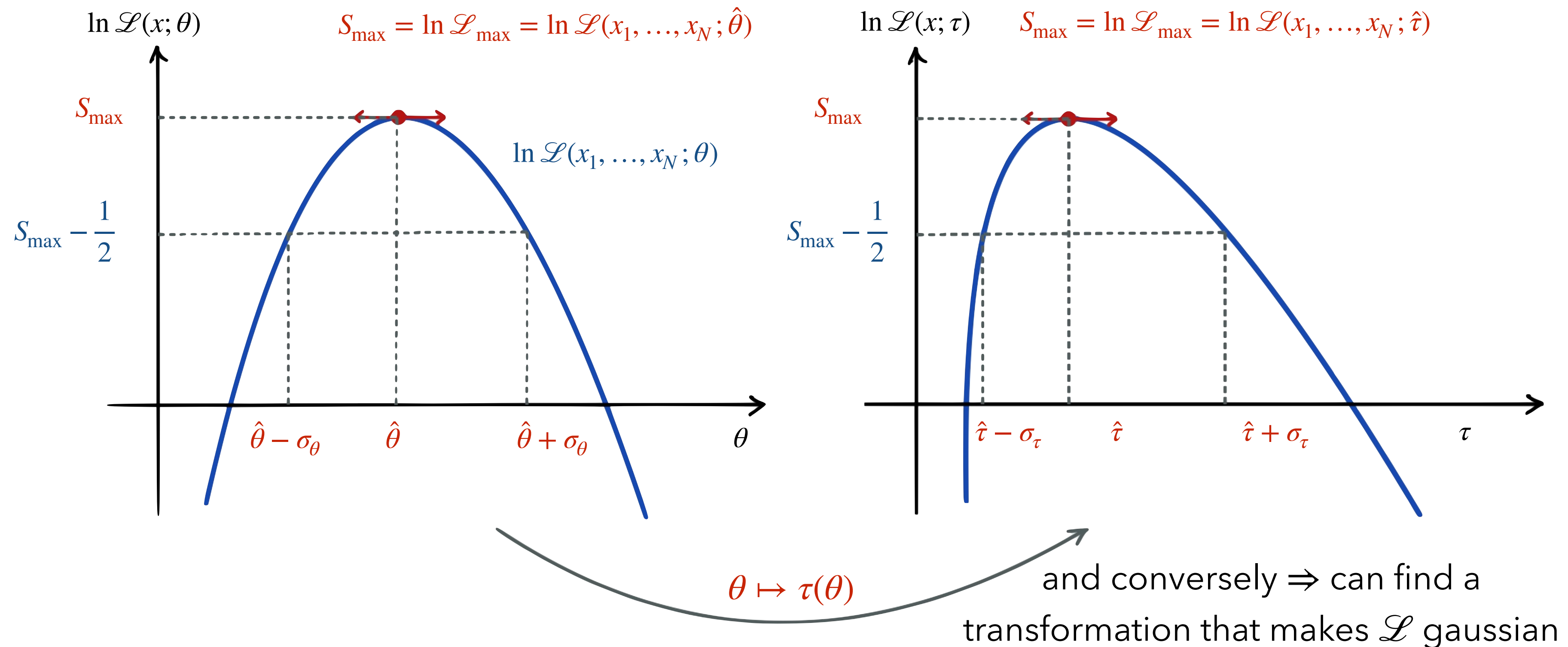
The log-likelihood is called the **score**, $S(x; \theta) = \ln \mathcal{L}(x; \theta) = \sum_{i=1}^N \ln f(x_i; \theta)$,

The **likelihood estimating equation** is $\frac{\partial S(x; \theta)}{\partial \theta} = \frac{\ln \partial \mathcal{L}(x; \theta)}{\partial \theta} = \sum_{i=1}^N \frac{\partial \ln f(x_i; \theta)}{\partial \theta} = 0$

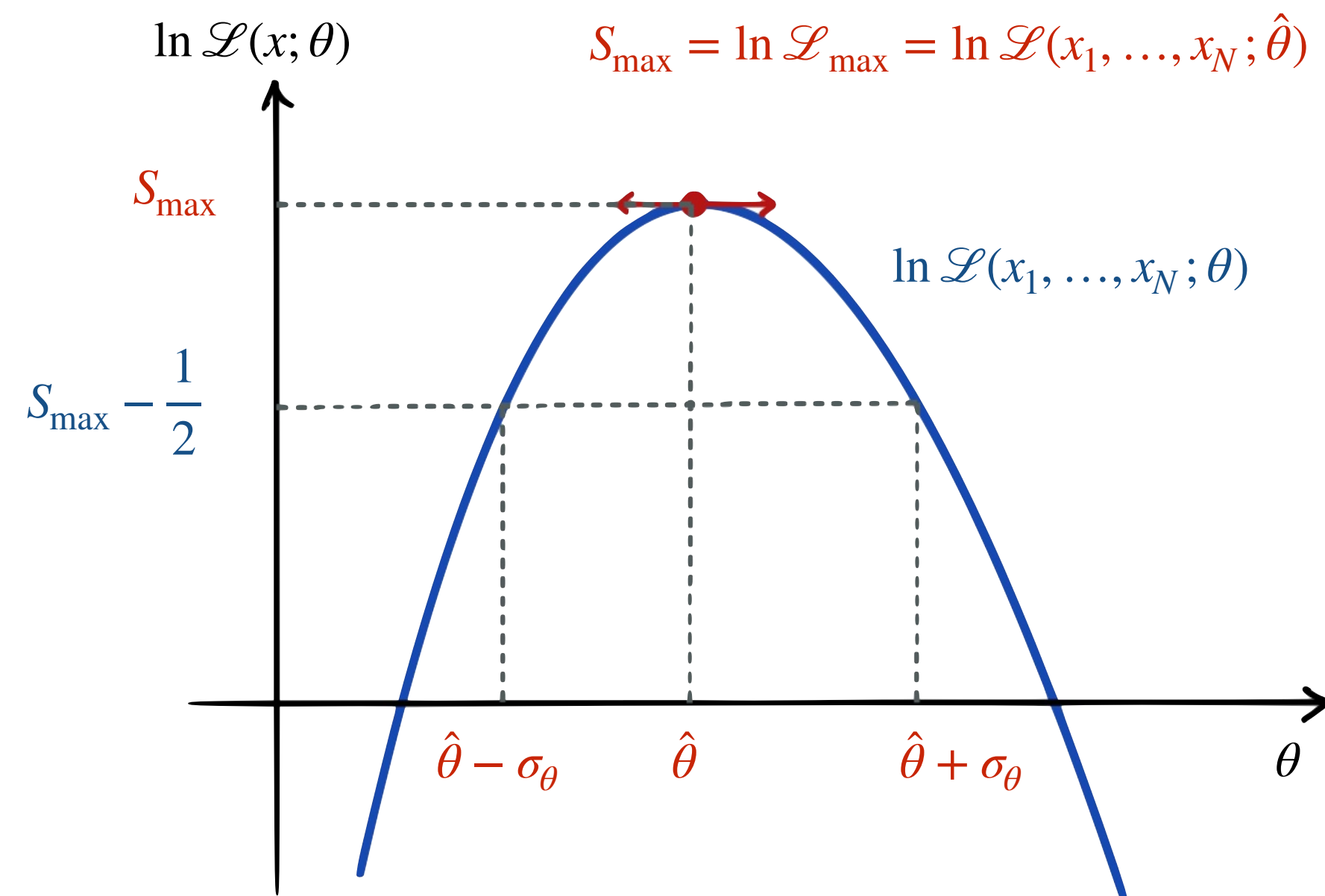
and the estimator $\hat{\theta}$ of θ is a root of the likelihood estimating equation, when it exists.

MLE properties

- consistent
- asymptotically normally distributed, with minimum variance
- for finite N , optimal under Darmois theorem with exponential family distributions
 $f(x; \theta) = \exp(a(x) \cdot \alpha(\theta) + b(x) + \beta(\theta))$, sufficient statistics, Cramer-Rao Lower Bound.
- invariant under transformation of the parameter: the MLE of $\hat{\tau}$ of $\tau(\theta)$ is $\hat{\tau} = \tau(\hat{\theta})$



MLE uncertainty on θ



Since $\mathcal{L}$ is asymptotically gaussian (CLT) and we can always reparameterize $\theta \mapsto \tau(\theta)$ to get a gaussian profile for the parameter. Then, $S_{\max} - \frac{1}{2} = \ln \mathcal{L}_{\max} - \frac{1}{2}$ gives the 1σ uncertainty on θ through:

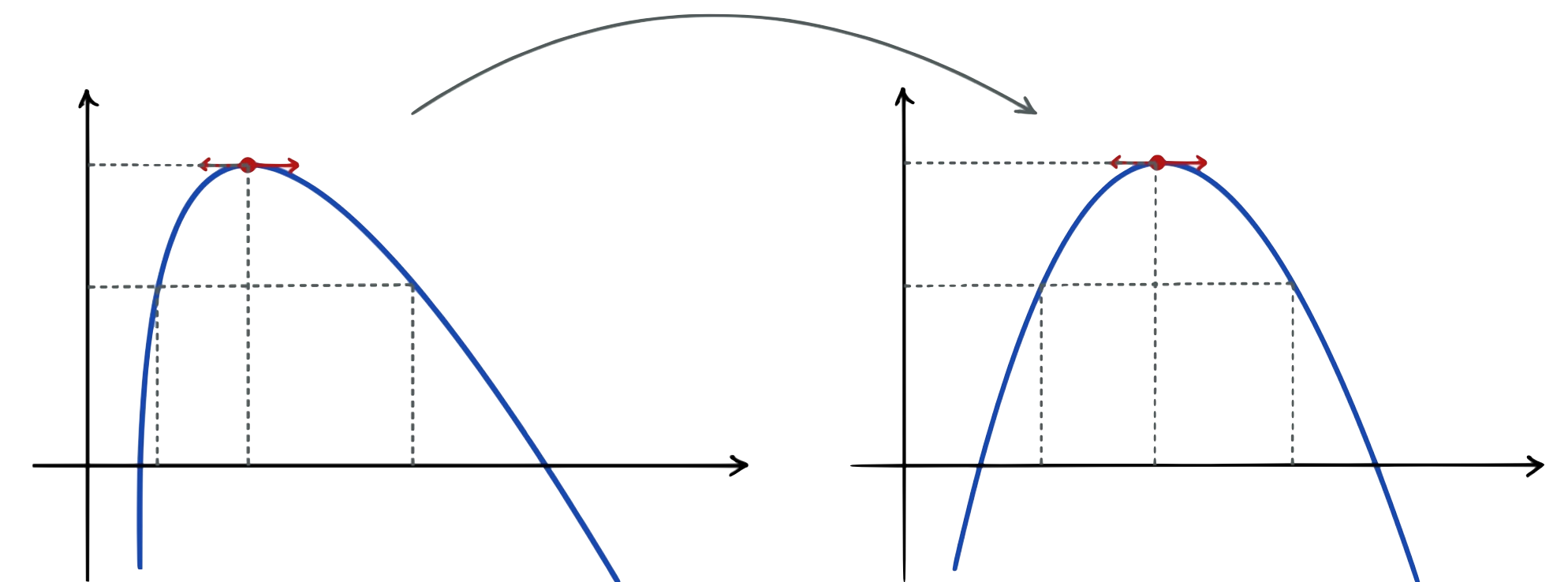
$$\ln \mathcal{L}(x; \theta \pm \sigma_{\theta}) = \ln \mathcal{L}_{\max} - \frac{1}{2}$$

The interval $[\hat{\theta} - \sigma_{\theta}; \hat{\theta} + \sigma_{\theta}]$ obtained this way is called the **likelihood interval**

Even with **non-gaussian likelihood function** (i.e. non-parabolic $\ln \mathcal{L}$)

Use of invariance property allows to find a transformation that makes $\mathcal{L}$ gaussian and the content of the interval is preserved. Can be used to determine the confidence intervals without actually making the transformation to gaussian.

Caution: these confidence intervals are only approximate for N finite! They are the **asymptotic likelihood confidence intervals**



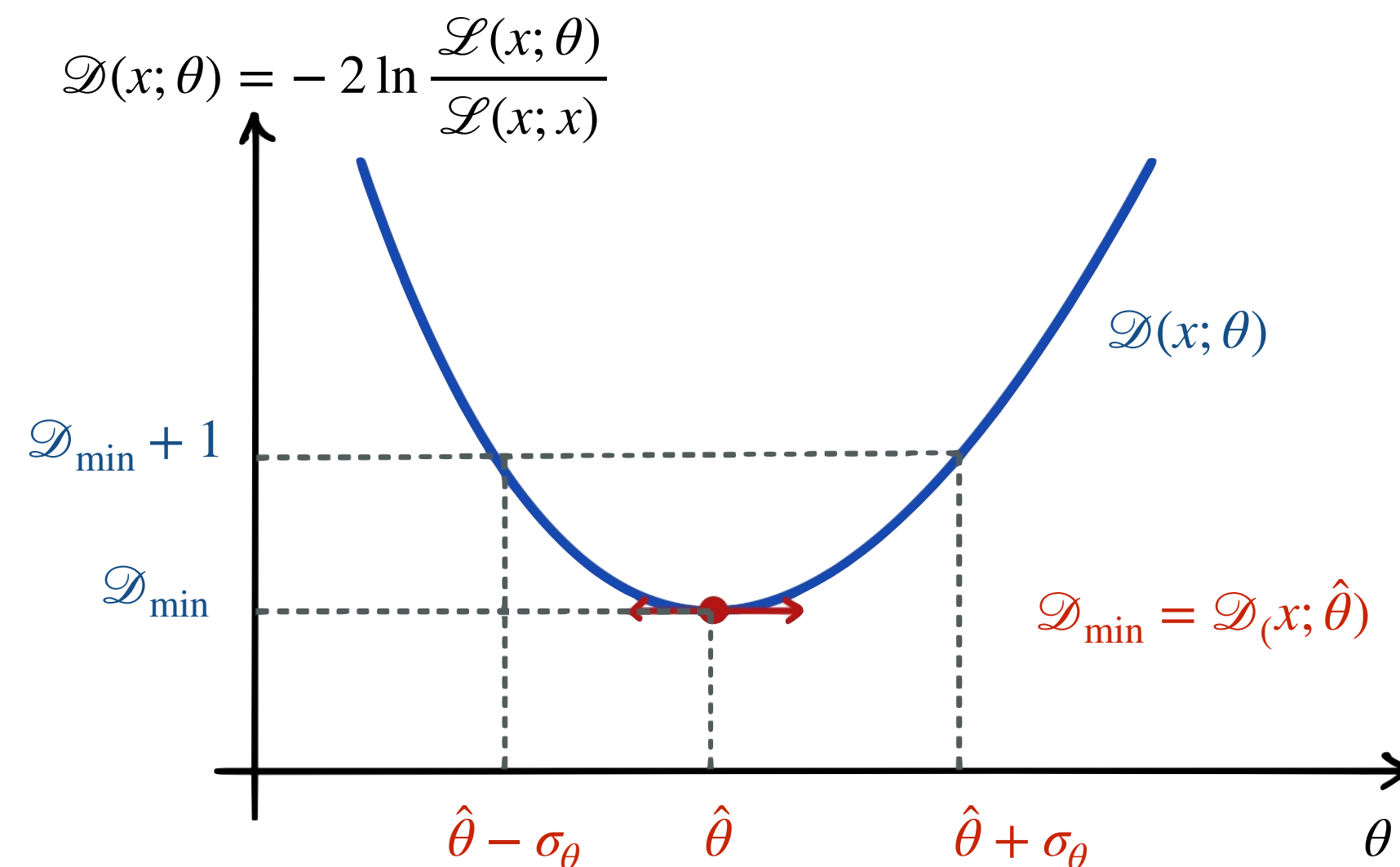
Approaching χ^2 with the likelihood

For several reason it is convenient to take $\ell(x; \theta) = -2 \ln \mathcal{L}(x; \theta)$

We further define the **deviance function** as: $\mathcal{D}(x; \theta) = \ell(x; \theta) - \ell(x; x)$

where $\ell(x; x)$ stands for the likelihood of the data sample with a saturated model (a model with a free parameter for each data point whose best fit reproduce exactly the data set). The deviance measures the departure of the model from data.

With a sample $(x_1, \dots, x_n)$ of independent random variables: $\mathcal{D}(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \mathcal{D}(x_i; \theta)$



Advantage of the deviance:
can check the agreement with the data with $\mathcal{D}_{\min}$ which should follow a χ^2 distribution with parameter (degrees of freedom) $n - p$ where n is the number of data points, and p the number of fitted parameters

With **deviance** you inherit from the χ^2 setup

Deviance of some standard distributions

Poisson deviance $\mathcal{D}(n; \lambda) = \ell(n; \lambda) - \ell(n; n) = -2 \left(n - \lambda + n \ln \left(\frac{\lambda}{n} \right) \right)$

$$\mathbb{P}[n; \lambda] = \frac{\lambda^n}{n!} e^{-\lambda} \quad \begin{aligned} \ell(n; \lambda) &= -2 \ln \mathbb{P}[n; \lambda] = 2\lambda - 2n \ln \lambda + 2 \ln(n!) \\ \ell(n; n) &= -2 \ln \mathbb{P}[n; n] = 2n - 2n \ln n + 2 \ln(n!) \end{aligned}$$

Case of a sample of N independent Poisson variables $(n_1, \dots, n_N)$

$$\sum_{i=1}^N \mathcal{D}(n_i; \lambda_i) = -2 \sum_{i=1}^N n_i - \lambda_i + n_i \ln \frac{\lambda_i}{n_i}$$

Binomial deviance $\mathcal{D}(n; N, p) = \ell(n; N, p) - \ell(n; N, p) = -2 \left(n \ln \left(\frac{Np}{n} \right) + (N - n) \ln \left(\frac{N - Np}{N - n} \right) \right)$

Multinomial deviance $\mathcal{D}(n_1, \dots, n_K; N, p_1, \dots, p_K) = \dots = -2 \sum_{k=1}^K n_k \ln \left(\frac{Np_k}{n_k} \right)$

Normal deviance $\mathcal{D}(y; \mu, \sigma) = \left(\frac{y - \mu}{\sigma} \right)^2 \longrightarrow \mathcal{D}(x; \mu) = \sum_{i=1}^N \mathcal{D}_i(x_i; \mu) = \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma_i} \right)^2$

Some usual ways to define likelihoods

S, B: expected signal & background

Description	Observable	Likelihood
Counting	n : measured number of events	Poisson $P(\textcolor{red}{n}; \textcolor{blue}{S}, \textcolor{blue}{B}) = \frac{(\textcolor{blue}{S} + \textcolor{blue}{B})^{\textcolor{red}{n}}}{\textcolor{red}{n}!} e^{-(\textcolor{blue}{S} + \textcolor{blue}{B})}$
Unbinned shape analysis	x_i : i=1,..., n_{events} observable value for each event	Extended Unbinned Likelihood $P(\textcolor{red}{X} \textcolor{blue}{S}, \textcolor{blue}{B}) = \frac{e^{-(\textcolor{blue}{S} + \textcolor{blue}{B})}}{\textcolor{red}{n}_{\text{evts}}!} \prod_{i=1}^{\textcolor{red}{n}_{\text{evts}}} (\textcolor{blue}{S} P(\textcolor{red}{x}_i \textcolor{blue}{S}) + \textcolor{blue}{B} P(\textcolor{red}{x}_i \textcolor{blue}{B}))$ $P(\textcolor{red}{x}_i \textcolor{blue}{S}), P(\textcolor{red}{x}_i \textcolor{blue}{B})$ PDFs for observing x_i in signal, background
Binned shape analysis	n_i : i=1,...,N bins measured events in each bin	Poisson product $P(\{\textcolor{red}{n}_i\}_{i=1,\dots,N} \textcolor{blue}{S}, \textcolor{blue}{B}) = \prod_{i=1}^N \left[\frac{(\textcolor{blue}{S} p_{S,i} + \textcolor{blue}{B} p_{B,i})^{\textcolor{red}{n}_i}}{\textcolor{red}{n}_i!} e^{-(\textcolor{blue}{S} p_{S,i} + \textcolor{blue}{B} p_{B,i})} \right]$ p_{S,i} & p_{B,i} : probability of signal, background in each bin

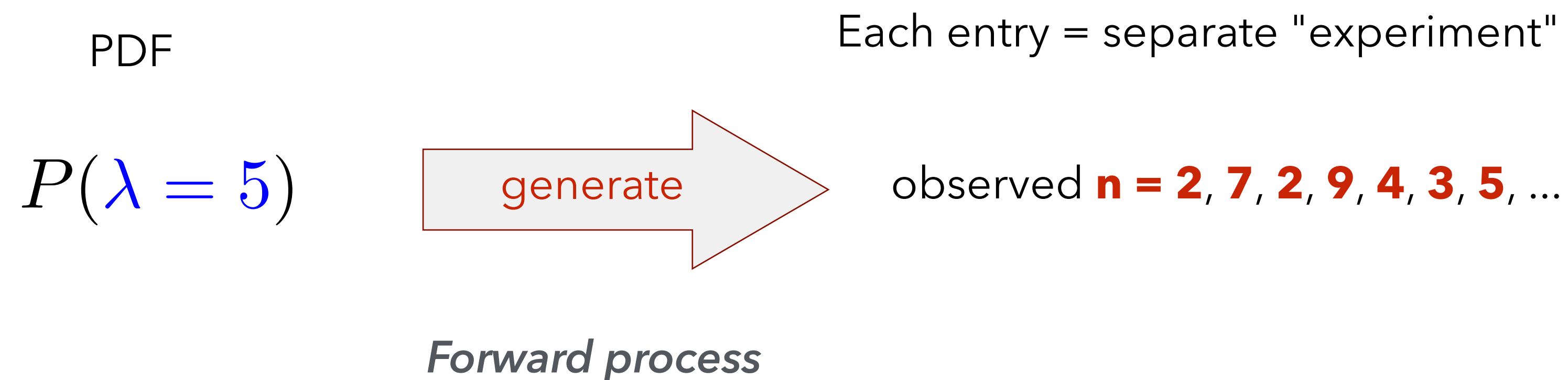
Use with increasing number of events

Illustrating the maximum likelihood estimation

Model describes the distribution of the observable: $P(\textcolor{red}{n}; \textcolor{blue}{S}) = P(\textcolor{red}{data}; \textcolor{blue}{parameters})$

➡ Possible outcomes of the experiment, for given parameter values

Can draw random events according to PDF: generate "***pseudo-data***"
or ***synthetic data***



Illustrating the maximum likelihood estimation

Model describes the distribution of the observable: $P(n; S) = P(\text{data}; \text{parameters})$

⇒ Possible outcomes of the experiment, for given parameter values

We want the other direction: use data to get information on parameters



⇒ **Likelihood**: function of **parameters** = $P(\text{data}; \text{parameters})$

Same as PDF, but "**seen**" as a function of the parameters only

(The likelihood is not a probability distribution over the parameters)

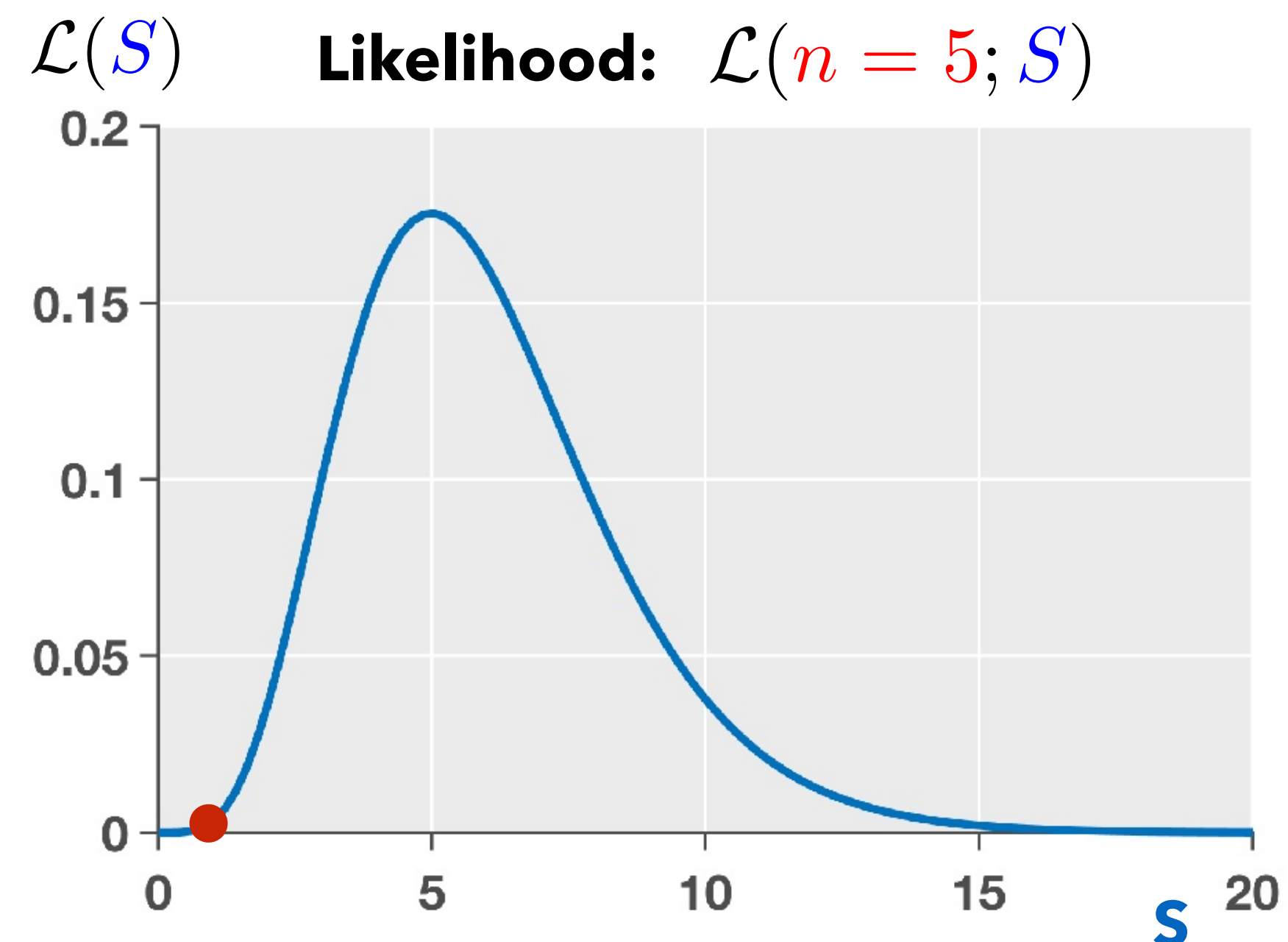
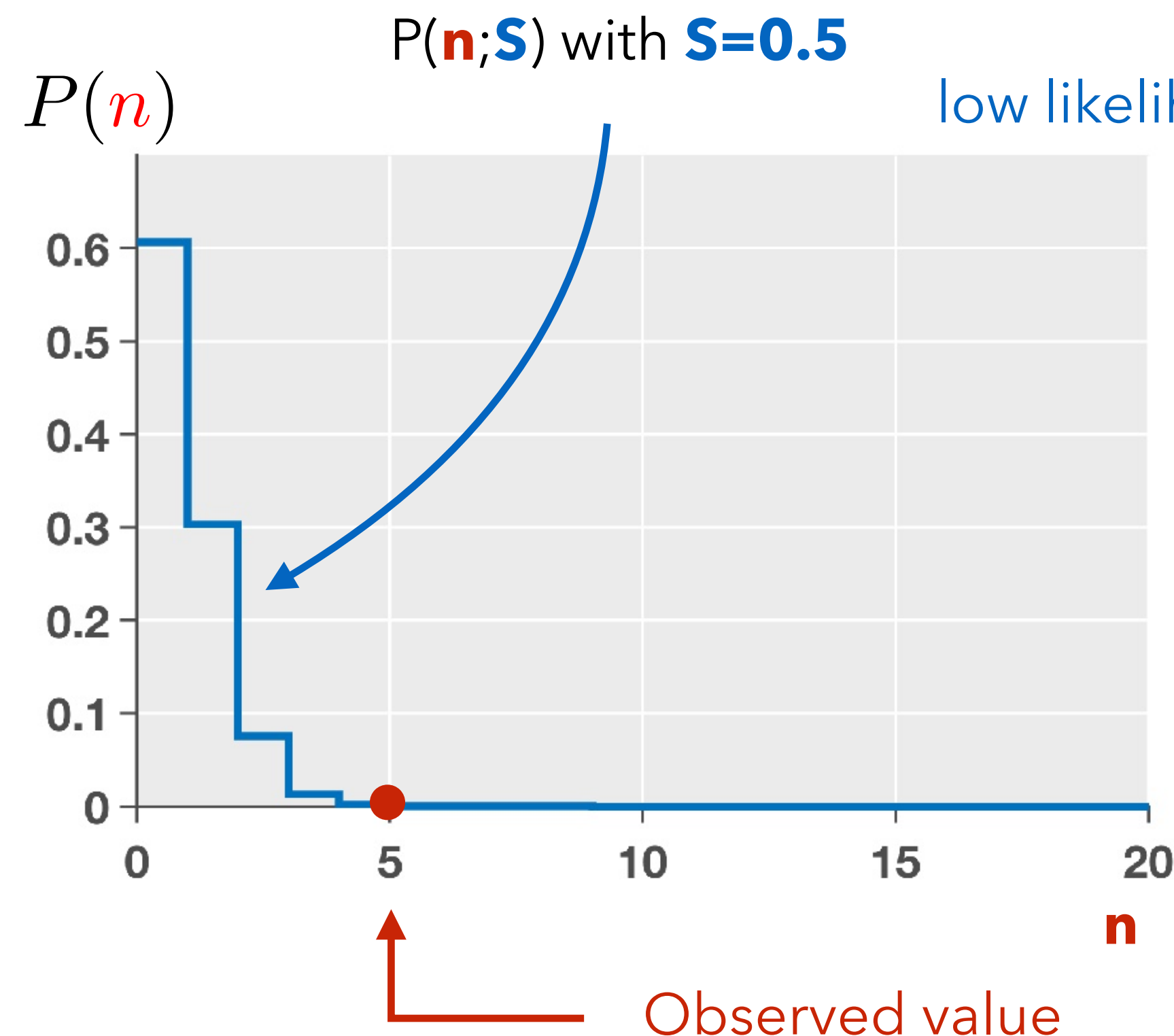
Likelihood - Poisson example

Assume Poisson distribution with $\mathbf{B=0}$ $P(\mathbf{n}; \mathbf{S}) = \frac{S^n}{n!} e^{-S}$

In a given experiment we observe $\mathbf{n=5}$ want to infer parameter $\mathbf{S}$ value

- Try different values of S for a fixed data value $n=5$
- Varying parameter, fixed data = Likelihood framework

$$\mathcal{L}(\mathbf{n} = \mathbf{5}; \mathbf{S}) = \frac{S^5}{5!} e^{-S}$$



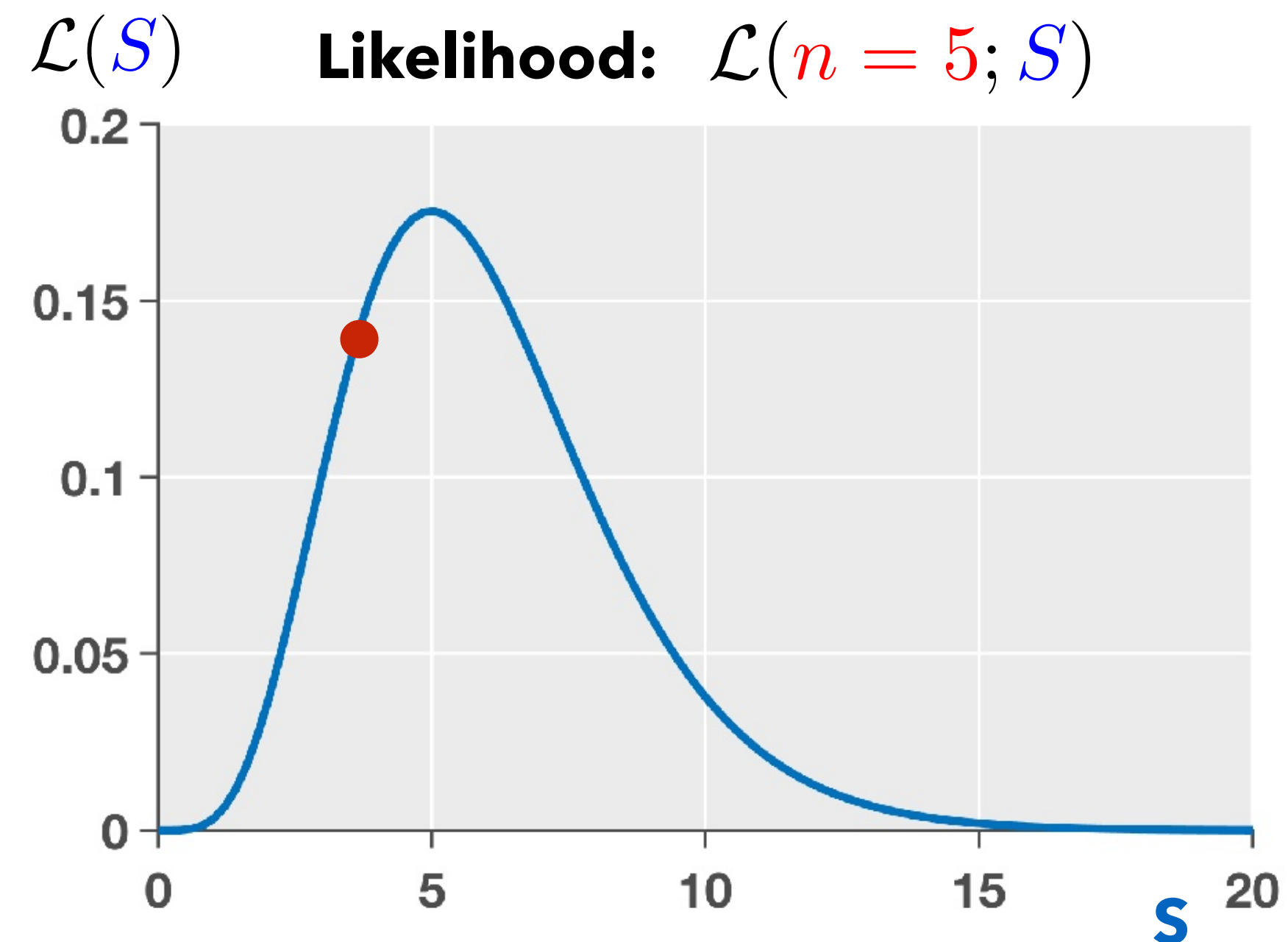
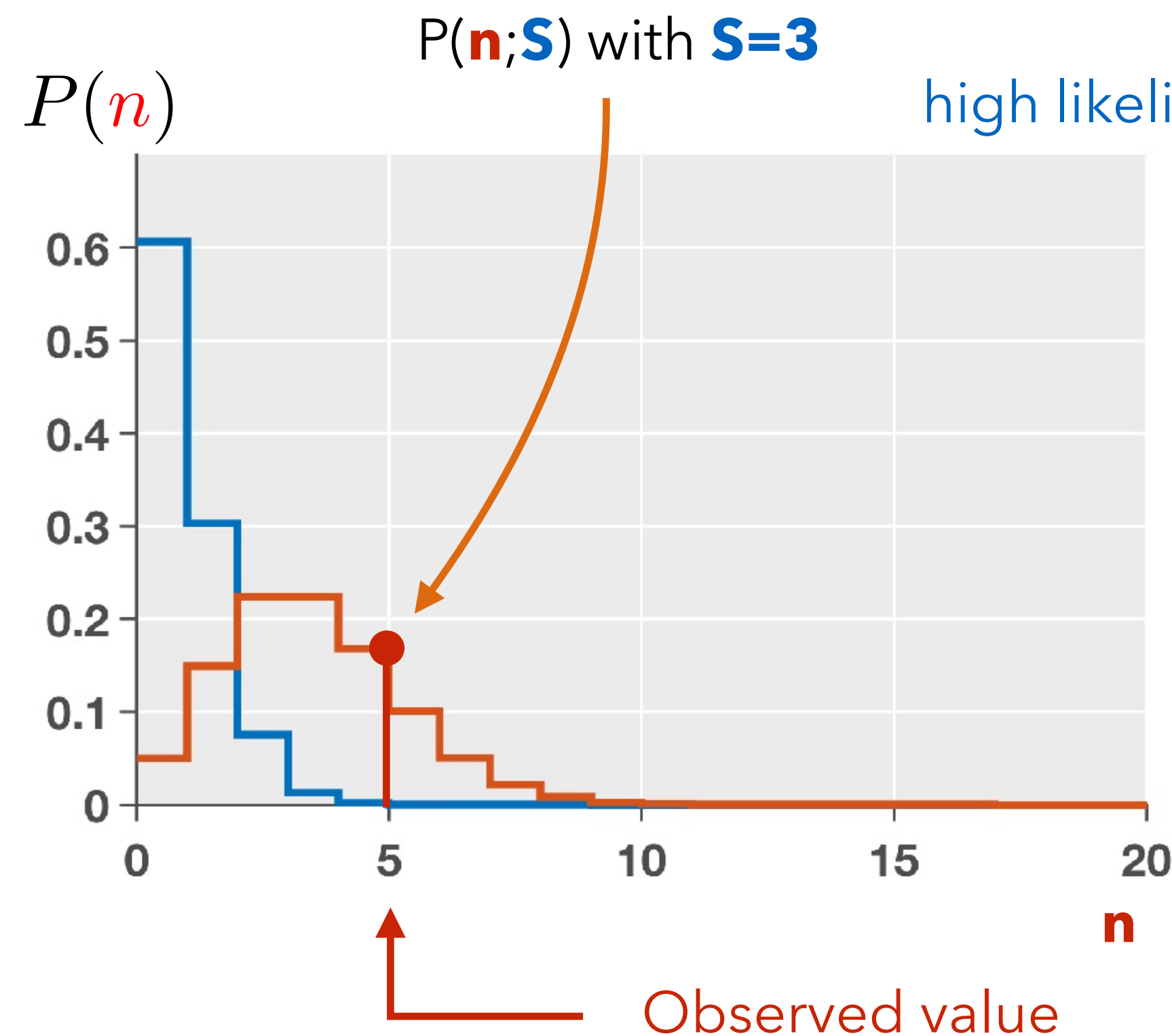
Likelihood - Poisson example

Assume Poisson distribution with $\mathbf{B=0}$ $P(\mathbf{n}; \mathbf{S}) = \frac{S^n}{n!} e^{-S}$

In a given experiment we observe $\mathbf{n=5}$ want to infer parameter $\mathbf{S}$ value

- Try different values of S for a fixed data value $n=5$
- Varying parameter, fixed data = Likelihood framework

$$\mathcal{L}(\mathbf{n} = \mathbf{5}; \mathbf{S}) = \frac{S^5}{5!} e^{-S}$$



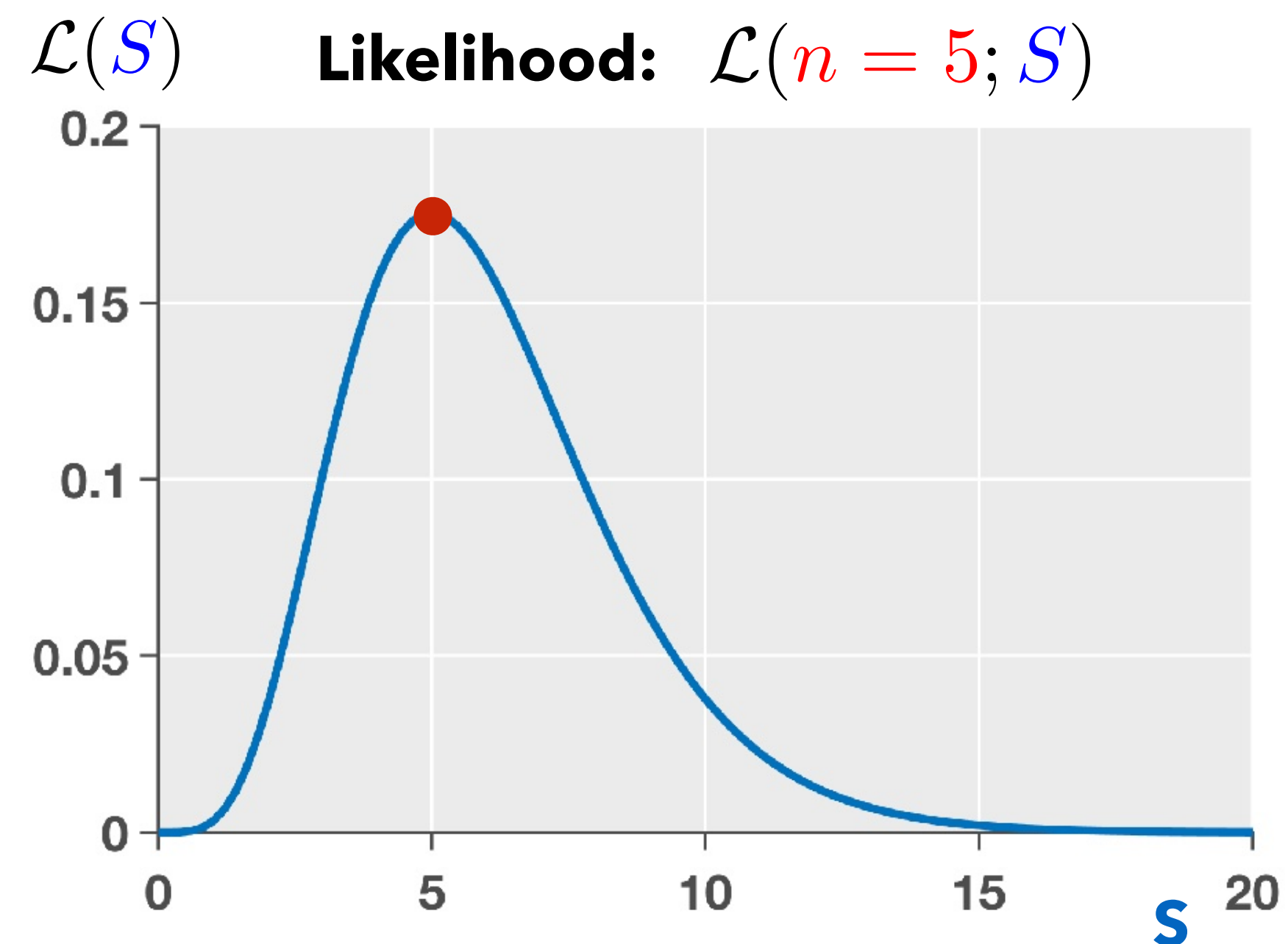
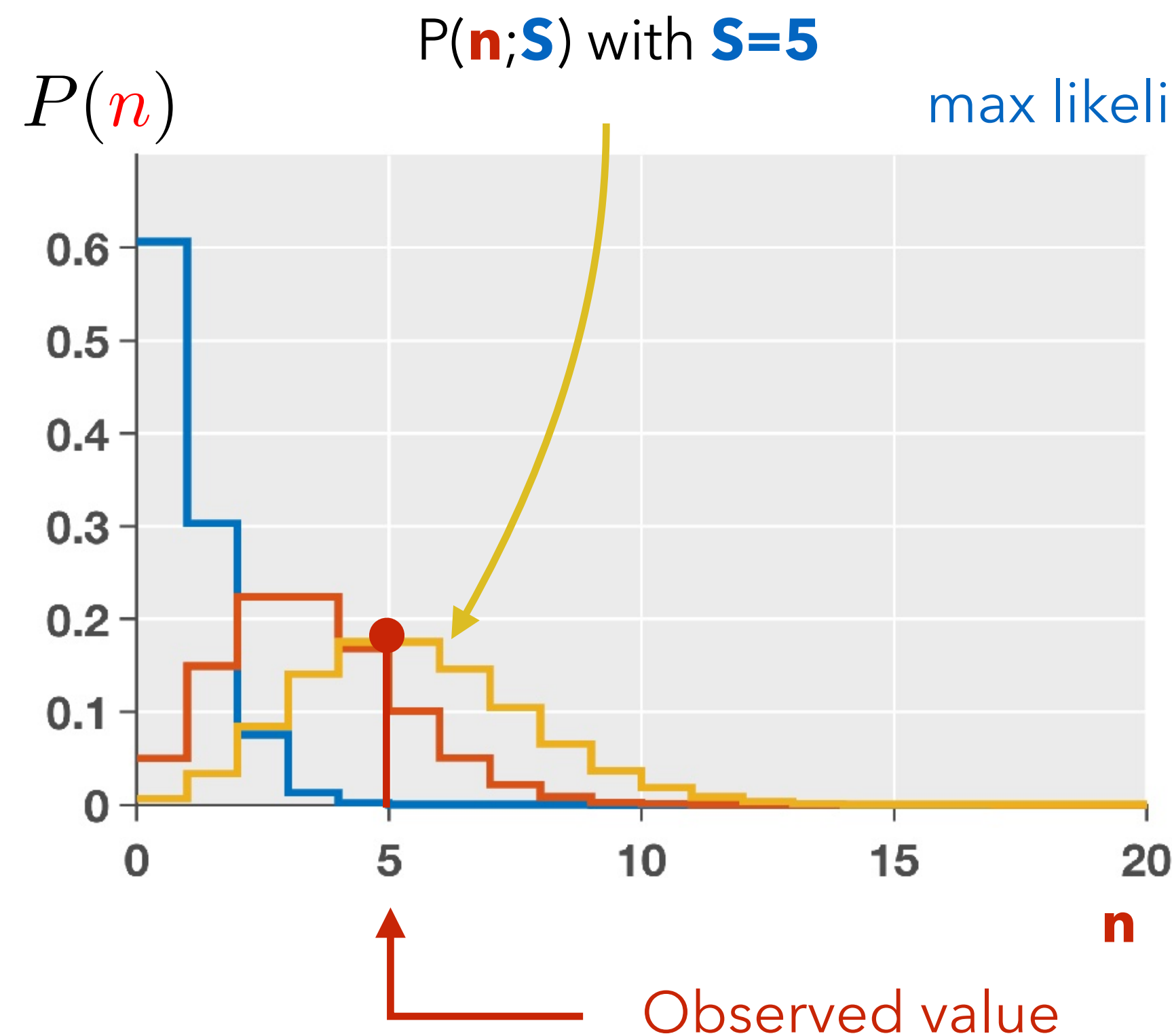
Likelihood - Poisson example

Assume Poisson distribution with $\mathbf{B=0}$ $P(\mathbf{n}; \mathbf{S}) = \frac{S^n}{n!} e^{-S}$

In a given experiment we observe $\mathbf{n=5}$ want to infer parameter $\mathbf{S}$ value

- Try different values of S for a fixed data value $n=5$
- Varying parameter, fixed data = Likelihood framework

$$\mathcal{L}(\mathbf{n} = \mathbf{5}; \mathbf{S}) = \frac{S^5}{5!} e^{-S}$$



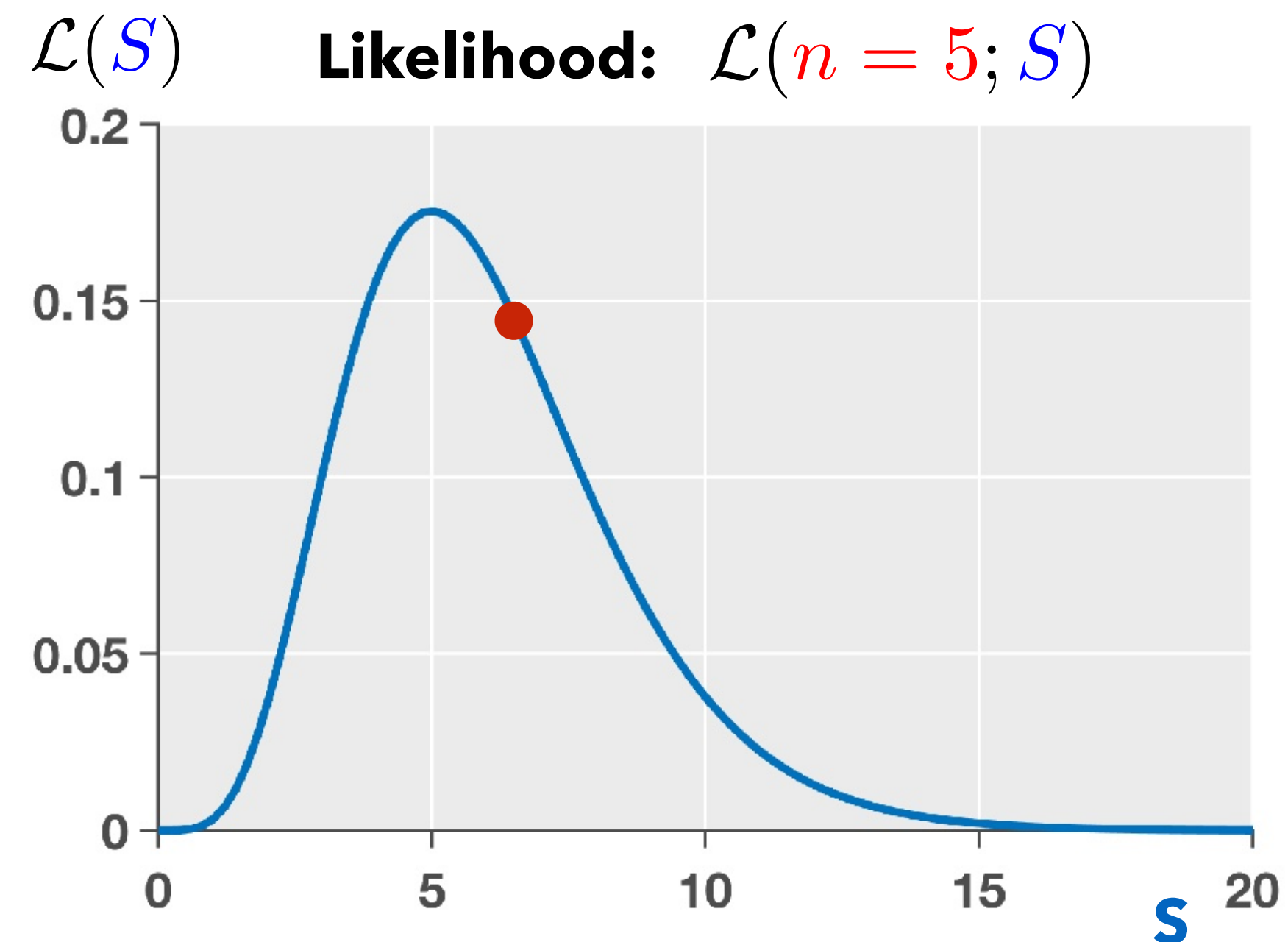
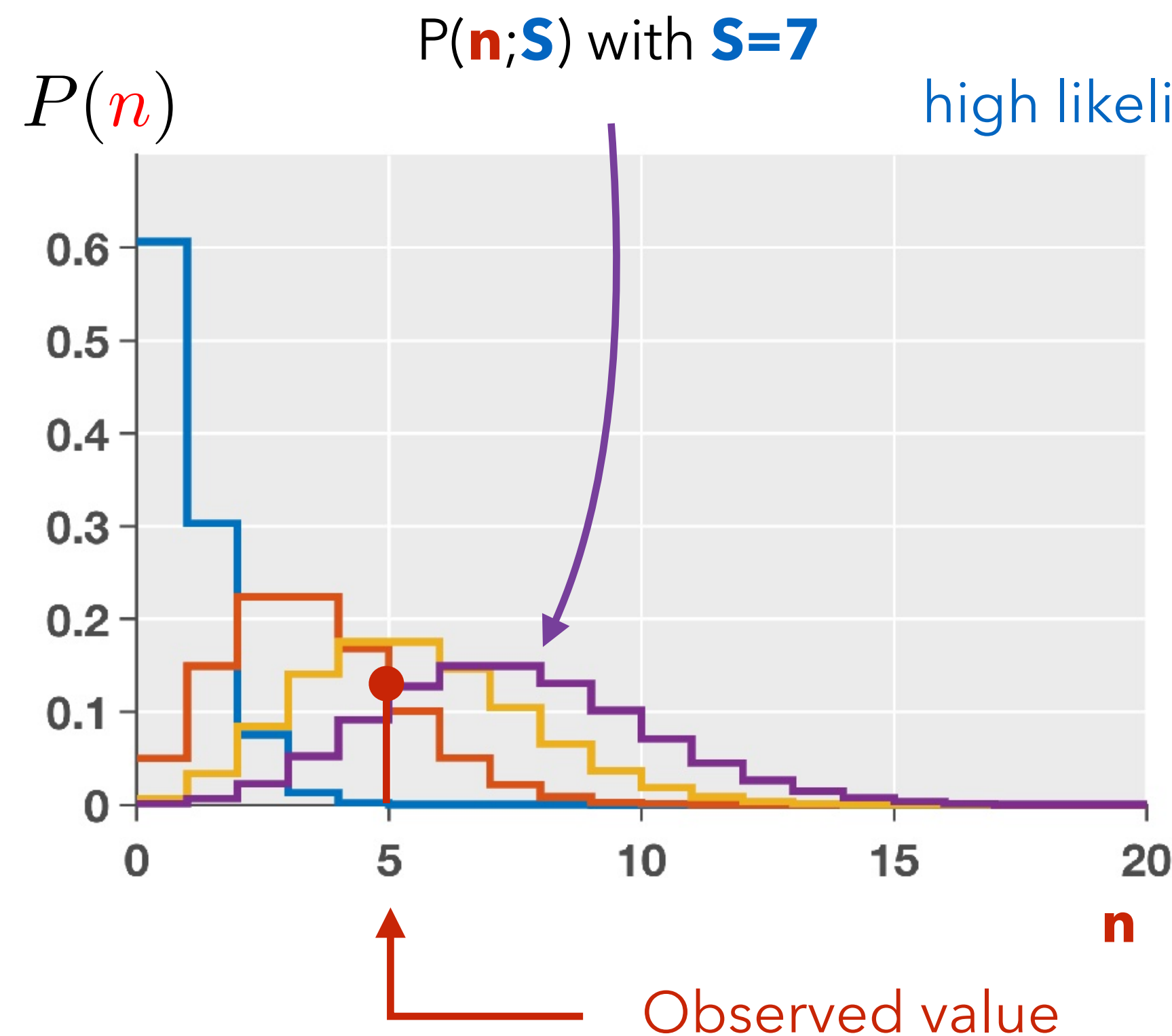
Likelihood - Poisson example

Assume Poisson distribution with $\mathbf{B=0}$ $P(\mathbf{n}; \mathbf{S}) = \frac{S^n}{n!} e^{-S}$

In a given experiment we observe $\mathbf{n=5}$ want to infer parameter $\mathbf{S}$ value

- Try different values of S for a fixed data value $n=5$
- Varying parameter, fixed data = Likelihood framework

$$\mathcal{L}(\mathbf{n} = \mathbf{5}; \mathbf{S}) = \frac{S^5}{5!} e^{-S}$$



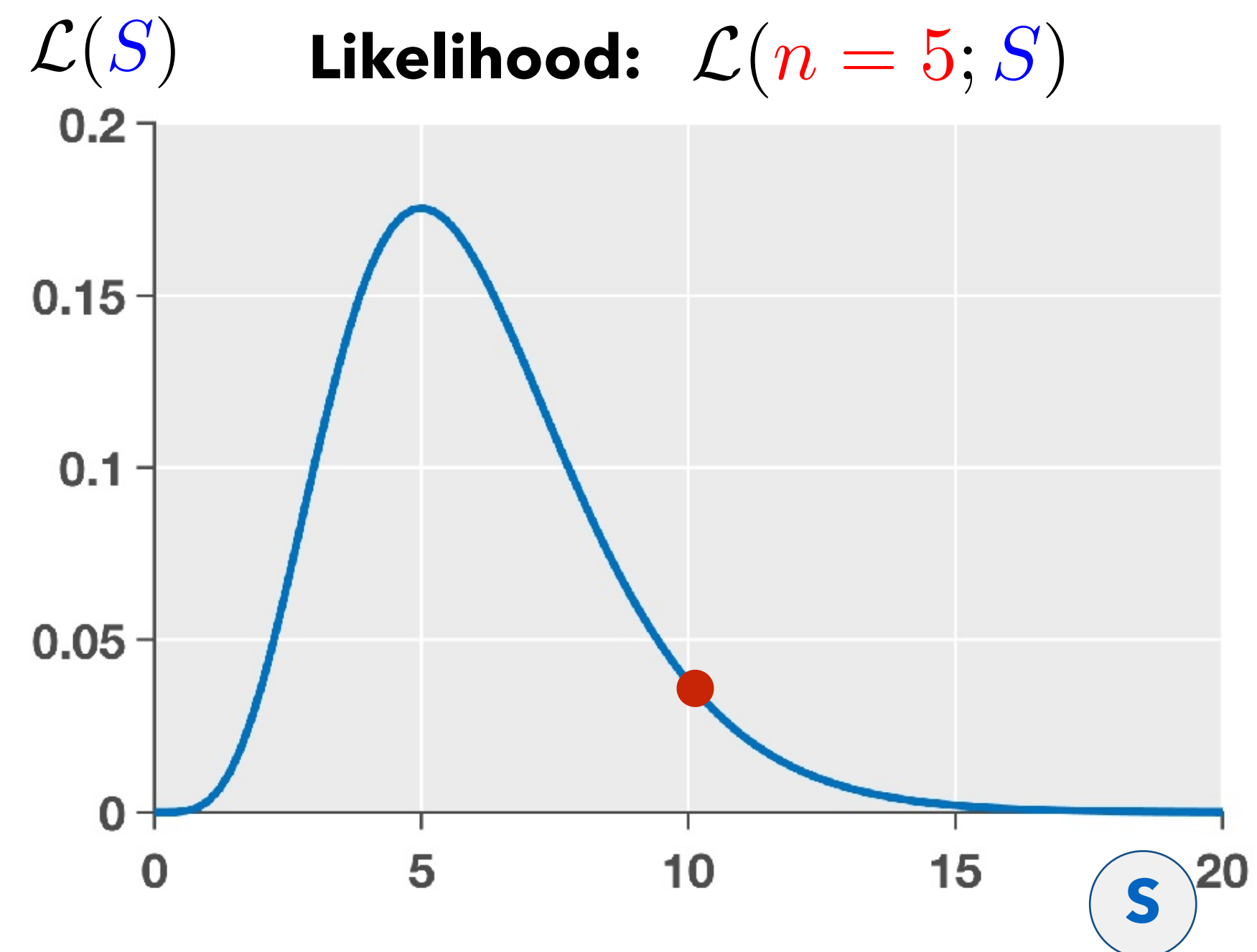
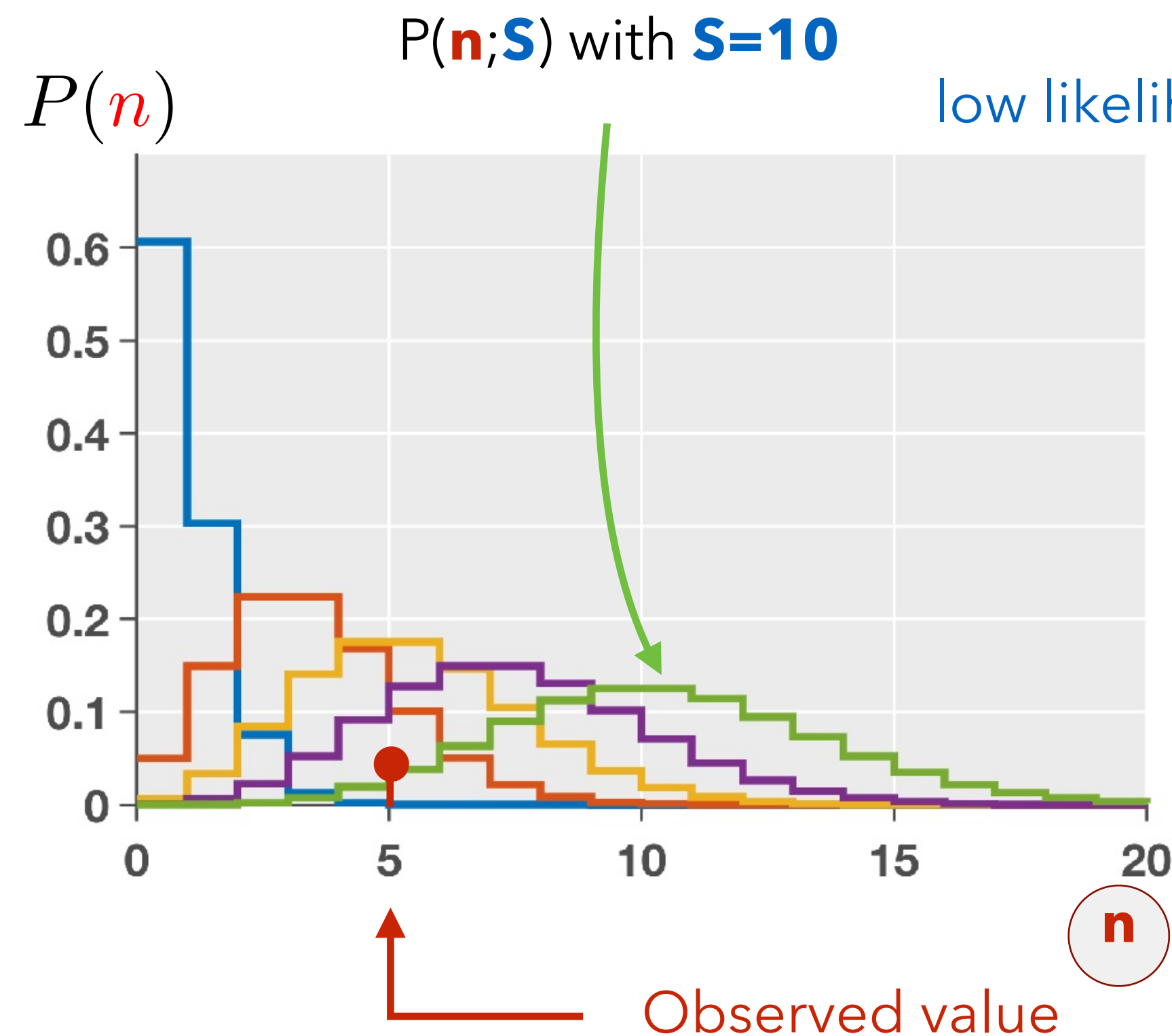
Likelihood - Poisson example

Assume Poisson distribution with $\mathbf{B=0}$ $P(\mathbf{n}; \mathbf{S}) = \frac{S^n}{n!} e^{-S}$

In a given experiment we observe $\mathbf{n=5}$ want to infer parameter $\mathbf{S}$ value

- Try different values of S for a fixed data value $n=5$
- Varying parameter, fixed data = Likelihood framework

$$\mathcal{L}(\mathbf{n} = \mathbf{5}; \mathbf{S}) = \frac{S^5}{5!} e^{-S}$$



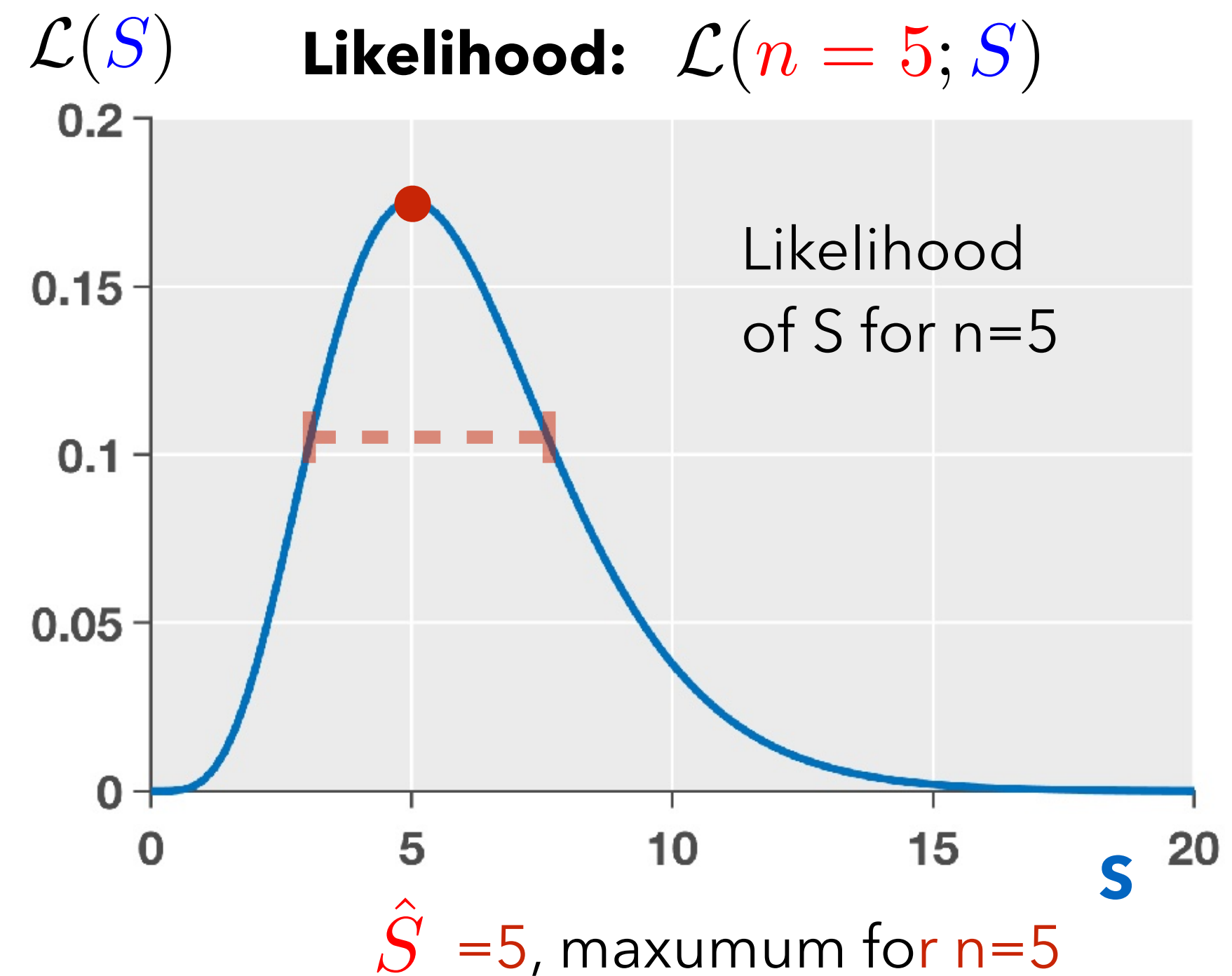
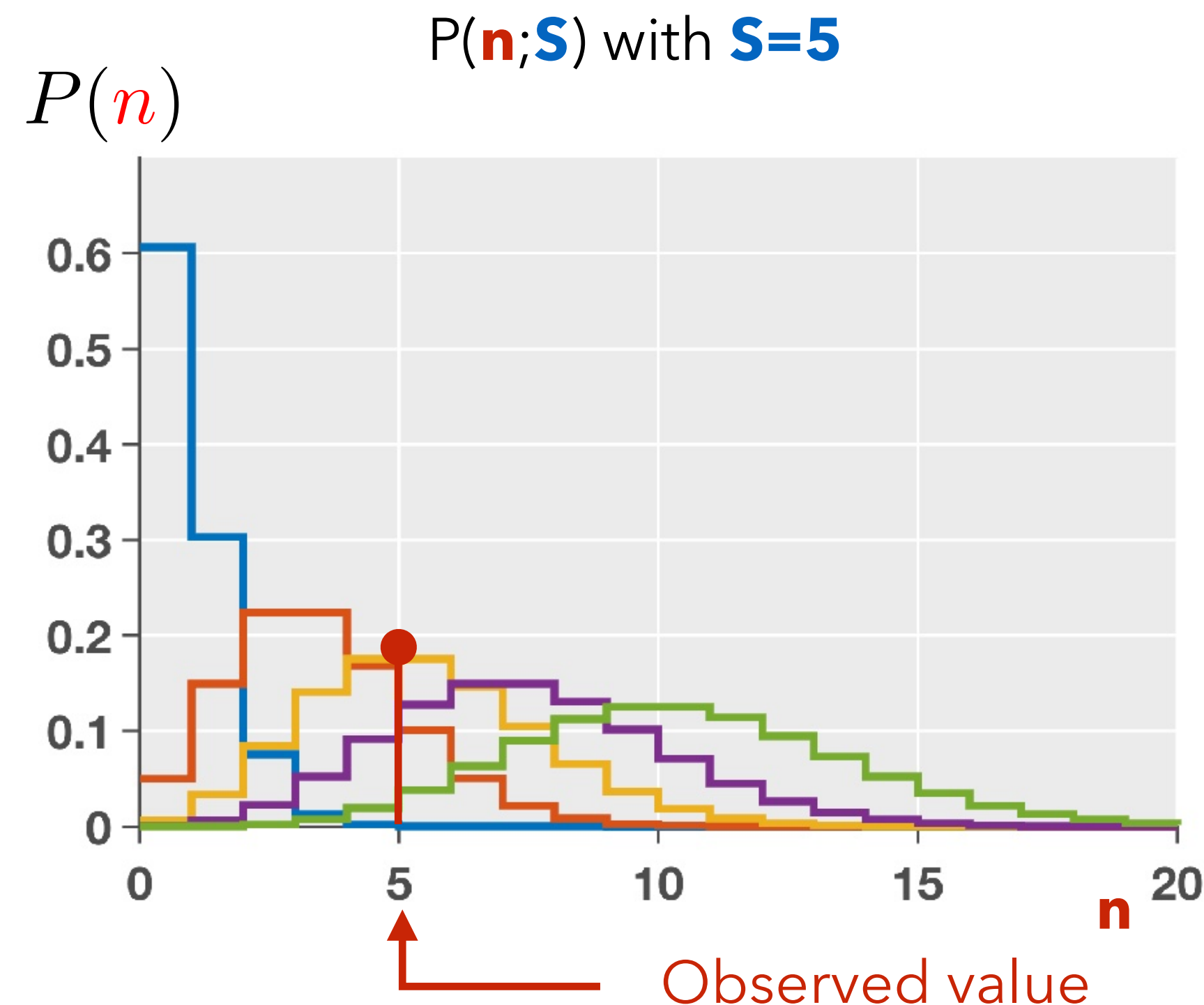
Maximum Likelihood Estimator (MLE)

Estimate a parameter μ = Find the value that maximizes $L(\mu)$

⇒ the value of μ for which this data was most likely to occur

⇒ **M**aximum **L**ikelihood **E**stimator

$$\hat{\mu} = \operatorname{argmax}_{\mu} \mathcal{L}(\mu)$$



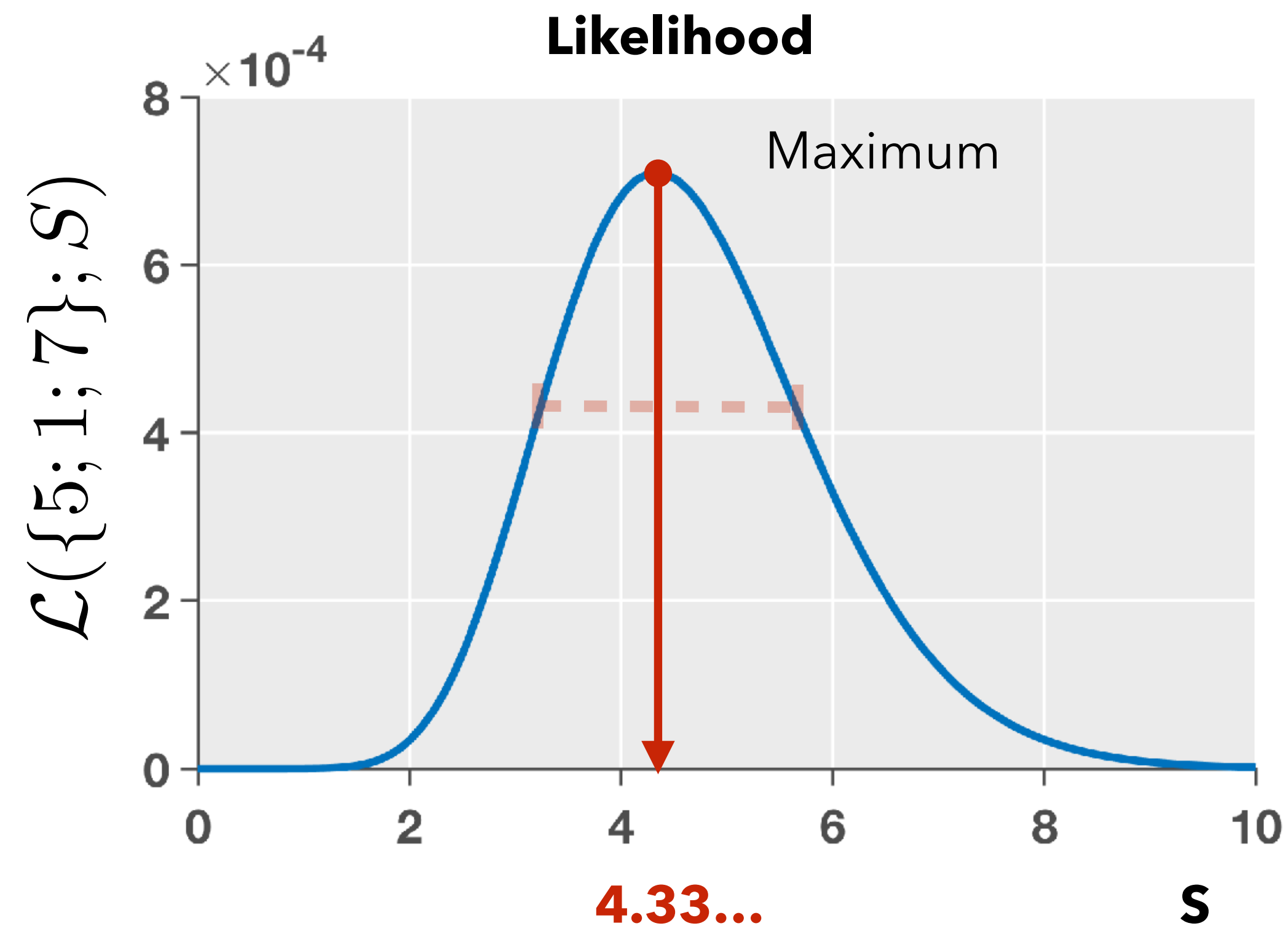
⇒ the **MLE** is a **function of the data** ⇒ it is itself an **observable**

⇒ no **guarantee** it is the true value (data may be "unlikely") but sensible estimate

Likelihood - Poisson example

with several measured values

Ex: 3 observed values : 5, 1, 7. $\mathcal{L}(\{5; 1; 7\}; S) = \frac{S^{13} e^{-3S}}{5! 1! 7!}$



S value when Likelihood is max: $\hat{S}$

Testing hypotheses

Hypotheses testing

Hypothesis: assumption on model (**parameters**), e.g. $H_0: S=0$

⇒ Goal = to determine if H_0 is **true** or **false** using a test based on data **= take a decision**

		Data output	
Possible outcomes		Data favors H_0	Data disfavors H_0
Truth	H_0 is true	 Nothing new = what we already know	 Error (Type I) ⇒ false discovery α p-value
	H_0 is false	 Error (Type II) ⇒ missed discovery $1-\beta$	 Discovery new effect, new signal $\neq 0$ decision

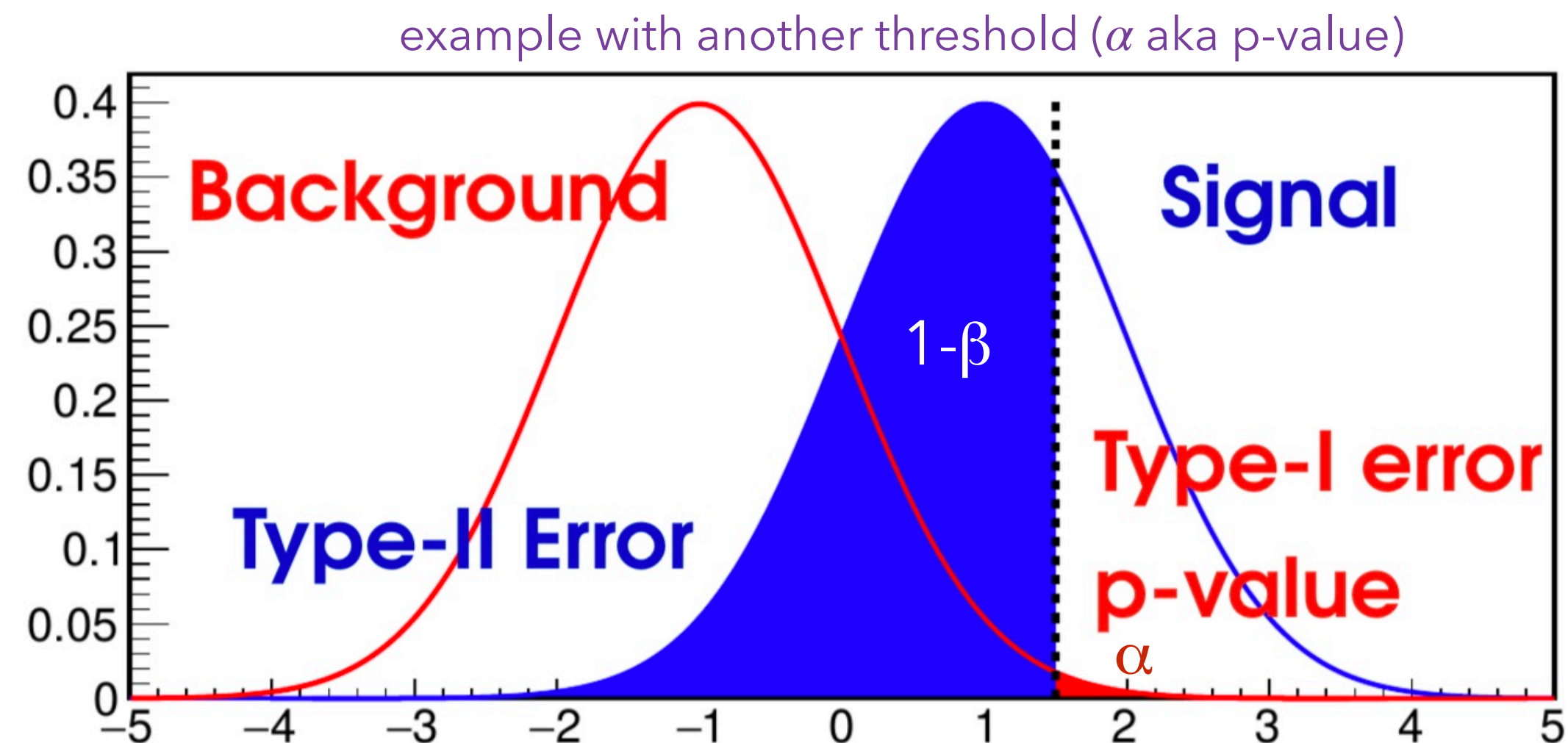
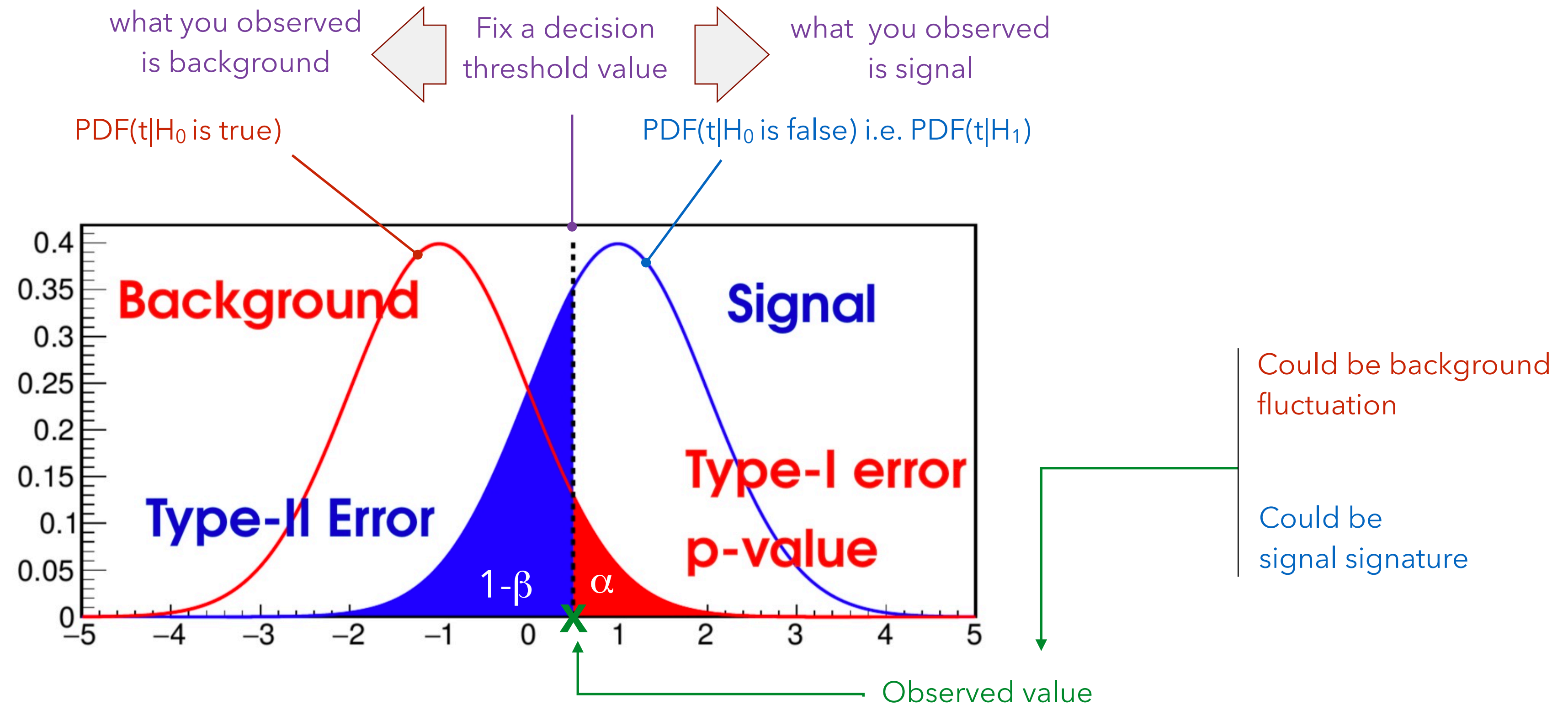
To make a discovery, you have to **prove** the null hypothesis, H_0 is false

⇒ You want to minimize Type I error (not to claim a discovery when it's false)

You do not want to publish a result you have to retract afterward... embarrassing...

However by **reducing Type I error**, there is a price to pay: it **increases the Type II error**

Before looking at data:
 1. Define a test statistics t
 2. Fix a decision threshold



Want to fix stringent discovery criteria

However:

lower Type-I errors, higher Type II errors

Find right balance

many tests possible...

Goal: Find test that minimizes **Type II error** for a given level of **Type I error**, fixed in advance

Hypotheses testing with Likelihood

Neyman-Pearson Lemma

When comparing two hypotheses H_0 and H_1 , the optimal discriminator is the **Likelihood ratio** (LR)

As for MLE, choose the hypothesis that is more likely for the data.

- **Minimizes Type-II uncertainties** for given level of Type-I errors
- Always need an **alternate hypothesis** (H_1) to test against.

Caveat: Strictly true only for simple hypotheses (no free parameters)

- In the following: all tests based on LR, focussing on p-values (Type-I errors), trusting that Type-II errors are anyway as small as they can be...

Note: finding better criteria is specific to the problem at hand

👉 need work to find some criteria better than the LR in composite hypotheses (converse of simple hypothesis: free parameters seen as composition of simple hypotheses)

$$LR = \frac{\mathcal{L}(\text{data}; H_0)}{\mathcal{L}(\text{data}; H_1)}$$

$$\Delta\chi^2 = -2 \ln \left(\frac{\mathcal{L}(\text{data}; H_0)}{\mathcal{L}(\text{data}; H_1)} \right)$$

Statistical result as hypothesis test

It is usual in particle/astroparticle physics to recast results in terms of **hypothesis testing**: $\Delta\chi^2 = -2 \ln \left(\frac{\mathcal{L}(\text{data}; \mathbf{H}_0)}{\mathcal{L}(\text{data}; \mathbf{H}_1)} \right)$

null hypothesis: $\left| \begin{array}{l} \mathbf{H}_0 \text{ different possibilities} \\ \text{following what we want} \end{array} \right.$ alternate hypothesis: $\left| \begin{array}{l} \mathbf{H}_1 \text{ is chosen as} \\ \text{the } \mathbf{best\ fit} \text{ fo the data} \end{array} \right.$

- **Discovery**: are the data compatible with background-only?
 - $\mathbf{H}_0$: only background is present
 - How well can we **reject** $\mathbf{H}_0$? → p-value (significance)
- **Upper limits**: no excess observed – how small must the signal be?
 - $\mathbf{H}_0(S)$: B + some signal S
 - How small can we make S, and still reject $\mathbf{H}_0(S)$ at 95% C.L. (p=5%) ?
C.L. = confidence level p-value
- **Parameter measurement**:
 - $\mathbf{H}_0(\mu)$: some parameter value μ
 - Which values μ are not rejected at 68% C.L. (p=32%) ?
 - ⇒ 1σ confidence interval on μ

In all cases, $\mathbf{H}_0$ is the **null hypothesis** – what we are trying to disprove

From test statistic to decision in astroparticle physics

Test statistics (Profile Likelihood Ratio) for a signal strength parameter (i.e. in a model $\mu S_i + B_i$ to explain observed data y_i):

$$q_0 = -2 \log \left(\frac{\mathcal{L}(\mu = 0, \hat{\theta}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \right)$$

Note: The notation $\hat{\theta}$ appears in some HEP references. It is equivalent to $\hat{\theta}_\mu$: the value of θ that maximises the likelihood for a given fixed μ .

- Asymptotic approximation (Wilk's theorem): significance is $Z \simeq \sqrt{q_0}$
(valid for large N and when the true value is not on a physical boundary)
- HEP/astroparticle convention: discovery if $Z \geq 5\sigma$; exclusion at 95% CL ($p=5\%$, i.e. nearly 2σ)
- Caution - when to calibrate with toys Monte Carlo (i.e. compute the distribution of q_0 without relying on asymptotic approximations):
 - Low-count regime (Poisson with small n)
 - Signal strength μ constrained near 0 (boundary effect)

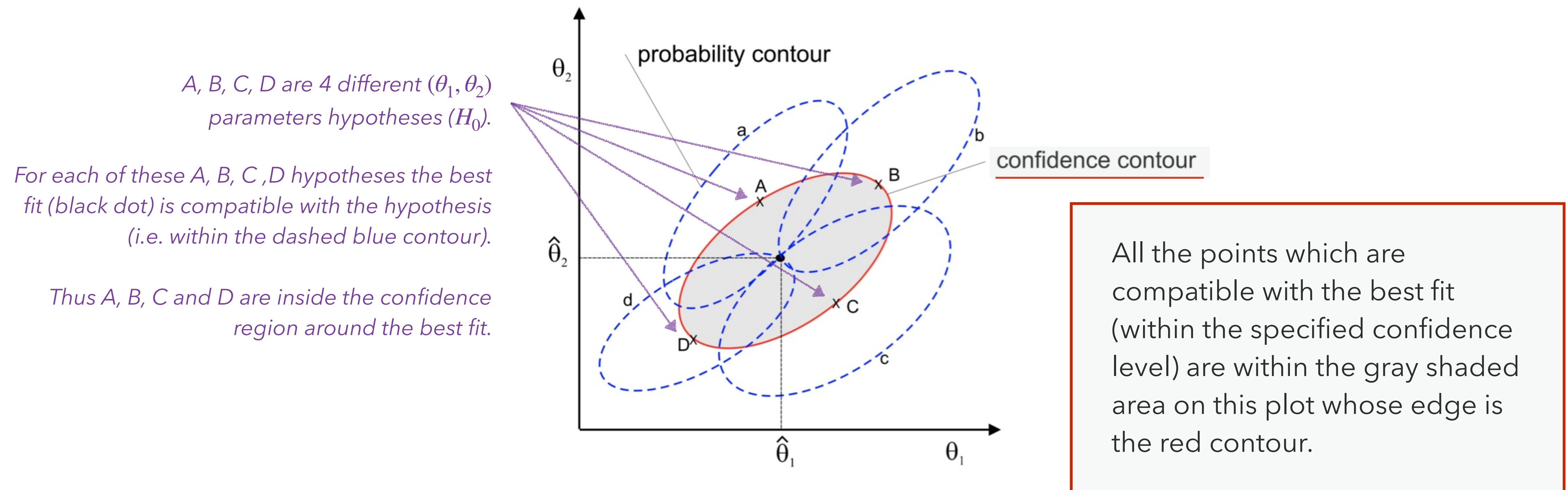
Use pseudo-experiments (toys Monte Carlo) under H_0 to get the true distribution of q_0 and use it to convert your observed q_0 into a calibrated p-value.

- Practical link: The Li & Ma (1983) formula gives Z directly for On-Off analyses (γ -ray, neutrino); implemented in ctools, Gammapy, etc.

Example of confidence region construction

First a confidence level is fixed. Say 90% CL.

Test every point in the plane (θ_1, θ_2) .



Take home concepts

Statistical inference is a vast topic. We focused on frequentist approach in this lecture. For a good introduction to Bayesian approach check reference [7] (next slide).

We reviewed basic probability, which are fundamental for statistical inference

We discussed two major ways to estimate parameters: the least squares and the maximum likelihood. We presented the approximate likelihood and χ^2 intervals, and goodness of fit to validate fitted model.

When addressing hypothesis testing, we also illustrated how to build confidence regions.

In astroparticle physics: always validate asymptotic approximations (e.g. Wilk's theorem) with Monte Carlo pseudo-experiments (toys MC) before quoting a significance σ in a publication.

Final advice: when you explore statistical questions,

1. simplify your problem to take the essence
2. simulate with a Monte Carlo to check your understanding.

References



[1] F. James, "Statistical methods in experimental physics", World Scientific

[2] G. Cowan, "Statistical Data Analysis", Oxford Science Publication

[3] L. Lyons, "Statistics for Nuclear and Particle Physicists", Cambridge University Press

[4] R. J. Barlow, "A guide to the use of statistical methods in the physical science", Wiley

[5] L. Lista, "Statistical Methods for Data Analysis in Particle Physics", Springer

[6] O. Behnke et al., "Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods", Wiley

[7] G. d'Agostini, "Bayesian Reasoning in Data Analysis": A Critical Introduction, World Scientific Publishing

[8] PDG Review, "Statistics" chapter, available online at pdg.lbl.gov (updated annually; essential reference for particle/astroparticle physics)

[9] B. Khélifi et al., "Analysis of the H.E.S.S. publica data release with ctools" – practical appendices illustrating On-Off likelihood, Li & Ma significance, residualls (see also Gammapy and ctools documentation)

Thank you!