Machine Learning 2020

Volker Roth

Department of Mathematics & Computer Science
University of Basel

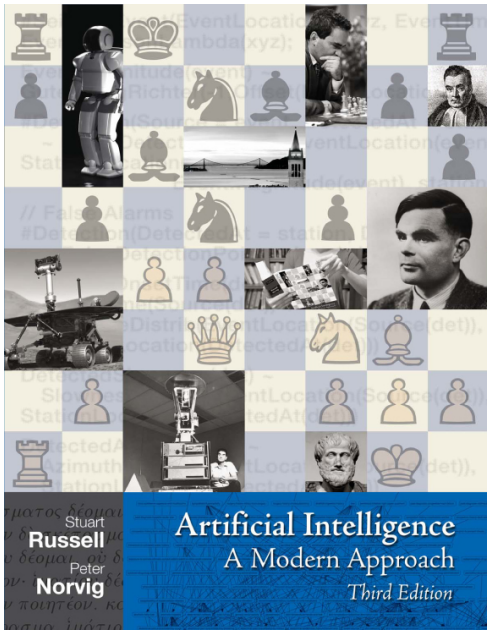
26th February 2020

Section 1

Probabilities

Outline

- **Probability basics**
- **Some important probability distributions**
- **Some important statistical concepts**



Probability theory vs Statistics

Definition (Probability Theory)

A branch of mathematics concerned with the analysis of random phenomena.

General $\Rightarrow$ Specific

Definition (Statistics)

The science of collecting, analyzing, presenting, and interpreting data.

Specific $\Rightarrow$ General

- Machine learning is closely related to (inferential) statistics.
- Learning algorithms are often probabilistic algorithms.

Probabilities

Definition (Probability Space)

A probability space is the triple

$$(\Omega, S, P)$$

where

- Ω is the sample/outcome space, $\omega \in \Omega$ is a **sample point/atomic event**.
  Example: 6 possible rolls of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- S is a collection of **events** to which we are willing to assign probabilities. An **event** $a \in S$ is any subset of Ω, e.g., die roll < 4: $a = \{1, 2, 3\}$
- P is a mapping from events in S to $\mathbb{R}$ that satisfies the probability axioms.

Axioms of Probability

1. $P(a) \geq 0 \; \forall a \in S$: probabilities are not negative,
2. $P(\Omega) = 1$: "trivial" event has maximal possible prob 1,
3. $a, b \in S$ and $a \cap b = \{\} \Rightarrow P(a \cup b) = P(a) + P(b)$: probability of two mutually disjoint events is the sum of their probabilities.
   Example:
   $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2.$

Random Variables

Definition (Random Variable)

A **random variable** X is a function from the sample points to some range, e.g., the reals

$$X : \Omega \to \mathbb{R},$$

or booleans $X : \Omega \to \{\text{true,false}\}$.

Real random variables are characterized by their distribution function.

Definition (Cumulative Distribution Function)

Let $X : \Omega \to \mathbb{R}$ be a real valued random variable. We define

$$F_X(x) = P(X \leq x).$$

This is the probability of the event $\{\omega \in \Omega : X(\omega) \leq x\}$

Probability and Propositions

- **Proposition**: event (set of sample points) where the **proposition is true.**
- Given Boolean random variables A and B:
  - event a = set of atomic events where $A(\omega) = \text{true}$
  - event $\neg a$ = set of atomic events where $A(\omega) = \text{false}$
  - event $a \wedge b$ = atomic events where $A(\omega) = \text{true}$ and $B(\omega) = \text{true}$
- With Boolean variables,
  event = propositional logic model
  e.g., $A = \text{true}$, $B = \text{false}$, or $a \wedge \neg b$.
- **Proposition = disjunction of events in which it is true**
  e.g., $(a \vee b) = (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$
  $\Longrightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Syntax for Propositions

- **Boolean** random variables
  e.g., *Cavity* (do I have a cavity?)
  $Cavity = true$ is a proposition, also written *cavity*
- **Discrete** random variables (*finite* or *infinite*)
  e.g., *Weather* is one of $\langle sunny, rain, cloudy, snow \rangle$
  $Weather = rain$ is a proposition
  Values must be exhaustive and mutually exclusive
- **Continuous** random variables (*bounded* or *unbounded*)
  e.g., $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.

Probability distribution

- **Unconditional probabilities** of propositions
  e.g., $P(Weather = sunny) = 0.72$. **Bayesian interpretation:** correspond to belief prior to arrival of any (new) evidence
- **Probability distribution** gives values for all possible assignments:
  $P(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (**normalized**, sums to 1)
- **Joint probability distribution** for a set of RVs gives the probability of every atomic event on those RVs (i.e., every sample point)
  $P(Weather, Cavity) =$ a 4×2 matrix of values:

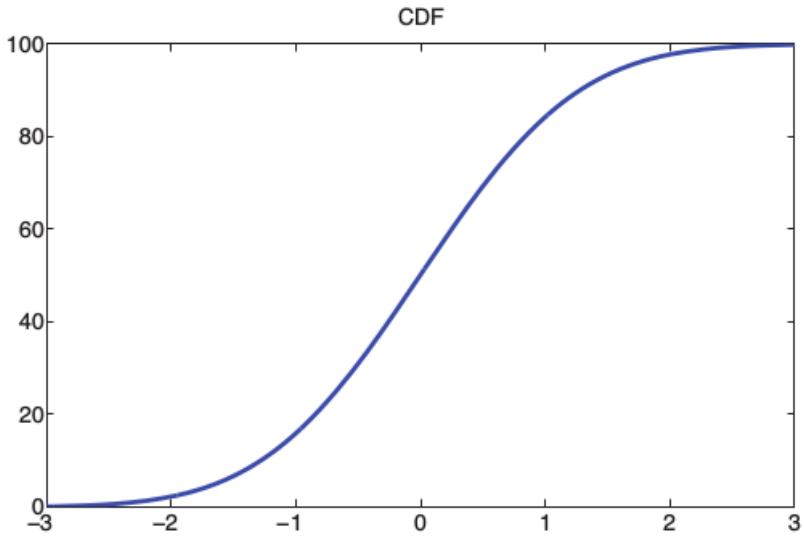
$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

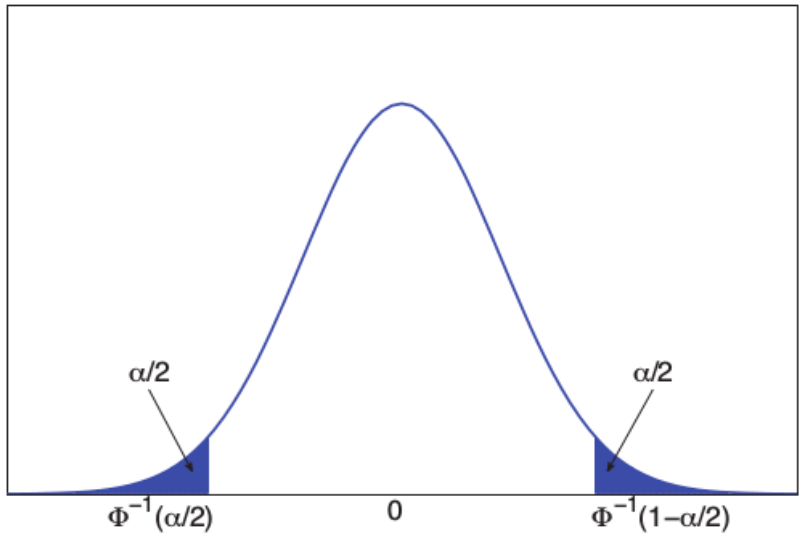
Probability for continuous variables

Suppose X describes some uncertain continuous quantity. What is the probability that $a < X \leq b$?

- define events $A = (X \leq a), B = (X \leq b), W = (a < X \leq b)$.
- $B = A \vee W$, A and W are **mutually exclusive** $\rightsquigarrow p(B) = p(A) + p(W) \rightsquigarrow p(W) = p(B) - p(A)$.
- Define the **cumulative distribution function (cdf)** as $F(q) := p(X \leq q) \rightsquigarrow p(a < X \leq b) = F(b) - F(a)$.
- Assume that F is absolutely continuous: define **probability density function (pdf)** $p(x) := \frac{d}{dx} F(x)$.
- Given a pdf, the probability of a continuous variable being **in a finite interval** is: $P(a < X \leq b) = \int_a^b p(x)\, dx$.
- As the size of the interval gets smaller, we can write $P(x < X \leq x + dx) \approx p(x)\, dx$.
- We require $p(x) \geq 0$, but it is possible for $p(x) > 1$, so long as the density integrates to 1.

Probability for continuous variables





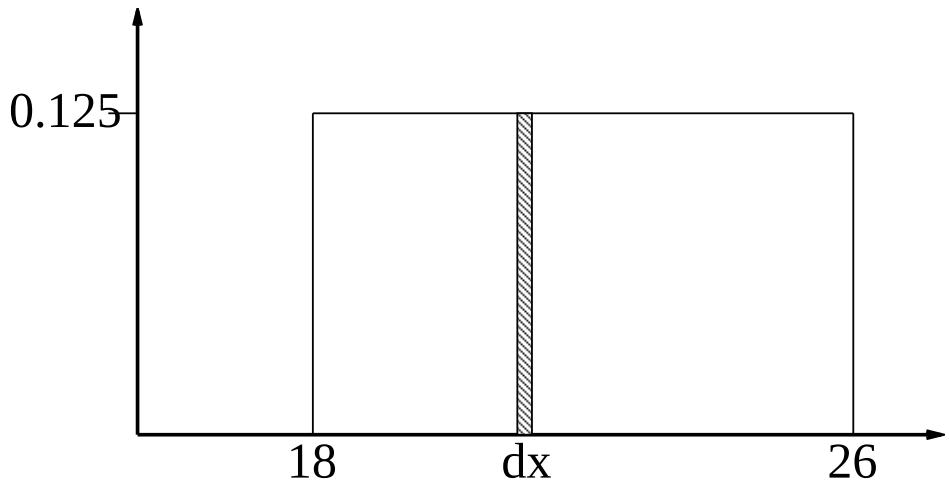
Left: cdf for the standard normal, $\mathcal{N}(0,1)$. Right: corresponding pdf.

- Shaded regions each contain $\alpha/2$ of the probability mass $\rightsquigarrow$ nonshaded region contains $1-\alpha$.
- Left cutoff point is $\Phi^{-1}(\alpha/2)$, Φ is cdf of standard Gaussian.
- By symmetry, the right cutoff point is $\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2)$.
- If $\alpha = 0.05$, the central interval is 95%, left cutoff is -1.96, right cutoff is 1.96.

Probability for continuous variables

Example: uniform distribution:

$$\text{Unif}(a, b) = \frac{1}{b-a}\mathbb{I}(a \leq x \leq b).$$



$p(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Mean and Variance

- Most familiar property of a distribution: **mean**, or **expected value**, denoted by μ.
- Discrete RVs:

$$E[X] = \sum_{x \in \mathcal{X}} x p(x),$$

- Continuous RVs:

$$E[X] = \int_{\mathcal{X}} x p(x)\, dx.$$

If this integral is not finite, the mean is not defined.

- The **variance** is a measure of the **spread** of a distribution:

$$\mathrm{var}[X] = E[(X - \mu)^2] = E[X^2] - \mu^2 =: \sigma^2.$$

- The square root $\sqrt{\mathrm{var}[X]}$ is the **standard deviation**.

Common discrete distributions: Binomial and Bernoulli

- Toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads.
- If the probability of heads is θ, then we say the RV X has a **binomial distribution**, $X \sim \text{Bin}(n, \theta)$:

$$\text{Bin}(k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

- Special case for $n = 1$: **Bernoulli distribution.**
  Let $X \in \{0, 1\} \rightsquigarrow$ binary random variable. Let θ be the probability of **success**. We write $X \sim \text{Ber}(\theta)$.

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)} (1-\theta)^{\mathbb{I}(x=0)}.$$

  In other words,

$$\text{Ber}(x|\theta) = \begin{cases} \theta, & \text{if } x = 1 \\ 1-\theta, & \text{if } x = 0. \end{cases}$$

Common discrete distributions: Multinomial

- Tossing a K-sided die $\rightsquigarrow$ can use the **multinomial distribution**.
- Let $\boldsymbol{X} = (X_1, X_2, \ldots X_K)$ be a **random vector**. Let x_j be the number of times side j of the die occurs.

$$\text{Mu}(\boldsymbol{x}|n, \boldsymbol{\theta}) = \binom{n}{x_1 \cdots x_K} \prod_{j=1}^{K} \theta_j^{x_j},$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \cdots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

is the **multinomial coefficient** (the number of ways to divide a set of size $n = \sum_{k=1}^{K} x_k$ into subsets with sizes x_1 up to x_K).

Common discrete distributions: Multinoulli

- Special case for $n = 1$: **Mutinoulli distribution.**
- Rolling a K-sided dice once, so $\boldsymbol{x}$ will be a vector of 0s and 1s, in which only one bit can be turned on.
- If the dice shows up as face k, then the k'th bit will be on
  $\rightsquigarrow$ think of x as being a scalar categorical random variable with K states, and $\boldsymbol{x}$ is its **dummy encoding**.
- Example: $K = 3$, encode the states 1, 2 and 3 as $(1, 0, 0), (0, 1, 0)$, and $(0, 0, 1)$.
- Also called a **one-hot encoding**, since we imagine that only one of the K "wires" is "hot" or on.

$$\text{Mu}(\boldsymbol{x}|1, \boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}.$$

Common discrete distributions: Empirical

- Given a set of data, $\mathcal{D} = \{x_1, \ldots, x_N\}$, define the **empirical distribution**, a.k.a. **empirical measure**:

$$p_{emp}(A) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A),$$

where

$$\delta_x(A) = \begin{cases} 0, & \text{if } x \notin A \\ 1, & \text{if } x \in A \end{cases}$$

- In general, we can associate **weights** with each sample:

$$p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$$

where we require $0 \leq wi \leq 1$ and $\sum_{i=1}^{N} w_i = 1$.

- We can think of this as a histogram, with "spikes" at the data points x_i, where w_i determines the height of spike i.
- This distribution assigns 0 probability to any point not in the data set.

Common continuous distributions: Normal

- The pdf of the normal distribution is

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

where μ is the mean, σ^2 is the variance. The inverse variance is sometimes called **precision**.

- The cdf of the standard normal distribution is the integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

It has no closed form expression.

- The cdf is sometimes expressed in terms of the **error function**

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt,$$

as follows:

$$\Phi(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right].$$

Common continuous distributions: Normal

- If σ tends to zero, $p(x)$ tends to zero at any $x \neq \mu$, but grows without limit if $x = \mu$, while its integral remains equal to 1.
- Can be defined as a generalized function: **Dirac's delta function** δ translated by the mean: $p(x) = \delta(x - \mu)$, where

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0, \end{cases}$$

additionally constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1.$$

- Sifting property: selecting out a single term from a sum or integral:

$$\int_{-\infty}^{\infty} f(x)\delta(x - z)\, dx = f(z)$$

since the integrand is only non-zero if $x - z = 0$.

Central Limit Theorem

- Under certain (fairly common) conditions, the sum of many random variables will have an **approximately normal distribution.**
- Let $X_1, \ldots, X_n$ be i.i.d. RVs with the same (arbitrary) distribution, zero mean, and variance σ^2.
- Let

$$Z = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

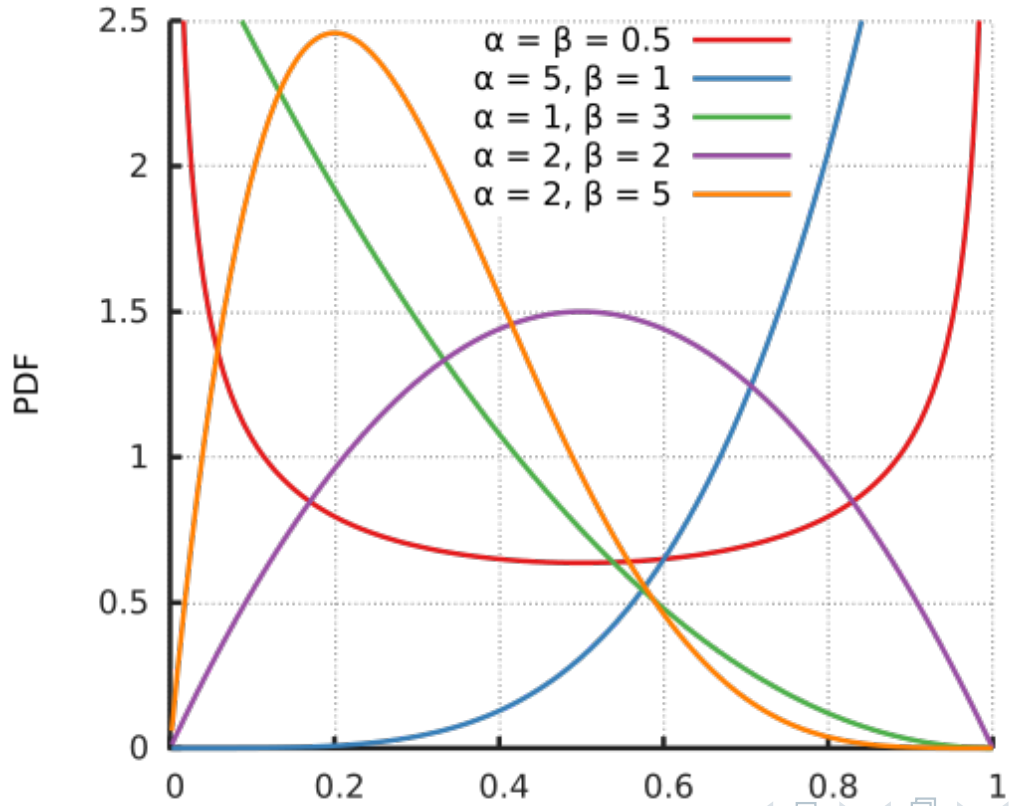
- Then, as n increases, the probability distribution of Z will tend to the normal distribution with zero mean and variance σ^2.

Common continuous distributions: Beta

- The beta distribution is supported on the unit interval $[0, 1]$
- For $0 \leq x \leq 1$, and shape parameters $\alpha, \beta > 0$, the pdf is

$$f(x; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

The beta function, B, is a normalization constant to ensure that the total probability is 1.



Common continuous distributions: Multivariate Normal

- The multivariate normal distribution of a k-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_k)^t$ can be written as: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with k-dimensional mean vector

  $$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]^{\mathrm{t}}$$

  and $k \times k$ covariance matrix

  $$\Sigma =: \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{t}}] = [\mathrm{Cov}[X_i, X_j]; 1 \leq i, j \leq k],$$

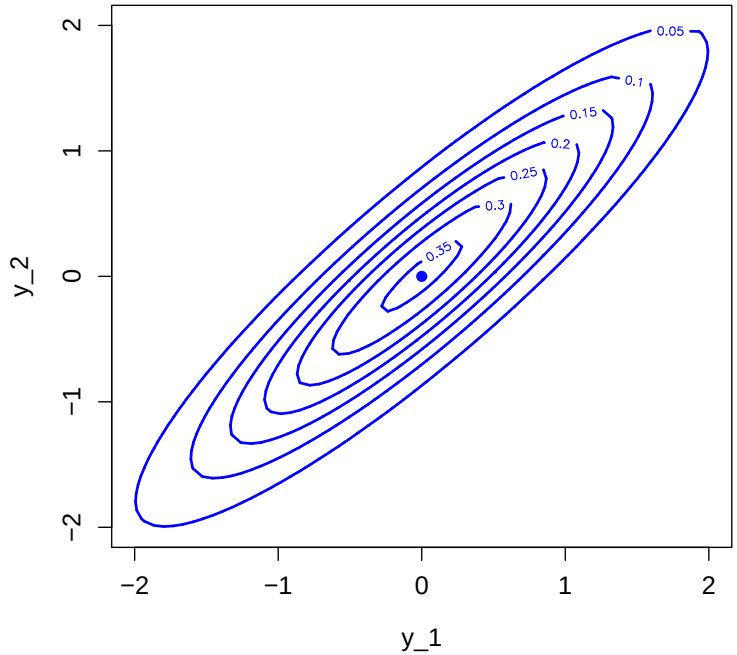
  where

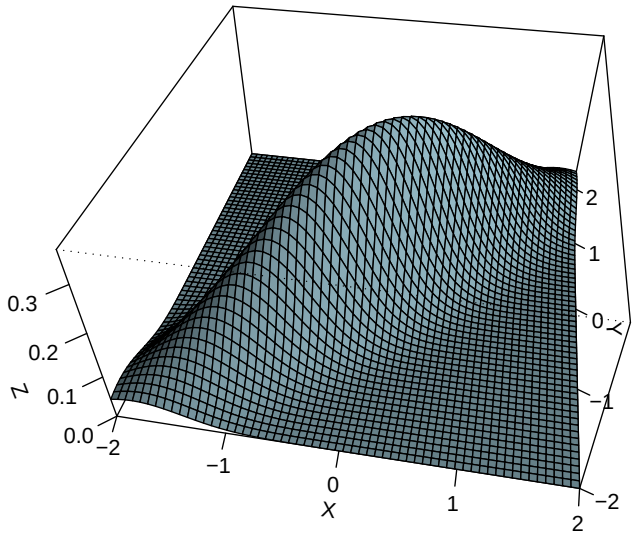
  $$\mathrm{Cov}[X_i, X_j] = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

- The inverse of the covariance matrix is the precision matrix $Q = \Sigma^{-1}$.
- The pdf of the multivariate normal distribution is

$$p(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{t}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

The 2D Normal distribution





Affine transformations:

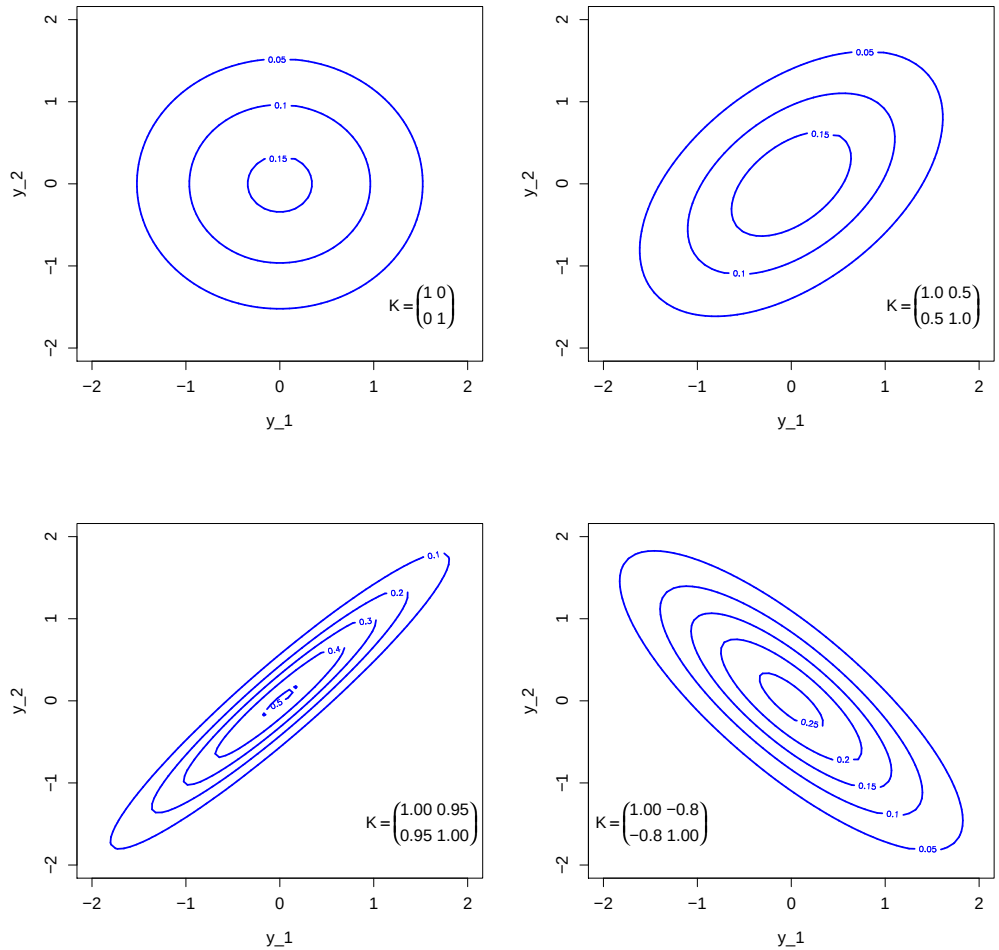
If $\boldsymbol{y} = \boldsymbol{c} + B\boldsymbol{x}$ is an affine transformation of $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{c} + B\boldsymbol{\mu}, B\Sigma B^{\mathrm{t}}\right)$

The 2D Gaussian distribution

2D Gaussian: $p(\boldsymbol{x}|\boldsymbol{\mu} = \mathbf{0}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp(-\frac{1}{2}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x})$

Covariance
(also written "*co*-variance")
is a measure of how much **two random variables vary together**:

- **+1**: perfect linear coherence,
- **-1**: perfect negative linear coherence,
- **0**: no linear coherence.



Common continuous distributions: Dirichlet

- The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ is a multivariate generalization of the beta distribution.
- Its pdf on $\mathbb{R}^{K-1}$ is

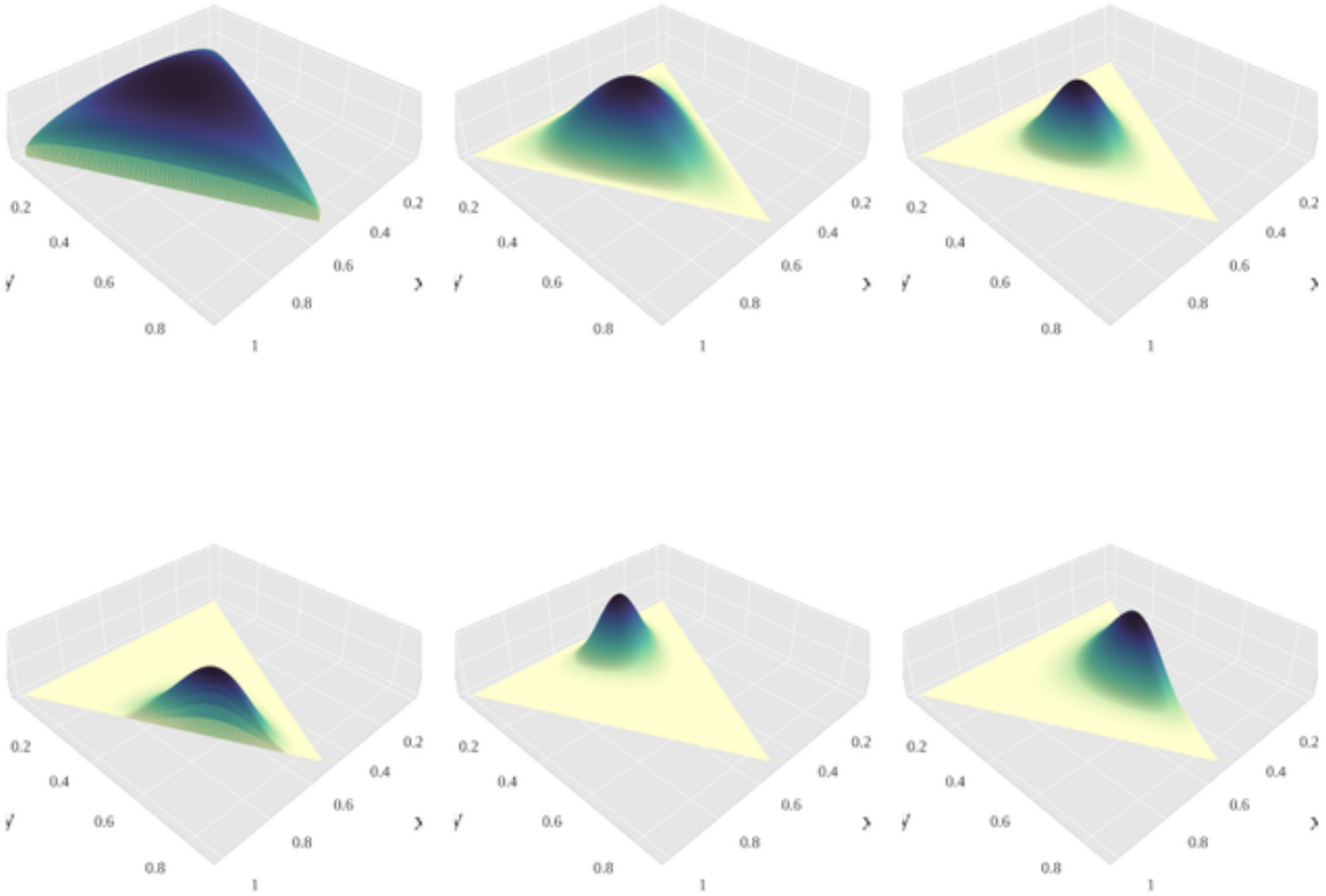
$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $\{x_k\}_{k=1}^{k=K}$ belong to the standard $K-1$ simplex:

$$\sum_{i=1}^{K} x_i = 1 \text{ and } x_i \geq 0 \; \forall i \in [1, K]$$

- The normalizing constant is the multivariate beta function.

Common continuous distributions: Dirichlet



wikimedia.org/w/index.php?curid=49908662

Conditional probability

- **Conditional** or **posterior probabilities**
  e.g., $P(cavity|toothache) = 0.8$
  i.e., **given that** *toothache* **is all I know**
  NOT "if *toothache* then 80% chance of *cavity*"
- Notation for conditional distributions:
  $P(Cavity|Toothache)$ = 2-element vector of 2-elem. vectors.
- If we **know more**, e.g., *cavity* is also given, then we have
  $P(cavity|toothache, cavity) = 1$
  Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**.
- New evidence may be **irrelevant**, allowing **simplification**:
  $P(cavity|toothache, die\ roll = 3) = P(cavity|toothache) = 0.8$

Conditional probability

- Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

  Product rule gives an alternative formulation:
  $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

- A general version holds for whole **distributions**, e.g.,
  $P(Weather, Cavity) = P(Weather|Cavity)P(Cavity)$

- **Chain rule** is derived by successive application of product rule:

$$\begin{aligned} & P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1})\ P(X_n|X_1, \ldots, X_{n-1}) \\ &= P(X_1, \ldots, X_{n-2})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ P(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a **normalization constant** α

$$\begin{aligned}
\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache) \\
&= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

General idea: compute distribution on query variable
by fixing **evidence variables** and summing over **hidden variables**

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want the posterior joint distribution of the **query variables** $\mathbf{Y}$ given specific values $\mathbf{e}$ for the **evidence variables** $\mathbf{E}$

Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by **summing out the hidden variables**:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

Joint probability $p(x) = p(x_1,\ldots,x_n) \rightsquigarrow$ number of states: $\prod_{i=1}^{n} |arity(x_i)|$.

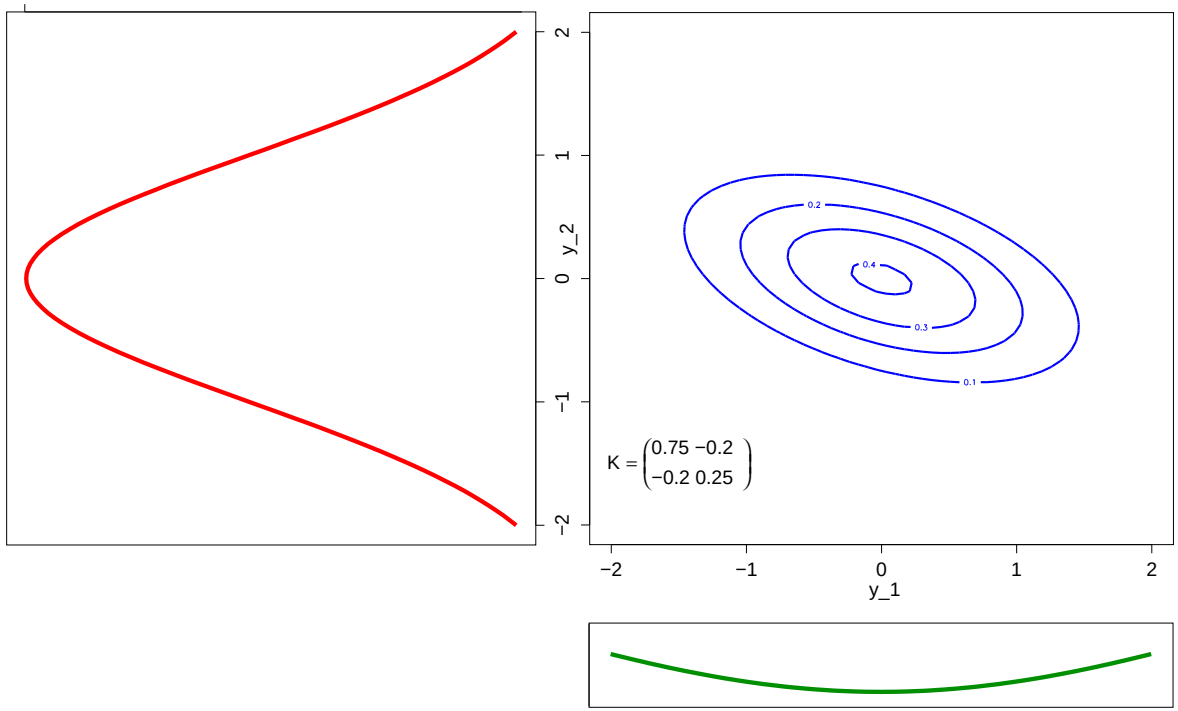
Obvious problems:

1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Inference in Jointly Gaussian Distributions: Marginalization

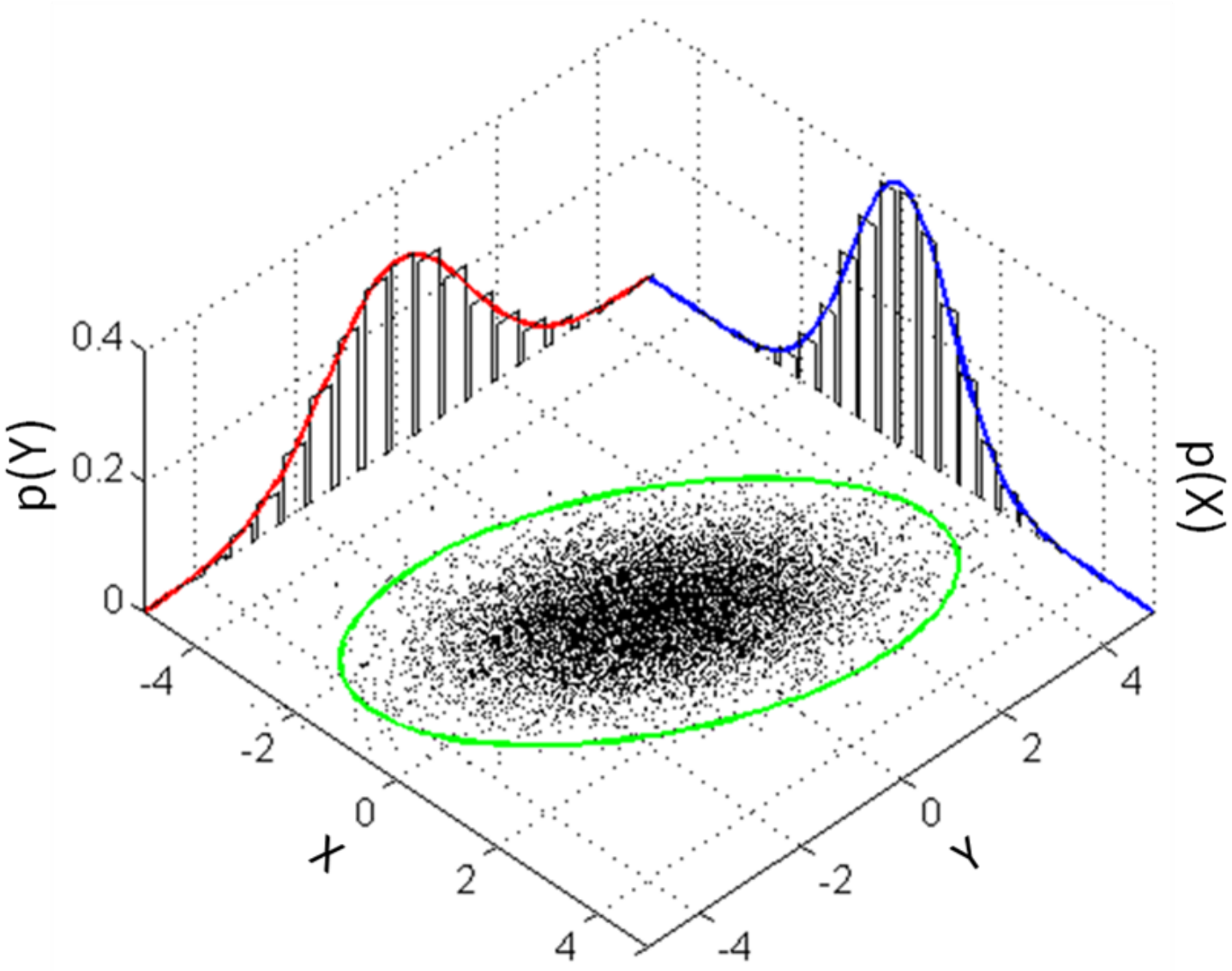
$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Let $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

Then $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22})$.



Marginals of Gaussians are again Gaussian!

Inference in Jointly Gaussian Distributions

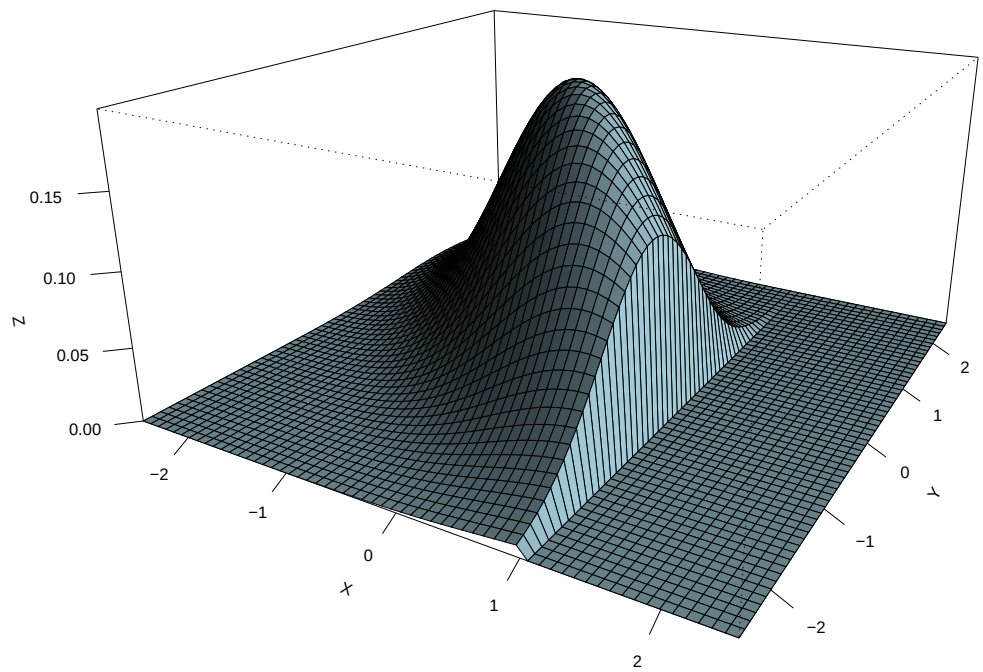


wikimedia.org/w/index.php?curid=25235145

Inference in Jointly Gaussian Distributions

$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Let $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

Then $\boldsymbol{x}_2 | \boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$.

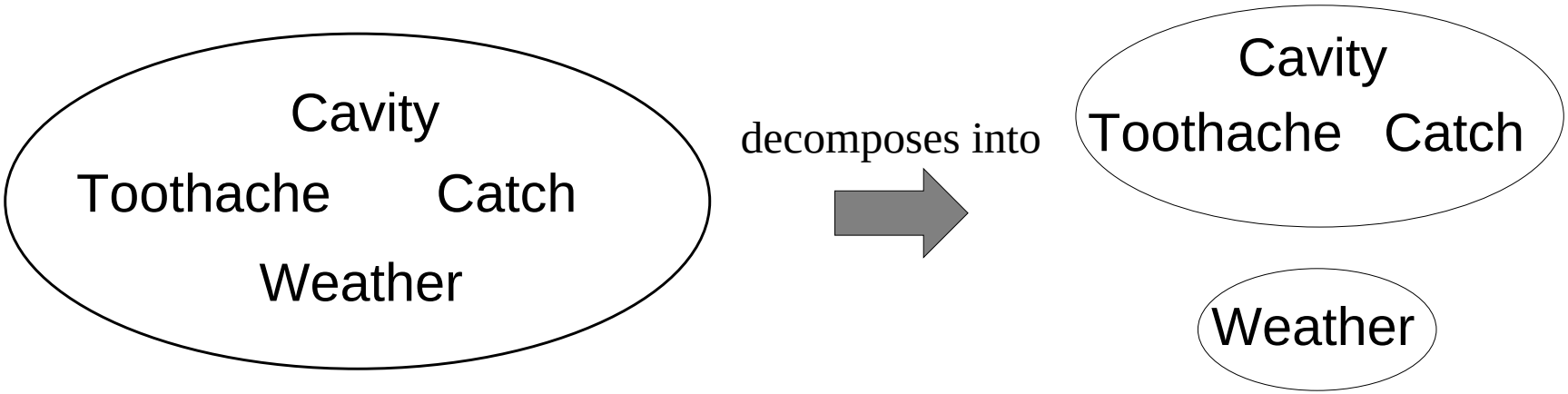


Conditionals of Gaussians are again Gaussian!

Independence

A and B are **independent** iff

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$



$\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}, \mathit{Weather})$
$= \mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})\mathbf{P}(\mathit{Weather})$

$\rightsquigarrow 4 \cdot 8 = 32$ entries reduced to $4 + 8 = 12$.

Absolute independence powerful but rare...

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:
(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is **conditionally independent** of $Toothache$ given $Cavity$:
$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:
$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$
$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Conditional independence contd.

Write out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache} | \mathit{Catch}, \mathit{Cavity})\mathbf{P}(\mathit{Catch}, \mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache} | \mathit{Catch}, \mathit{Cavity})\mathbf{P}(\mathit{Catch} | \mathit{Cavity})\mathbf{P}(\mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache} | \mathit{Cavity})\mathbf{P}(\mathit{Catch} | \mathit{Cavity})\mathbf{P}(\mathit{Cavity})
\end{aligned}$$

I.e., only $2 + 2 + 1 = 5$ independent numbers.

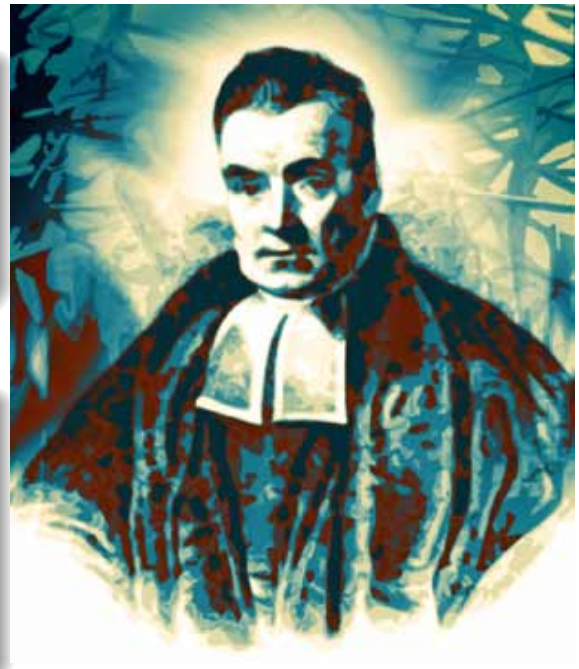
Often, conditional independence reduces the size of the representation of the joint distribution from **exponential** in n to **linear** in n.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes Rule

Bayes Rule

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



Proof.

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

□

Bayes Rule (cont'd)

- Useful for assessing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = \frac{\overbrace{P(Effect|Cause)}^{\text{Prob(symptoms)}}\overbrace{P(Cause)}^{\text{Prevalence}}}{P(Effect)}$$

- E.g., let M be **meningitis** (acute inflammation of the protective membranes covering the brain and spinal cord),
  S be **stiff neck**. Assume the doctor knows that the prevalence of meningitis is $1/50,000$, that the prior probability of a stiff neck is 0.01, and that the symptom *stiff neck* occurs with a probability of 0.7.

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014.$$

- Note: the posterior probability of meningitis is still very small (1 in 700 patients)!

Bayes rule (cont'd)

Question: Why should it be easier to estimate the conditional probabilities in the causal direction $P(Effect|Cause)$, as compared to the diagnostic direction, $P(Cause|Effect)$?
There are two possible answers (in a medical setting):

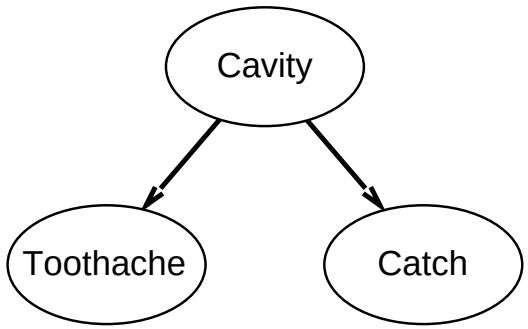
- We might have access to a collection of health records for patients having meningitis. This collections will provide us with estimates of $P(s|m)$. For directly estimating $P(m|s)$ we would need a database of all cases of the **very unspecific symptom**.
- Diagnostic knowledge might be **more fragile** than causal knowledge. Assume a doctor has directly estimated $P(m|s)$. If there is a sudden epidemic of meningitis, $P(m)$ will go up, but this doctor will have no idea how to update $P(m|s)$. The other doctor who uses Bayes rule knows that $P(m|s)$ should go up proportionately with $p(m)$.
  Note that causal information $P(s|m)$ is **unaffected by the epidemic** (it simply reflects the way how meningitis works)!

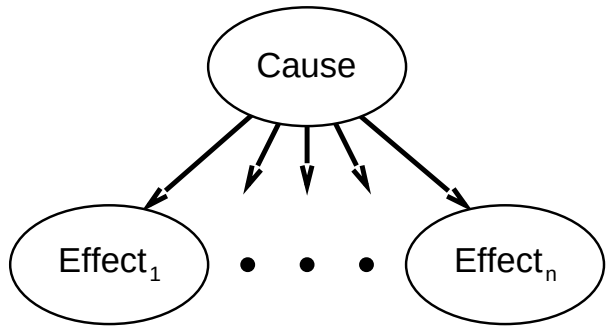
Bayes' Rule and conditional independence

$$
\begin{aligned}
&P(Cavity|toothache \wedge catch) \\
&\quad = \quad \alpha\, P(toothache \wedge catch|Cavity)P(Cavity) \\
&\quad = \quad \alpha\, P(toothache|Cavity)P(catch|Cavity)P(Cavity)
\end{aligned}
$$

This is an example of a **naive Bayes** model:

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause)\prod_i P(Effect_i|Cause)$$

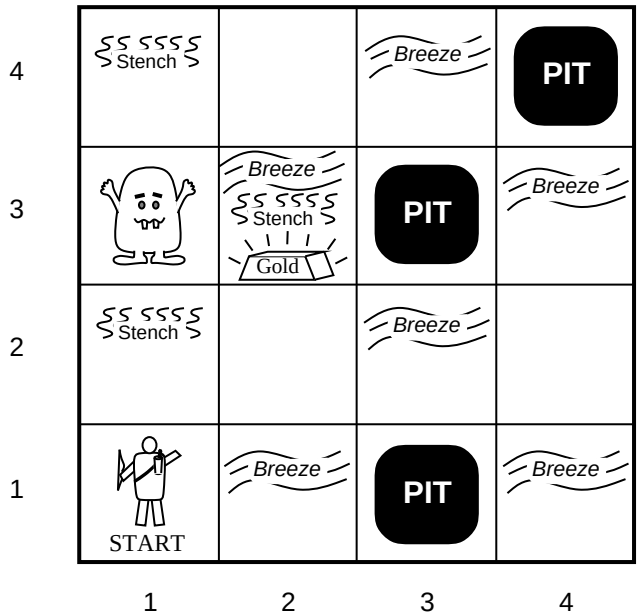




Total number of parameters is **linear** in n

Example: Wumpus World

- The **wumpus** is a beast that eats anyone who enters the room.
- Some rooms contain bottomless **pits** that will trap anyone entering the room (except for the wumpus, which is too big to fall in!)
- The only positive aspect is the possibility of finding **gold**...



Squares adjacent to wumpus are **smelly**.
Squares adjacent to pit are **breezy**.
Glitter if and only if gold is in the same square.
Shooting kills the wumpus if you are facing it.
Shooting uses up the only **arrow**.
Grabbing picks up the gold if in the same square.
Releasing drops the gold in the same square.

Goal: Get gold back to start without entering pit or wumpus square

Wumpus World

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 **B** **OK**	2,2	3,2	4,2
1,1 **OK**	2,1 **B** **OK**	3,1	4,1

$P_{ij} = true$ iff $[i, j]$ contains a pit

$B_{ij} = true$ iff $[i, j]$ is breezy

Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the probability model

The full joint distribution is $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$
Apply product rule: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$
(Do it this way to get $P(\mathit{Effect}|\mathit{Cause})$.)
First term: 1 if pits are adjacent to breezes, 0 otherwise
Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and query

We know the following facts:

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

Query is $\mathbf{P}(P_{1,3}|known, b)$

Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$

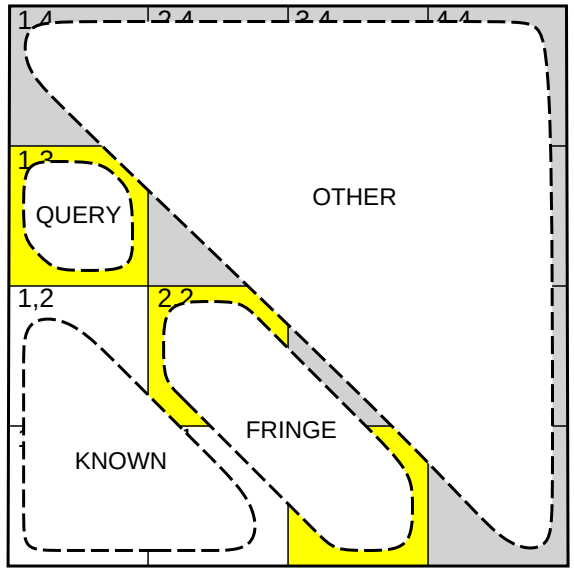
For inference by enumeration, we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Using conditional independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares



Define $Unknown = Fringe \cup Other$

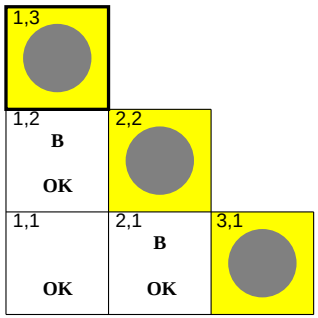
$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

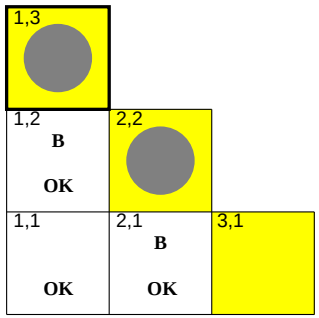
Using conditional independence contd.

$$\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\
&= \alpha'\, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)
\end{aligned}$$

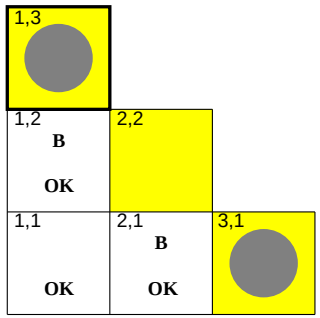
Using conditional independence contd.



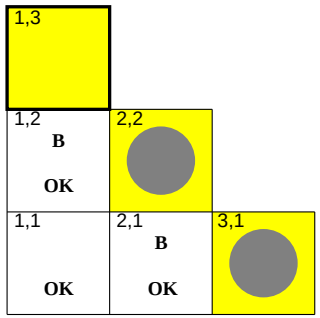
0.2 x 0.2 = 0.04



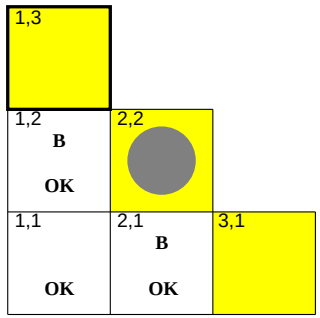
0.2 x 0.8 = 0.16



0.8 x 0.2 = 0.16



0.2 x 0.2 = 0.04



0.2 x 0.8 = 0.16

$$\begin{aligned}\mathbf{P}(P_{1,3}|known, b) &= \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16)\rangle \\ &\approx \langle 0.31, 0.69\rangle\end{aligned}$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14\rangle$$

Summary

- Probability is a rigorous formalism for uncertain knowledge
- **Joint probability distribution** specifies probability of every **atomic event**
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- **Independence** and **conditional independence** provide the tools

Subsection 1

Origins of probabilities

Origins of probabilities I

Historically speaking, probabilities have been regarded in a number of different ways:

- **Frequentist position:** probabilities come from **measurements**. The assertion $P(\mathit{cavity}) = 0.05$ means that 0.05 is the fraction that would be observed in the limit of infinitely many samples. From a finite sample, we can estimate this true fraction and also calculate how accurate this estimates is likely to be.
- **Objectivist view:** probabilities are actual **properties of the universe**
  An excellent example: quantum phenomena.
  A less clear example: coin flipping – the uncertainty is probably due to our uncertainty about the initial conditions of the coin.

Origins of probabilities II

- **Subjectivist view**: probabilities are an **agent's degrees of belief**, rather than having any external physical significance.
- The **Bayesian view** allows any self-consistent ascription of prior probabilities to propositions, but then insists on proper **Bayesian updating** as evidence arrives.
- For example $P(\mathit{cavity}) = 0.05$ denotes the degree of belief that a random person has a cavity **before we make any actual observation** of that person.
  Updating in the light of **further evidence** "person has a *toothache*":

$$P(\mathit{cavity}|\mathit{toothache}) = \alpha\, P(\mathit{toothache}|\mathit{cavity})P(\mathit{cavity})$$

The reference class problem

- Even a strict frequentist position involves **subjective** analysis.
- **Example:** Say a doctor takes a frequentist approach to diagnosis.
  She examines a large number of people to establish the probability of whether or not they have heart disease.
  To be accurate she tries to measure "similar people" (she knows for example that gender might be important.)
- Taken to an extreme, all people are different and therefore the **reference class is empty**.
  This has been a vexing problem in the philosophy of science.

Frequentist and Bayesian view: a simple example

- Assume $x_1, \ldots, x_n$ are drawn i.i.d. from normal $N(\mu, \sigma^2)$ with known variance σ^2. What can be said about μ?
- Frequentist view: no further probabilistic assumptions
  $\rightsquigarrow$ treat μ as an **unknown constant**.
- The sample mean $\bar{x} = \sum_i x_i / n$ is the observed value of the RV
  $$\bar{X} \sim N(\mu, \bar{\sigma}^2), \quad \text{with } \bar{\sigma}^2 = \sigma^2/n.$$
  Now define the linearly transformed random variable
  $$B := \frac{\mu - \bar{X}}{\bar{\sigma}} \sim N(0,1), \text{ (i.e. standard normal).}$$
  Use normal cdf $\Phi(k_c) = P(B < k_c)$ to derive an upper limit for μ:
  $$\begin{aligned} P(B < k_c) &= \Phi(k_c) = 1 - c \\ &= P(-\bar{\sigma} B > -\bar{\sigma} k_c) \\ &= P(\underbrace{\mu - \bar{\sigma} B}_{\bar{x}} > \mu - \bar{\sigma} k_c) \\ &= P(\bar{X} + \bar{\sigma} k_c > \mu). \end{aligned}$$

Frequentist and Bayesian view: a simple example

- The statement $P(\mu < \bar{X} + \bar{\sigma} k_c) = 1 - c$ can be interpreted as specifying a **hypothetical long run of statements** about the **constant** μ, **a portion** $1 - c$ **of which is correct.**
  (Note that $\bar{X}$ is a RV!)
- Plugging in the **observed** $\bar{x}$, the statement $\mu < \bar{x} + \bar{\sigma} k_c$ can be interpreted as **one** of a long run of such statements about μ.
- Arguments involving probability **only via its (hypothetical) long-run frequency interpretation** are called **frequentist**.
- That is, in the frequentist world we define procedures for assessing evidence that are calibrated by how they would perform **were they used repeatedly.**

Frequentist and Bayesian view: a simple example

- From the Bayesian viewpoint, we treat μ as having a probability distribution **both with and without the data.** That is, μ is the unobserved value of the random variable M.
- Bayes' theorem: $p_{M|X}(\mu|x) = \alpha \times p_{X|M}(x|\mu)p_M(\mu)$.
- Intuitive idea:
  - all relevant information about μ is in the conditional distribution, given the data;
  - this distribution is determined by the elementary formulae of probability theory;
  - remaining problems are solely computational.
- Example: choose $p(\mu) = N(m, \nu^2) \rightsquigarrow p(\mu|x) = N(\tilde{m}, \tilde{\nu}^2)$ with

$$\tilde{m} = \frac{\bar{x}/\bar{\sigma}^2 + m/\nu}{1/\bar{\sigma}^2 + 1/\nu^2}, \quad \tilde{\nu}^2 = \frac{1}{1/\bar{\sigma}^2 + 1/\nu^2}$$

"Normal likelihood times normal prior gives normal posterior"

Frequentist and Bayesian view: a simple example

- Same reasoning as before: define transfomed $\tilde{B} := \frac{\mu - \tilde{m}}{\tilde{\nu}} \sim N(0,1)$
- Upper limit for μ: $P(\mu < \tilde{m} + k_c \tilde{\nu}) = 1 - c$.
- If the prior variance $\nu^2 \gg \bar{\sigma}^2$ and the prior mean m is not too different from $\bar{X}$, this limit agrees closely with the one obtained by the frequentist method (because then $\tilde{m} \approx \bar{X}$ and $\tilde{\nu} \approx \bar{\sigma}$).
- This broad parallel between the different types of analysis is in no way specific to the normal distribution.
- See the beautiful book (D.R. Cox, Principles of statistical inference, Cambridge, 2006) for further details.

Subsection 2

Some important statistical concepts

Convergence of random variables

Definition (Convergence in Probability)

Let $X_1, X_2, \ldots$ be random variables. We say that X_n **converges in probability** to the random variable X as $n \to \infty$, iff, for all $\varepsilon > 0$,

$$P(|X_n - X| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

We write $X_n \xrightarrow{p} X$ as $n \to \infty$.

Example: Weak law of large numbers

Theorem (Bernoulli's Theorem, Weak law of large numbers)

Let $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed (i.i.d.) random variables, each having mean μ (and standard deviation σ). Let $S_n = X_1 + \ldots + X_n$. Then

$$P(|S_n/n - \mu| > \varepsilon) \to 0$$

as $n \to \infty$.

- Given sufficiently many observations x_i, the sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ will approach the true mean μ.
- Note that $|S_n/n - \mu| > \varepsilon$ might happen an **infinite number of times**, although at **infrequent intervals**.
- The **strong law** even says that for any $\varepsilon > 0$ the inequality $|S_n/n - \mu| < \varepsilon$ holds for all large enough n, but we will not discuss this further...

Example: convergence of empirical CDF

Definition (Empirical cumulative distribution function)

Let $X_1, X_2, \ldots, X_n$ be iid real random variables with the common cdf $F(t)$. Then the empirical distribution function is defined as

$$\hat{F}_n(t) = \frac{\#(\text{elements}) \text{ in the sample} \leq t}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq t\},$$

where $\mathbf{1}\{A\}$ is the indicator of event A.

- For a fixed t, the indicator $\mathbf{1}\{X_i \leq t\}$ is a Bernoulli random variable with mean $\mu = F(t)$.
- Weak law of large numbers $\Rightarrow$ estimator $\hat{F}_n(t)$ converges in probability to $F(t)$ as $n \to \infty$, for every value of t:

$$\hat{F}_n(t) \xrightarrow{p} F(t).$$

Expectation

Definition (Expectation)

Let X be a random variable with probability density function f_X.
The expectation is

$$E[X] := \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Definition (Sample mean)

Let a sample $x = \{x_1, x_2, \ldots, x_n\}$ be given. The **sample mean** is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample mean is an **unbiased estimator** of $\mu = E[X]$.

Variance

Definition (Variance)

Let X be a random variable with density function f_X. The **variance** is

$$\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

The square root $\sqrt{\mathrm{Var}[X]}$ is the **standard deviation**.

Definition (Sample Variance)

Let the sample $x = \{x_1, x_2, \ldots, x_n\}$ with sample mean $\overline{x}$ be given.
The **sample variance** is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

The sample variance is an **unbiased estimator** of $\mathrm{Var}[X]$.

Bias and Variance of an Estimator

- Assume that a statistical model parametrized by θ gives rise to a **probability distribution for observed data**, $P(x|\theta)$.
- Let $\hat{\theta}$ be an estimator of θ based on any observed data x, i.e. $\hat{\theta}$ maps observed x to values that we hope are close to θ.
- The bias of $\hat{\theta}$ is defined to be

$$\text{Bias}[\hat{\theta}] = E_{P(x|\theta)}[\hat{\theta}] - \theta = E_{P(x|\theta)}[\hat{\theta} - \theta],$$

  where $E_{P(x|\theta)}[\cdot]$ denotes expected value over the distribution $P(x|\theta)$, i.e. averaging over all possible observations x.
- An estimator is **unbiased** if its bias is zero **for all values of parameter** θ.
- The variance of $\hat{\theta}$ is the expected value of the squared sampling deviations: $\text{var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$.