

Data Science 2

Chris Mathys



Master's Degree Programme in Cognitive Science

Spring 2023

A surprising piece of information



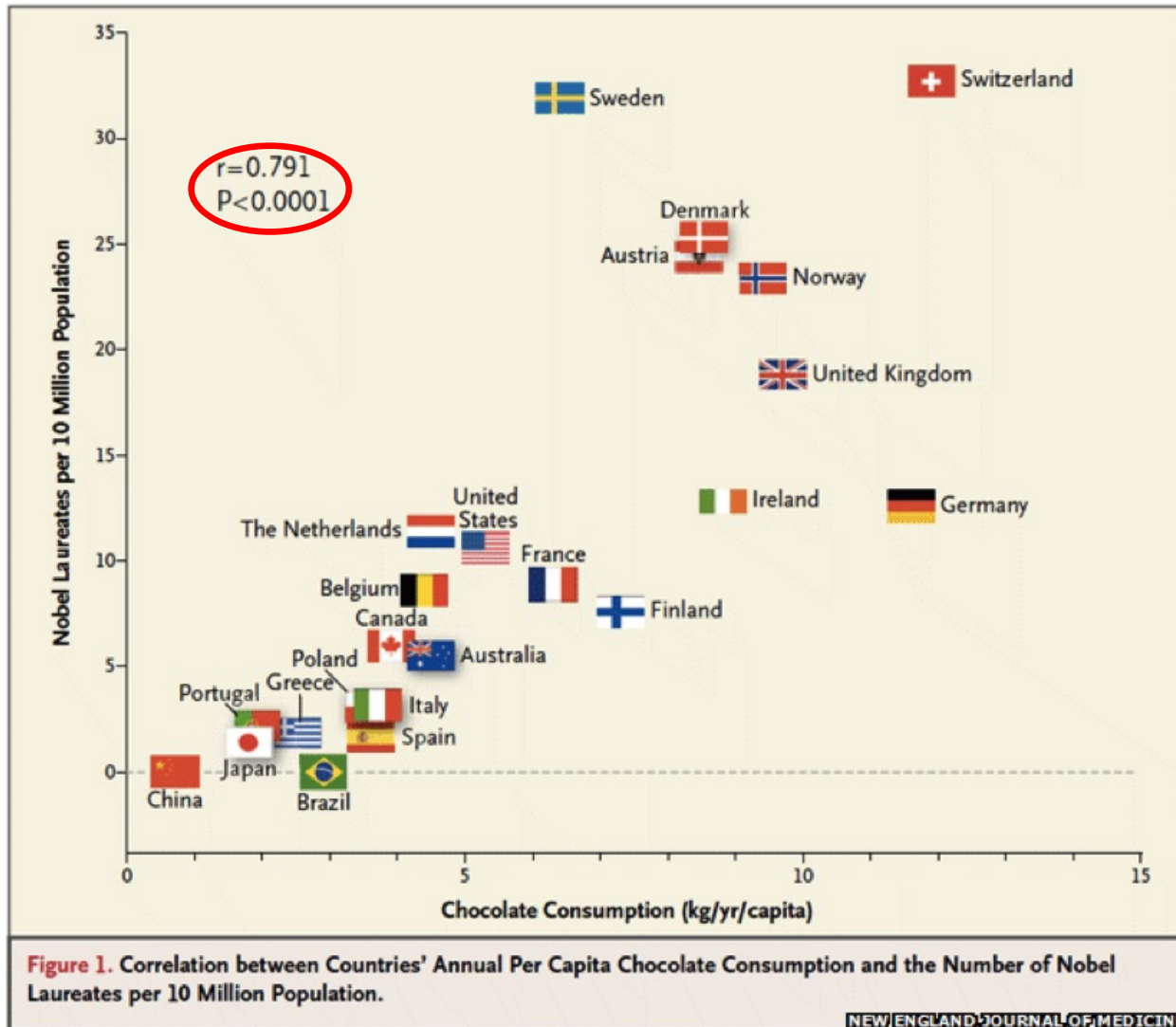
Does chocolate make you clever?

By Charlotte Pritchard
BBC News

Eating more chocolate improves a nation's chances of producing Nobel Prize winners - or at least that's what a recent study appears to suggest. But how much chocolate do Nobel laureates eat, and how could any such link be explained?

A surprising piece of information

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates.
New England Journal of Medicine, 367(16), 1562–1564.



So will I win the Nobel prize if I eat lots of chocolate?

This is a question referring to **uncertain quantities**. Like almost all scientific questions, it cannot be answered by deductive logic. *Nonetheless, quantitative answers can be given – but they can only be given in terms of probabilities.*

Our question here can be rephrased in terms of a conditional probability:

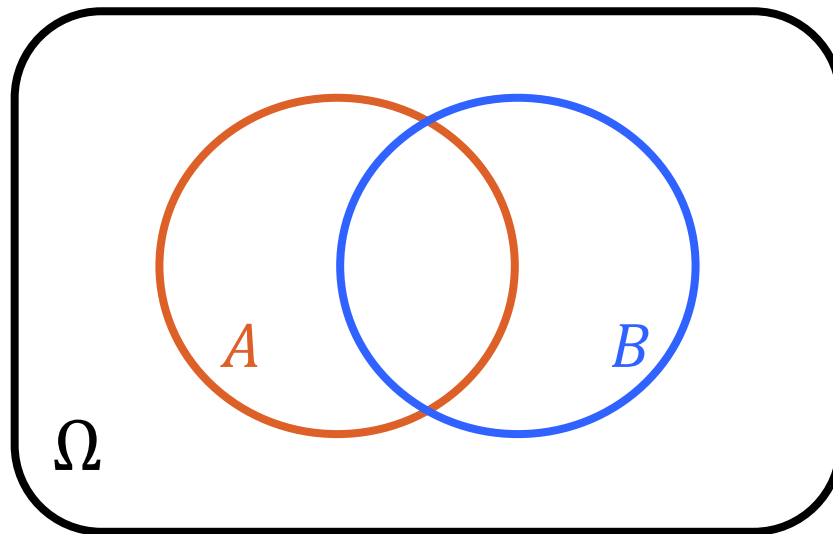
$$p(\text{Nobel} \mid \text{amount of chocolate}) = ?$$

To answer it, we have to learn to calculate such quantities. The tool for this is **Bayesian inference**.

However: note that no amount of statistical analysis will tell you anything about the causal mechanism behind this if you don't have **a hypothesis about that mechanism and a causal scientific model of it!**

Calculating with probabilities: the setup

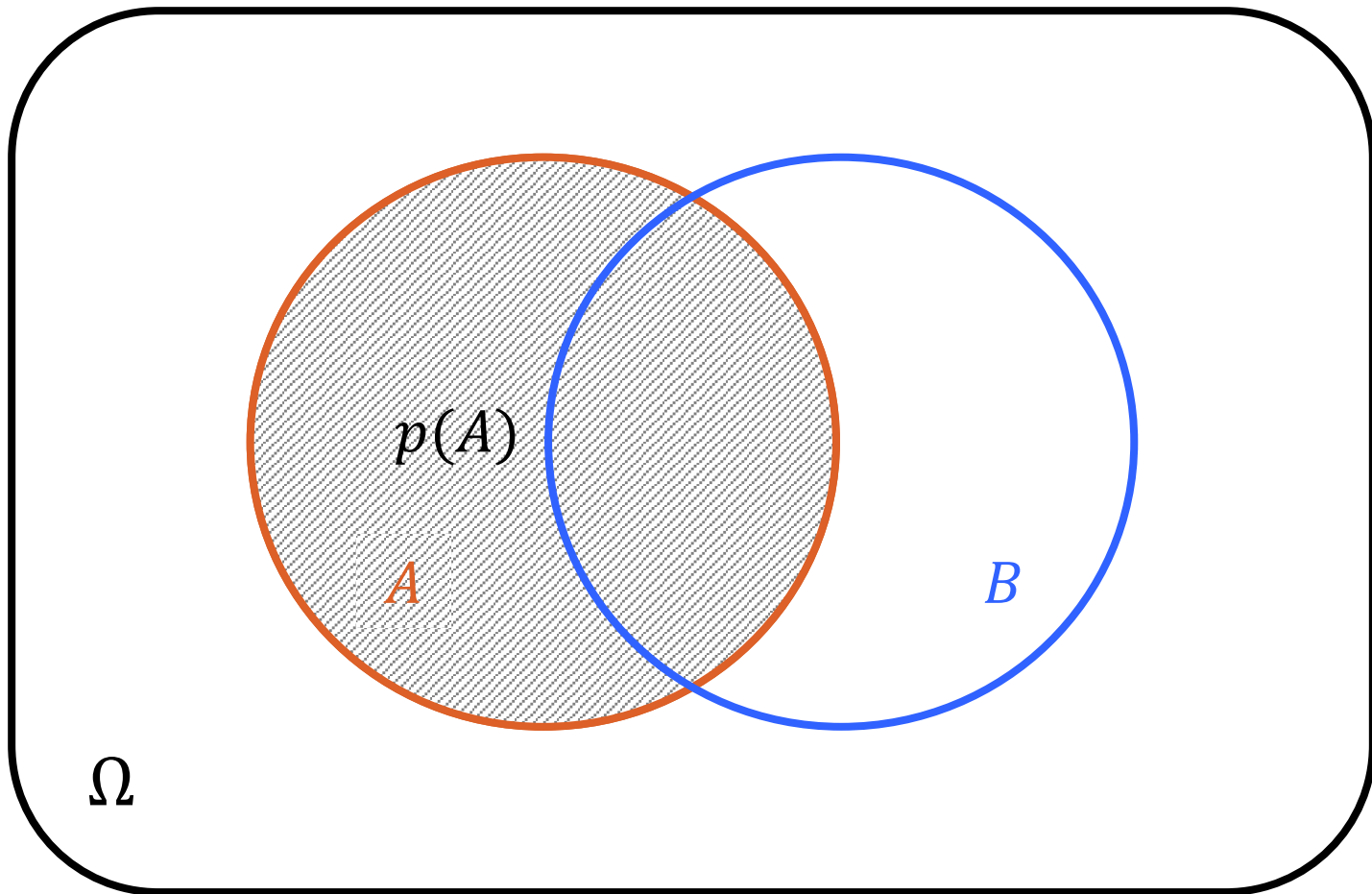
We assume a probability space Ω with subsets A and B



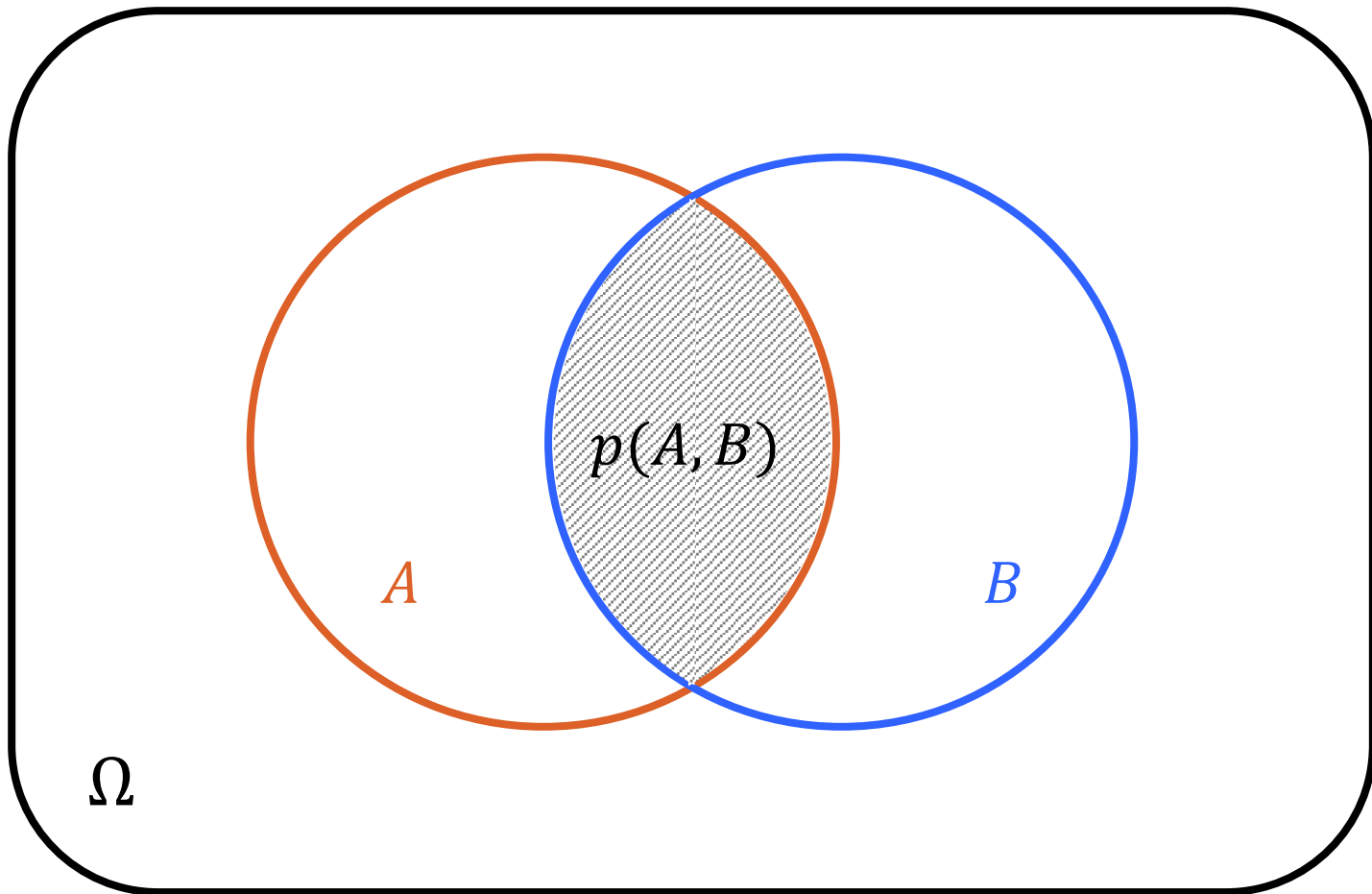
In order to understand *the rules of probability*, we need to understand **three kinds of probabilities**

- *Marginal* probabilities like $p(A)$
- *Joint* probabilities like $p(A, B)$
- *Conditional* probabilities like $p(B|A)$

Marginal probabilities



Joint probabilities



What is ‘marginal’ about marginal probabilities?

- Let A be the statement ‘the sun is shining’
- Let B be the statement ‘it is raining’
- \bar{A} negates A , \bar{B} negates B

Consider the following table of joint probabilities:

	B	\bar{B}	Marginal probabilities
A	$p(A, B) = 0.1$	$p(A, \bar{B}) = 0.5$	$p(A) = 0.6$
\bar{A}	$p(\bar{A}, B) = 0.2$	$p(\bar{A}, \bar{B}) = 0.2$	$p(\bar{A}) = 0.4$
Marginal probabilities	$p(B) = 0.3$	$p(\bar{B}) = 0.7$	Sum of all probabilities $\sum p(\cdot, \cdot) = 1$

Marginal probabilities get their name from being at the margins of tables such as this one.

Conditional probabilities

- In the previous example, what is the probability that the sun is shining given that it is not raining?
- This question refers to a conditional probability: $p(A|\bar{B})$
- You can find the answer by asking yourself: out of all times where it is not raining, which proportion of times will the sun be shining?

	B	\bar{B}	Marginal probabilities
A	$p(A, B) = 0.1$	$p(A, \bar{B}) = 0.5$	$p(A) = 0.6$
\bar{A}	$p(\bar{A}, B) = 0.2$	$p(\bar{A}, \bar{B}) = 0.2$	$p(\bar{A}) = 0.4$
Marginal probabilities	$p(B) = 0.3$	$p(\bar{B}) = 0.7$	Sum of all probabilities $\sum p(\cdot, \cdot) = 1$

- This means we have to divide the joint probability of ‘sun shining, not raining’ by the sum of all joint probabilities where it is not raining:

$$p(A|\bar{B}) = \frac{p(A, \bar{B})}{p(A, \bar{B}) + p(\bar{A}, \bar{B})} = \frac{p(A, \bar{B})}{p(\bar{B})} = \frac{0.5}{0.7} \approx 0.71$$

The rules of probability

Considerations like the ones above led to the following definition of the **rules of probability**:

1. $\sum_a p(a) = 1$ (*Normalization*)
2. $p(B) = \sum_a p(a, B)$ (*Marginalization* – the **sum rule**)
3. $p(A, B) = p(A|B)p(B) = p(B|A)p(A)$ (*Conditioning* – the **product rule**)

These are **axioms**, ie they are assumed to be true. Therefore, we cannot test them the way we could test a theory. However, we can see if they turn out to be useful.

Bayes' rule

- The product rule of probability states that

$$p(A|B)p(B) = p(B|A)p(A)$$

- If we divide by $p(B)$, we get **Bayes' rule**:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{\sum_a p(B|a)p(a)}$$

- The last equality comes from unpacking $p(B)$ according to the product and sum rules:

$$p(B) = \sum_a p(B, a) = \sum_a p(B|a)p(a)$$

Bayes' rule: what problem does it solve?

- Why is Bayes' rule important?
- It allows us to invert conditional probabilities, ie to pass from $p(B|A)$ to $p(A|B)$:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- In other words, it allows us to update our belief about A in light of observation B

Bayes' rule: the chocolate example

In our example, it is immediately clear that $P(\text{Nobel}|\text{chocolate})$ is very different from $P(\text{chocolate}|\text{Nobel})$. While the first is hopeless to determine directly, the second is much easier to find out: ask Nobel laureates how much chocolate they eat. Once we know that, we can use Bayes' rule:

The diagram illustrates Bayes' rule with the following components and annotations:

- posterior** (green oval): $p(\text{Nobel}|\text{chocolate})$
- evidence** (blue oval): $p(\text{chocolate})$
- likelihood** (red oval): $p(\text{chocolate}|\text{Nobel})$
- model** (yellow oval): $P(\text{Nobel})$
- prior** (green oval): $P(\text{Nobel})$

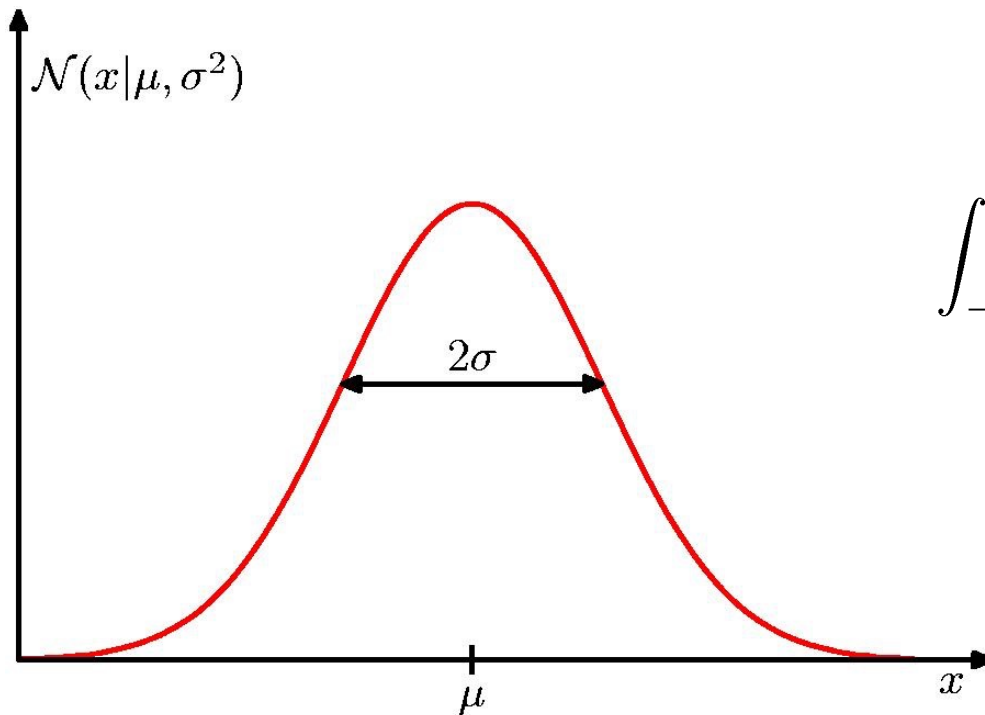
The equation is written as:

$$p(\text{Nobel}|\text{chocolate}) = \frac{p(\text{chocolate}|\text{Nobel})P(\text{Nobel})}{p(\text{chocolate})}$$

However: note that no amount of statistical analysis will tell you anything about the causal mechanism behind this if you don't have **a hypothesis about that mechanism and a causal scientific model of it!**

The Gaussian Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1$$

Gaussian Mean and Variance

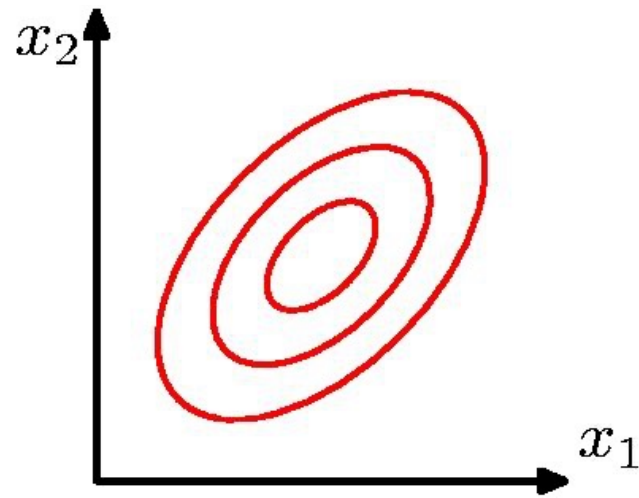
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2$$

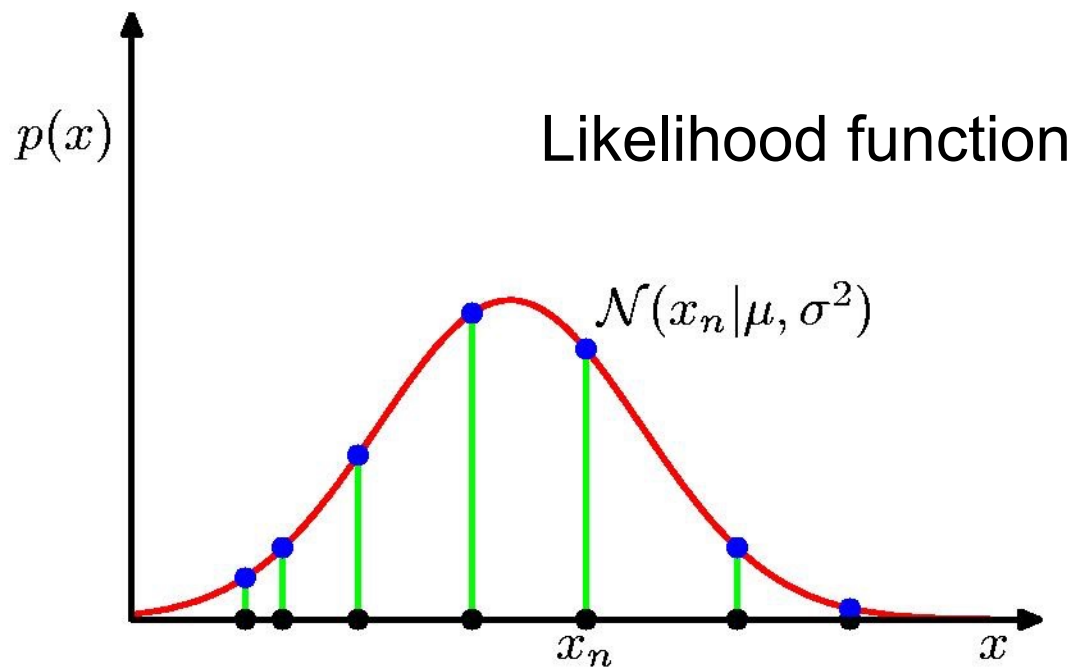
$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$



Gaussian Parameter Estimation



$$p(\mathbf{x} | \mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n | \mu, \sigma^2)$$

Maximum (Log) Likelihood

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

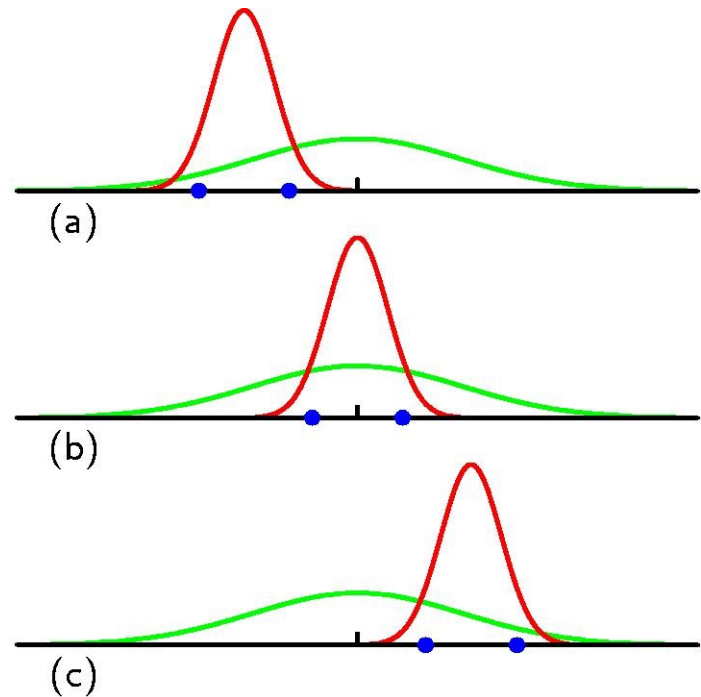
$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \qquad \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

Properties of μ_{ML} and σ_{ML}^2

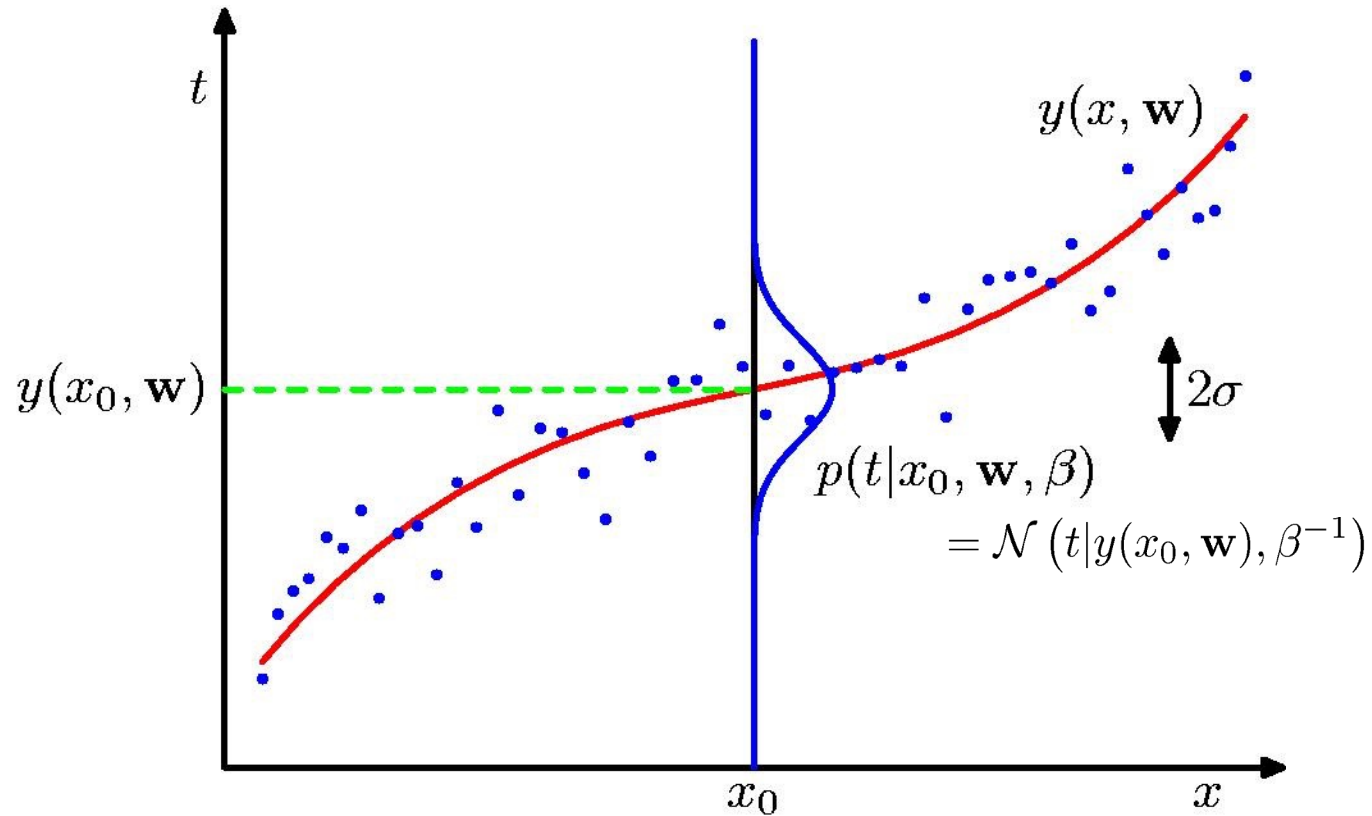
$$\mathbb{E}[\mu_{\text{ML}}] = \mu$$

$$\mathbb{E}[\sigma_{\text{ML}}^2] = \left(\frac{N-1}{N}\right) \sigma^2$$

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{N}{N-1} \sigma_{\text{ML}}^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2\end{aligned}$$



Curve Fitting Re-visited



Maximum Likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = - \underbrace{\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

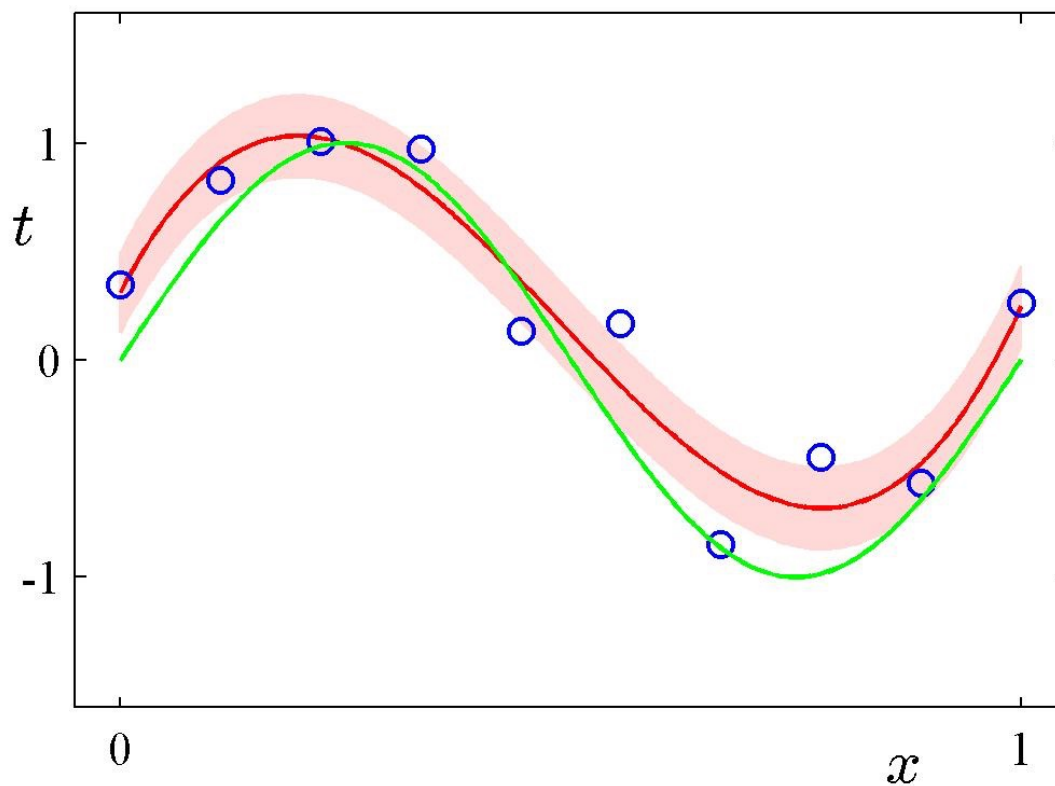
Determine \mathbf{w}_{ML} by minimizing sum-of-squares error $E(\mathbf{w})$

.

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2$$

Predictive Distribution

$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$



MAP: A Step towards Bayes

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$\beta\tilde{E}(\mathbf{w}) = \frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Determine \mathbf{w}_{MAP} by minimizing regularized sum-of-squares $\tilde{E}(\mathbf{w})$.

Bayesian Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w} = \mathcal{N}(t|m(x), s^2(x))$$

$$m(x) = \beta \phi(x)^T \mathbf{S} \sum_{n=1}^N \phi(x_n) t_n \qquad s^2(x) = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(x_n) \phi(x_n)^T \qquad \phi(x_n) = (x_n^0, \dots, x_n^M)^T$$

Bayesian Predictive Distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$

