



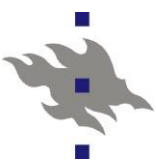
■ Winter School on 'Big Data Challenges to
■ Modern Statistics'
■ Geilo Jan, 2014

(A Light Appetizer before Dinner)

Bayesian Inference: What, and Why?

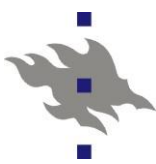
Elja Arjas

UH, THL, UiO



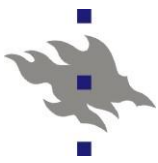
Understanding the concepts of randomness and probability: Does it make a difference?

- In the Bayesian approach to statistics, a crucially important distinction is made between variables/quantities depending on whether their true values are **known** or **unknown** (to me, or to you, as an observer).
- In the Bayesian usage/semantics, the epithet “**random**”, as in “**random variable**”, means that “**the exact value** of this variable **is not known**”.



Understanding the concepts of randomness and probability: Does it make a difference?

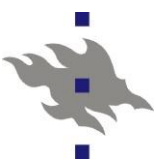
- Stated briefly: In Bayesian inference
"random" = "uncertain to me (or to you) as an observer"
- Correct understanding of the semantics that are used is important – and it helps you to avoid obvious misunderstandings!



Understanding the concepts of randomness and probability: Does it make a difference?

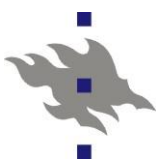
- For example:
 - "An event (in the future) is **random** (to me) if I am **uncertain** about whether it will occur or not".
 - "An event (in the past) is **random** (to me) if I am **uncertain** about whether it has occurred or not".

- "Randomness" does not require "variability", e.g. in measurements over time or in samples drawn from a population.



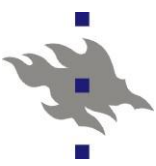
Understanding the concepts of randomness and probability: Does it make a difference?

- Even unique events, statements, or quantities can be "random": The number of balls in this box now is "random" to (any of) you.
- It may not be "random" for me (because I put the balls into the box before this lecture, and I might remember ...).



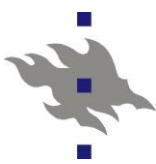
Understanding the concepts of randomness and probability: Does it make a difference?

- The characterization of the concept of a **parameter** that is found in many textbooks of statistics, as being something that is 'fixed but unknown', would for a Bayesian – by definition - mean that it is a **random variable**!
- **Data**, on the other hand, **after their values have been observed**, are **no longer "random"**.
- The dichotomy (**population**) **parameters** vs. **random variables**, which is fundamental in classical / frequentist statistical modeling and inference, has lost its significance in the Bayesian approach.



Understanding the concepts of randomness and probability: Does it make a difference?

- Probability = degree of uncertainty, expressed as my / your subjective assessment, based on the available **information**.
- All probabilities are conditional. To make this aspect explicit in the notation we write systematically $P(. | I)$, where I is the **information** on which the assessment is based. Usually, however, the role of I is left implicit, and I is dropped from the probability expressions. (Not here ...!)



Understanding the concepts of randomness and probability: Does it make a difference?

- Note: In probability calculus it is customary to define **conditional probabilities** as ratios of 'absolute' probabilities, via the formula $P(B | A) = P(A \cap B) / P(A)$. Within the Bayesian framework, such 'absolute' probabilities do not exist.
- But the same formula, in the form of the 'chain multiplication rule' $P(A \cap B | I) = P(A | I) P(B | A, I)$ still holds!



Understanding the concepts of randomness and probability: Does it make a difference?

“There are no unknown probabilities in a Bayesian analysis, only unknown - and therefore random - quantities *for which you have a probability* based on your background information” (O'Hagan 1995).



Understanding the concepts of randomness and probability: Does it make a difference?

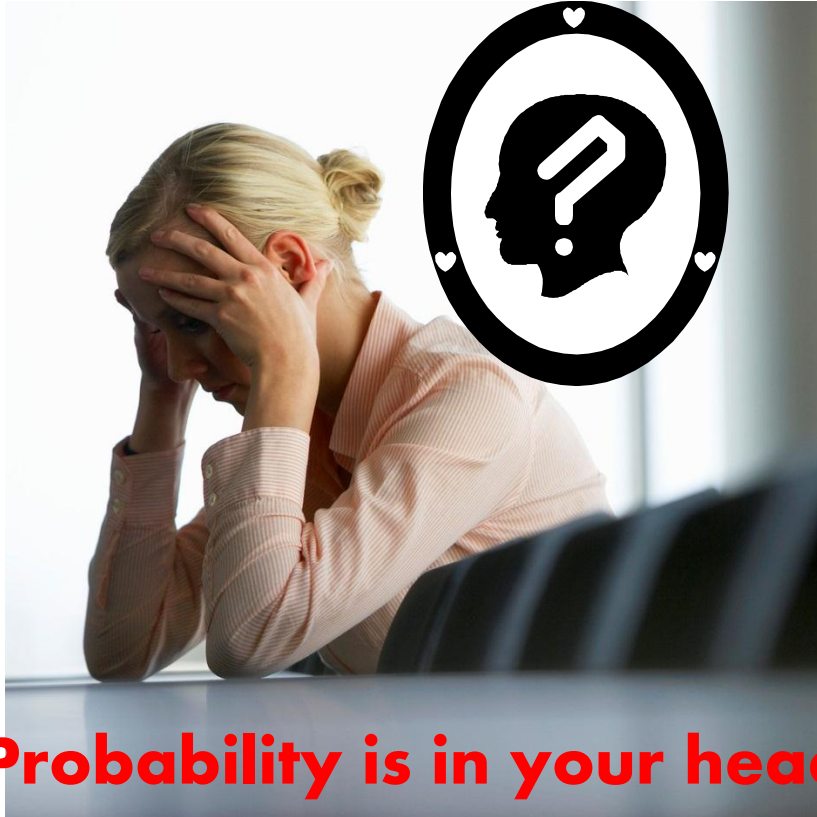
- Note here the wording
'probability for ...', not 'probability of ...'
- This corresponds to an understanding, where probabilities are not quantities which have an objective existence in the physical world (as would be, for example, the case if they were identified with observable frequencies).



Understanding the concepts of randomness and probability: Does it make a difference?

- **Probability does not exist !** (Bruno de Finetti, 1906-1985)
- **Projection fallacy !** (Edwin T Jaynes, 1922 – 1998)
- (Convey the idea that probability is an expression of an observer's view of the world, and as such it has no existence of its own).

Bayesian probability: P



Probability is in your head

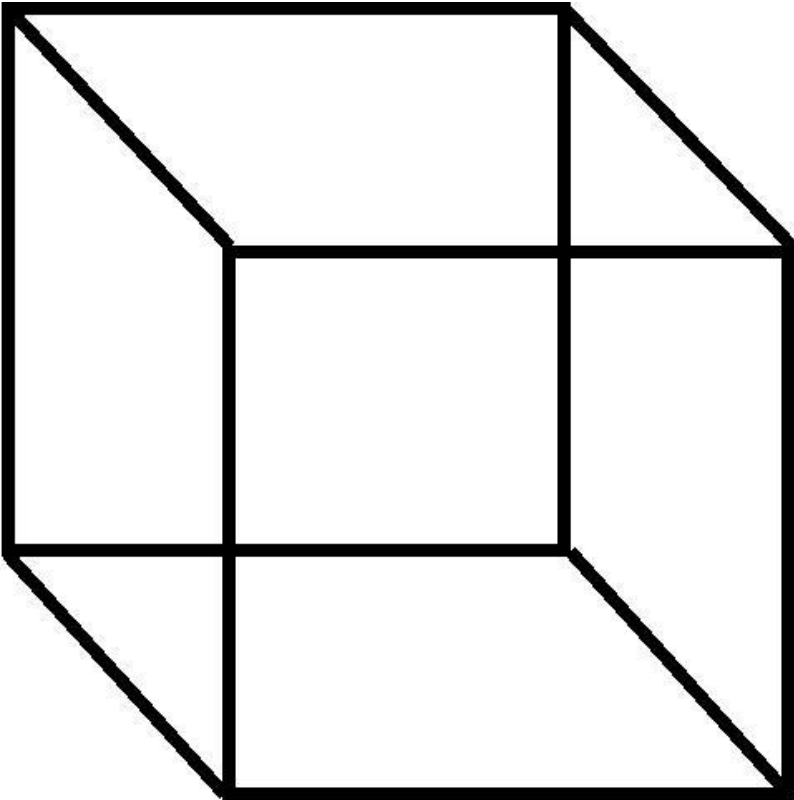
State of the World: θ



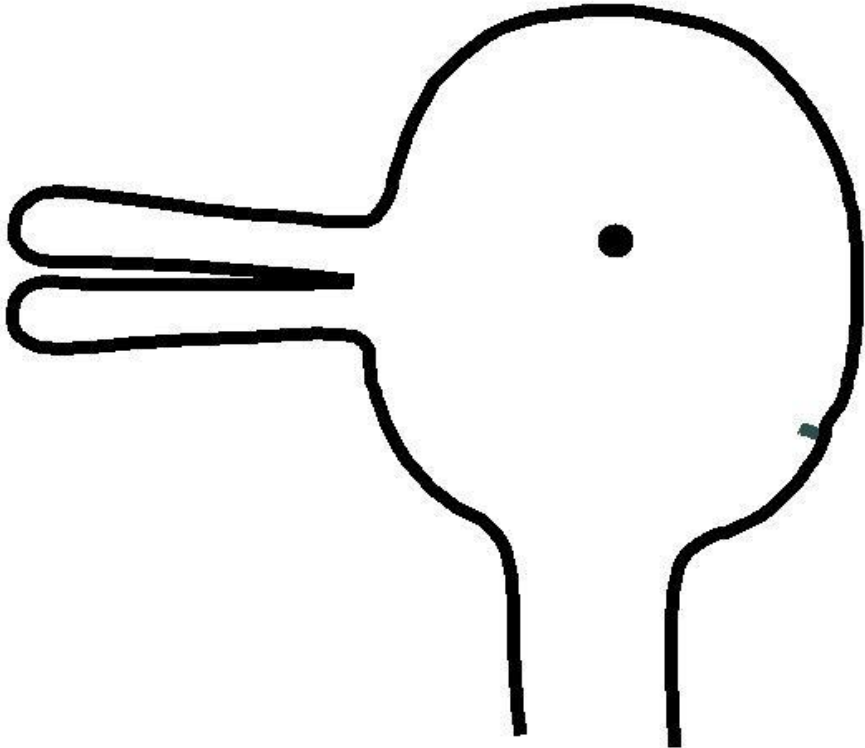
$P(\theta$ | your information I)



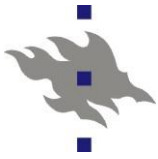
Two ways to look at things – and it is not always easy to switch from one to the other



Necker cube

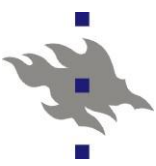


Duck-rabbit



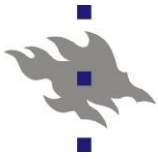
Obvious reservation ...

- This view of the concept of probability applies in the macroscopic scale, and does not say anything about the role of probability in describing quantum phenomena.
- Still OK for me, and perhaps for you as well ...



Understanding the concepts of randomness and probability: Does it make a difference?

- Understanding the meaning of the concept of probability, in the above sense, is crucial for Bayesian statistics.
- This is because: All Bayesian statistics involves in practice is actually evaluating such probabilities!



Simple motivation for Bayesian inference:

- Use probability as a description of your uncertainty about things which **you don't** know – such as values of **parameters** or of future observables – instead of things which you **do know** – such as the **data** you have.



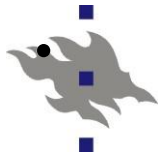
Understanding the concepts of randomness and probability: Does it make a difference?

- 'Ordinary' probability calculus (based on Kolmogorov's axioms) applies without change, apart from that the usual definition of conditional probability $P(A | B) = P(A \cap B) / P(B)$ becomes 'the chain multiplication rule'

$$\underline{P(A \cap B | I) = P(A | I) P(B | A, I) = P(B | I) P(A | B, I).}$$

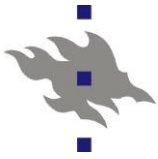
- Expressed in terms of probability densities, this becomes

$$\underline{p(x, y | I) = p(x | I) p(y | x, I) = p(y | I) p(x | y, I).}$$



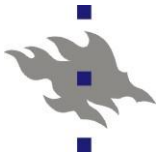
Controversy between statistical paradigms

It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. (L J Savage 1972).



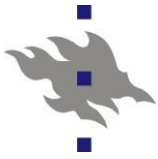
Simple example: Balls in a box

- Suppose there are N 'similar' balls (of the same size, made of the same material, ...) in a box.
- Suppose further that K of these balls are white and the remaining $N - K$ are yellow.
- Shake the contents of the box thoroughly. Then draw – blindfolded – one ball from the box and check its colour!



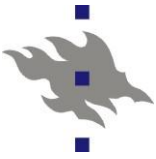
Balls in a box (cont'd)

- This is the background information I , which is given for an assessment of the probability for $P(\text{'the colour is white' } | I)$.
- What is your answer?



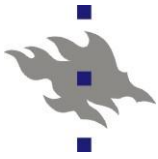
Balls in a box (cont'd)

- Each of the N balls is as likely to be drawn as any other (exchangeability), and K of such draws will lead to the outcome 'white' (additivity). Answer: K / N .
- Note that K and N are here assumed to be known values, provided by I , and hence 'non-random'.
- We can write
$$P(\text{'the colour is white'} | I) = P(\text{'the colour is white'} | K, N) = K / N.$$



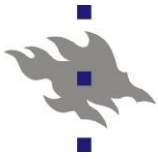
Balls in a box (cont'd):

- Shaking the contents ... being blindfolded: no idea of how the balls in the box are arranged when one is chosen.
- If the person drawing a ball were allowed to look into the box and check the colours of the individual balls, 'randomness' in the experiment would disappear.
- "What is the probability that the present Pope is Chinese?" (Stephen Hawking, in "The Grand Design", 2010)



Balls in a box (cont'd): conditional independence

- **Balls in a box (cont'd):** Consider then a sequence of such draws, such that the ball that was drawn is put back into the box, and the contents of the box are shaken thoroughly.
- Because of the thorough mixing, any information about the positions of the previously drawn balls is lost. Memorizing the earlier results does not help beyond what we know already: N balls, out of which K are white.
- Hence, denoting by X_i the color of the i^{th} draw, we get the crucially important **conditional independence** property
$$P(X_i | X_1, X_2, \dots, X_{i-1}, I) = P(X_i | I).$$



Balls in a box (cont'd): conditional independence

- **Balls in a box (cont'd):** Hence, for any $i \geq 1$,

$$P(X_1, X_2, \dots, X_i | I)$$

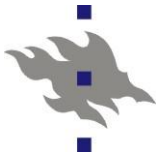
$$= P(X_1 | I) P(X_2 | X_1, I) \dots P(X_i | X_1, X_2, \dots, X_{i-1}, I) \quad \text{chain rule}$$

$$= P(X_1 | I) P(X_2 | I) \dots P(X_i | I) \quad \text{conditional independence}$$

$$= P(X_1 | K, N) P(X_2 | K, N) \dots P(X_i | K, N)$$

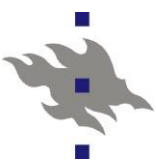
$$= (K/N)^{\#\{\text{white balls in } i \text{ draws}\}} [1 - (K/N)]^{\#\{\text{yellow balls in } i \text{ draws}\}} .$$

- Old result, as if there were no difference to the 'classical' concept of probability based on symmetry.



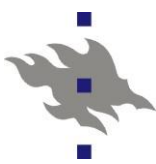
Balls in a box (cont'd): from parameters to data

- Indeed: The variables N and K , whose values are here taken to be contained in the background information I , could be called 'parameters' of the distribution of each X_i .



Balls in a box (cont'd): number of white balls not known

- However: Consider now a situation in which the value of N is fixed by I , but **the value of K is not**. (Otherwise the experiment is as before: the ball that was drawn is put back into the box, and the contents of the box are shaken thoroughly.)
- **This makes K** , whose value is 'fixed but unknown', **a random variable** in a Bayesian problem formulation.
- Assigning numerical values to $P(K = k | I)$, $1 \leq k \leq N$, will then correspond to my (or your) uncertainty ('degree of belief') about the correctness of each of the events $\{K = k\}$.



Balls in a box (cont'd): distinguishing between physical and logical independence

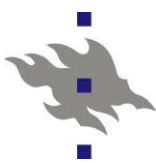
- According to the 'law of total probability' therefore, for any $i \geq 1$,

$$\begin{aligned} P(X_i | I) &= E(P(X_i | K, I) | I) \\ &= \sum_k P(K = k | I) P(X_i | K = k, I). \end{aligned}$$

- But now these probabilities cannot be multiplied to give probability $P(X_1, X_2, \dots, X_i | I)!$

- The chain multiplication rule still holds (it holds always!), but conditional independence is lost:

$$P(X_i | I, X_1, X_2, \dots, X_{i-1}) \neq P(X_i | I), i \geq 1. .$$



Balls in a box (cont'd): distinguishing between physical and logical independence

- The consecutive draws from the box are still – to a good approximation – **physically independent** of each other, but not **logically independent**.
- The outcome of any $\{X_1, X_2, \dots, X_{j-1}\}$ will contain information on likely values of K , and will thereby – if observed - influence what values should be assigned to the probabilities for $\{X_j \text{ is 'white'}\}$ and $\{X_j \text{ is 'yellow'}\}$.



Balls in a box (cont'd): considering joint distribution

- Instead, we get the following

$$P(X_1, X_2, \dots, X_i | I)$$

$$= E(P(X_1, X_2, \dots, X_i | K, I) | I)$$

$$= \sum_k P(K = k | I) P(X_1, X_2, \dots, X_i | K = k, I)$$

$$= \sum_k P(K = k | I) \{ (k/N)^{\#\{\text{white balls in } i \text{ draws}\}} [1 - (k/N)]^{\#\{\text{yellow balls in } i \text{ draws}\}} \},$$

where we have used, inside the sum, the previously derived result for $P(X_1, X_2, \dots, X_i | K, I)$, i.e., corresponding to the situation in which the value of K is known.

- Technically, this is 'mixing' according (or taking an expectation with respect to) the 'prior' probabilities $\{P(K = k | I): 1 \leq k \leq N\}$.



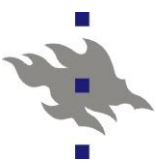
Balls in a box (cont'd): considering joint distribution

- Remember this result when we later consider 'exchangeability' and deFinetti's representation theorem!



Probabilistic inference: from observed data to unknown parameters

- **Intuitively obvious:** By keeping track on the observed values of $X_1, X_2, \dots, X_i, i \geq 1$, we can learn something about the unknown correct value of K .
- **New question:** If we can in this way learn something about the unknown correct value of K , is there some systematic way in which this could be done?



Probabilistic inference: from observed data to unknown parameters

- It could be thought of as providing an example of reversing the direction of the reasoning from the usual 'from parameters to observations', to 'from observations to parameters'.
- Such 'inverse probability' would be a form of [Statistical Inference](#).
- And yes, there is such a systematic way: Bayes' formula!



Balls in a box (cont'd): from observed data to unknown parameters

- The task is to evaluate conditional probabilities of events $\{K = k\}$, given the observations (data) X_1, X_2, \dots, X_i , i.e. probabilities of the form $P(K = k | X_1, X_2, \dots, X_i, I)$.
- By applying the chain multiplication rule twice, in both directions, we get the identity

$$P(K = k, X_1, X_2, \dots, X_i | I)$$

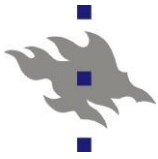
$$= P(K = k | I) P(X_1, X_2, \dots, X_i | K = k, I) \quad \text{chain rule one way}$$

$$= P(X_1, X_2, \dots, X_i | I) P(K = k | X_1, X_2, \dots, X_i, I), \quad \text{...and another way}$$

so that

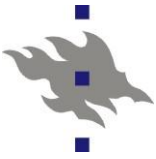
$$P(K = k | X_1, X_2, \dots, X_i, I)$$

$$= P(K = k | I) P(X_1, X_2, \dots, X_i | K = k, I) [P(X_1, X_2, \dots, X_i | I)]^{-1}.$$



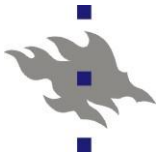
Balls in a box (cont'd): Bayes' formula

- This is **Bayes' formula**.
- By writing $(X_1, X_2, \dots, X_i) = \text{'data'}$, it can be stated simply as
$$P(K = k \mid \text{data}, I)$$
$$= P(K = k \mid I) P(\text{data} \mid K = k, I) [P(\text{data} \mid I)]^{-1}$$
$$\propto P(K = k \mid I) P(\text{data} \mid K = k, I),$$
where ' \propto ' means proportionality in k .



Balls in a box (cont'd): Bayes' formula

- The value of the **constant factor** $P(\text{data} / I)$ in the denominator of Bayes' formula can be obtained 'afterwards' by a simple summation, over the values of k , of the terms $P(K = k / I) P(\text{data} / K = k, I)$ appearing in the numerator.



Bayes' formula

- Using the terminology

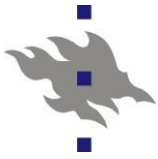
$P(K = k | I) = \text{'prior'}$

$P(\text{data} | K = k, I) = \text{'likelihood'}$

$P(K = k | \text{data}, I) = \text{'posterior'}$,

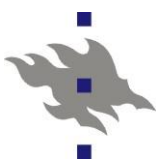
we can write Bayes' formula simply as

'posterior' \propto 'prior' x 'likelihood'



Bayes' formula

- **Stated in words:** By using the information contained in the **data**, as provided by the corresponding **likelihood**, the distribution expressing (**prior**) uncertainty about the true value of K **has been updated into** a corresponding **posterior** distribution.
- Thus Bayesian statistical inference can be viewed as forming a framework, based on probability calculus, for **learning from data**.
- This aspect has received considerable attention in the **Artificial Intelligence** and **Machine Learning** communities, and the corresponding recent literature.



Balls in a box (cont'd): Bayesian inference and prediction

- An interesting aspect in the Bayesian statistical framework is its direct way of leading to probabilistic predictions of future observations.
- Considering the 'Balls in the box' –example, we might be interested in direct evaluation of **predictive probabilities** of the form $P(X_{i+1} / X_1, X_2, \dots, X_i, I)$.
- There is a 'closed form solution' for this problem!



Balls in a box (cont'd): Bayesian inference and prediction

- Writing again $(X_1, X_2, \dots, X_i) = \text{'data'}$, we get

$$P(X_{i+1} \text{ is 'white' } | \text{ data, } I)$$

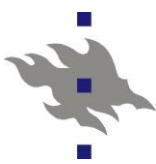
$$= E(P(X_{i+1} \text{ is 'white' } | K, \text{ data, } I) | \text{ data, } I) \text{ by (un)conditioning}$$

$$= E(P(X_{i+1} \text{ is 'white' } | K, I) | \text{ data, } I) \quad \text{conditional independence}$$

$$= (K / N | \text{ data, } I)$$

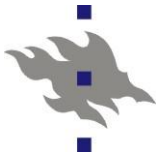
$$= E(K | \text{ data, } I) / N.$$

- In other words, the evaluation $P(X_{i+1} \text{ is 'white' } | K, I) = K / N$, which applies when the value of K is known, is replaced in the prediction by (posterior expectation of K)/ N .



Extensions: continuous variables, multivariate distributions ...

- The probabilistic structure of the 'Balls in a box' –example remains valid in important extensions.
- When considering **continuous variables**, the point-masses appearing of the discrete distributions are changed into **probability density functions**, and the sums appearing in the formulas will be replaced by corresponding **integrals** (in the parameter space).
- In the **multivariate** case (vector parameters) possible redundant parameter coordinates are 'integrated out' from the joint posterior distribution. (Compare this with how nuisance parameters are handled in frequentist inference by maximization, profile likelihood, etc.)



Practical implementation

- **Determining the posterior** in a closed analytic form is possible in some special cases.
- They are restricted to situations in which the prior and the likelihood belong to so-called **conjugate distribution families**: the posterior belongs to the same class of distributions as the prior, but with parameter values updated from data.
- In general, numerical methods leading to approximate solutions are needed (e.g. WinBUGS/OpenBUGS for Monte Carlo approximation).

Table A.1 Continuous distributions		
Distribution	Notation	Parameters
Uniform	$\theta \sim U(a, b)$ $p(\theta) = U(\theta a, b)$	boundaries a, b with $b > a$
Normal	$\theta \sim N(\mu, \sigma^2)$ $p(\theta) = N(\theta \mu, \sigma^2)$	location μ scale $\sigma > 0$
Multivariate normal	$\theta \sim N(\mu, \Sigma)$ $p(\theta) = N(\theta \mu, \Sigma)$ (implicit dimension d)	symmetric, pos. definite, $d \times d$ variance matrix Σ
Gamma	$\theta \sim \text{Gamma}(\alpha, \beta)$ $p(\theta) = \text{Gamma}(\theta \alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$
Inverse-gamma	$\theta \sim \text{Inv-gamma}(\alpha, \beta)$ $p(\theta) = \text{Inv-gamma}(\theta \alpha, \beta)$	shape $\alpha > 0$ scale $\beta > 0$
Chi-square	$\theta \sim \chi^2_\nu$ $p(\theta) = \chi^2_\nu(\theta)$	degrees of freedom $\nu > 0$
Inverse-chi-square	$\theta \sim \text{Inv-}\chi^2_\nu$ $p(\theta) = \text{Inv-}\chi^2_\nu(\theta)$	degrees of freedom $\nu > 0$
Scaled inverse-chi-square	$\theta \sim \text{Inv-}\chi^2(\nu, s^2)$ $p(\theta) = \text{Inv-}\chi^2(\theta \nu, s^2)$	degrees of freedom $\nu > 0$ scale $s > 0$
Exponential	$\theta \sim \text{Expon}(\beta)$ $p(\theta) = \text{Expon}(\theta \beta)$	inverse scale $\beta > 0$
Wishart	$W \sim \text{Wishart}_\nu(S)$ $p(W) = \text{Wishart}_\nu(W S)$ (implicit dimension $k \times k$)	degrees of freedom ν symmetric, pos. definite $k \times k$ scale matrix S
Inverse-Wishart	$W \sim \text{Inv-Wishart}_\nu(S^{-1})$ $p(W) = \text{Inv-Wishart}_\nu(W S^{-1})$ (implicit dimension $k \times k$)	degrees of freedom ν symmetric, pos. definite $k \times k$ scale matrix S

Density function	Mean, variance, and mode
$p(\theta) = \frac{1}{b-a}, \theta \in [a, b]$	$E(\theta) = \frac{a+b}{2}, \text{var}(\theta) = \frac{(b-a)^2}{12}$ no mode
$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)^2\right)$	$E(\theta) = \mu, \text{var}(\theta) = \sigma^2$ $\text{mode}(\theta) = \mu$
$p(\theta) = (2\pi)^{-d/2} \Sigma ^{-1/2} \times \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu)\right)$	$E(\theta) = \mu, \text{var}(\theta) = \Sigma$ $\text{mode}(\theta) = \mu$
$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \theta > 0$	$E(\theta) = \frac{\alpha}{\beta}$ $\text{var}(\theta) = \frac{\alpha}{\beta^2}$ $\text{mode}(\theta) = \frac{\alpha-1}{\beta}, \text{ for } \alpha \geq 1$
$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \theta > 0$	$E(\theta) = \frac{\beta}{\alpha-1}, \text{ for } \alpha > 1$ $\text{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$ $\text{mode}(\theta) = \frac{\beta}{\alpha+1}$
$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{\nu/2-1} e^{-\theta/2}, \theta > 0$ same as $\text{Gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2})$	$E(\theta) = \nu, \text{var}(\theta) = 2\nu$ $\text{mode}(\theta) = \nu-2, \text{ for } \nu \geq 2$
$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{-(\nu/2+1)} e^{-1/(2\theta)}, \theta > 0$ same as $\text{Inv-gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2})$	$E(\theta) = \frac{1}{\nu-2}, \text{ for } \nu > 2$ $\text{var}(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}, \nu > 4$ $\text{mode}(\theta) = \frac{1}{\nu+2}$
$p(\theta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^\nu \theta^{-(\nu/2+1)} e^{-\nu s^2/(2\theta)}, \theta > 0$ same as $\text{Inv-gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2} s^2)$	$E(\theta) = \frac{\nu}{\nu-2} s^2$ $\text{var}(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4$ $\text{mode}(\theta) = \frac{\nu}{\nu+2} s^2$
$p(\theta) = \beta e^{-\beta\theta}, \theta > 0$ same as $\text{Gamma}(\alpha = 1, \beta)$	$E(\theta) = \frac{1}{\beta}, \text{var}(\theta) = \frac{1}{\beta^2}$ $\text{mode}(\theta) = 0$
$p(W) = \left(2^{\nu k/2} \pi^{k(k-1)/4} \prod_{i=1}^k \Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \times S ^{-\nu/2} W ^{-(\nu-k-1)/2} \times \exp\left(-\frac{1}{2} \text{tr}(S^{-1}W)\right), W \text{ pos. definite}$	$E(W) = \nu S$
$p(W) = \left(2^{\nu k/2} \pi^{k(k-1)/4} \prod_{i=1}^k \Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \times S ^{-\nu/2} W ^{-(\nu+k+1)/2} \times \exp\left(-\frac{1}{2} \text{tr}(SW^{-1})\right), W \text{ pos. definite}$	$E(W) = (\nu - k - 1)^{-1} S$

Table A.1 Continuous distributions *continued*

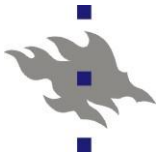
Distribution	Notation	Parameters
Student- <i>t</i>	$\theta \sim t_\nu(\mu, \sigma^2)$ $p(\theta) = t_\nu(\theta \mu, \sigma^2)$ t_ν is short for $t_\nu(0, 1)$	degrees of freedom $\nu > 0$ location μ scale $\sigma > 0$
Multivariate Student- <i>t</i>	$\theta \sim t_\nu(\mu, \Sigma)$ $p(\theta) = t_\nu(\theta \mu, \Sigma)$ (implicit dimension d)	degrees of freedom $\nu > 0$ location $\mu = (\mu_1, \dots, \mu_d)$ symmetric, pos. definite $d \times d$ scale matrix Σ
Beta	$\theta \sim \text{Beta}(\alpha, \beta)$ $p(\theta) = \text{Beta}(\theta \alpha, \beta)$	'prior sample sizes' $\alpha > 0, \beta > 0$
Dirichlet	$\theta \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ $p(\theta) = \text{Dirichlet}(\theta \alpha_1, \dots, \alpha_k)$	'prior sample sizes' $\alpha_j > 0; \alpha_0 \equiv \sum_{j=1}^k \alpha_j$

Table A.2 Discrete distributions

Distribution	Notation	Parameters
Poisson	$\theta \sim \text{Poisson}(\lambda)$ $p(\theta) = \text{Poisson}(\theta \lambda)$	'rate' $\lambda > 0$
Binomial	$\theta \sim \text{Bin}(n, p)$ $p(\theta) = \text{Bin}(\theta n, p)$	'sample size' n (positive integer) 'probability' $p \in [0, 1]$
Multinomial	$\theta \sim \text{Multin}(n; p_1, \dots, p_k)$ $p(\theta) = \text{Multin}(\theta n; p_1, \dots, p_k)$	'sample size' n (positive integer) 'probabilities' $p_j \in [0, 1];$ $\sum_{j=1}^k p_j = 1$
Negative binomial	$\theta \sim \text{Neg-bin}(\alpha, \beta)$ $p(\theta) = \text{Neg-bin}(\theta \alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$
Beta-binomial	$\theta \sim \text{Beta-bin}(n, \alpha, \beta)$ $p(\theta) = \text{Beta-bin}(\theta n, \alpha, \beta)$	'sample size' n (positive integer) 'prior sample sizes' $\alpha > 0, \beta > 0$

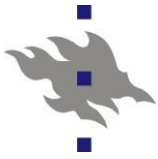
Density function	Mean, variance, and mode
$p(\theta) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} (1 + \frac{1}{\nu}(\frac{\theta-\mu}{\sigma})^2)^{-(\nu+1)/2}$	$E(\theta) = \mu$, for $\nu > 1$ $\text{var}(\theta) = \frac{\nu}{\nu-2}\sigma^2$, for $\nu > 2$ $\text{mode}(\theta) = \mu$
$p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2} \Sigma ^{-1/2}} \times (1 + \frac{1}{\nu}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu))^{-(\nu+d)/2}$	$E(\theta) = \mu$, for $\nu > 1$ $\text{var}(\theta) = \frac{\nu}{\nu-2}\Sigma$, for $\nu > 2$ $\text{mode}(\theta) = \mu$
$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\theta \in [0, 1]$	$E(\theta) = \frac{\alpha}{\alpha+\beta}$ $\text{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ $\text{mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$
$p(\theta) = \frac{\Gamma(\alpha_1+\dots+\alpha_k)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$ $\theta_1, \dots, \theta_k \geq 0; \sum_{j=1}^k \theta_j = 1$	$E(\theta_j) = \frac{\alpha_j}{\alpha_0}$ $\text{var}(\theta_j) = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$ $\text{cov}(\theta_i, \theta_j) = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$ $\text{mode}(\theta_j) = \frac{\alpha_j-1}{\alpha_0-k}$

Density function	Mean, variance, and mode
$p(\theta) = \frac{1}{\theta!} \lambda^\theta \exp(-\lambda)$ $\theta = 0, 1, 2, \dots$	$E(\theta) = \lambda$, $\text{var}(\theta) = \lambda$ $\text{mode}(\theta) = \lfloor \lambda \rfloor$
$p(\theta) = \binom{n}{\theta} p^\theta (1-p)^{n-\theta}$ $\theta = 0, 1, 2, \dots, n$	$E(\theta) = np$ $\text{var}(\theta) = np(1-p)$ $\text{mode}(\theta) = \lfloor (n+1)p \rfloor$
$p(\theta) = \binom{n}{\theta_1, \theta_2, \dots, \theta_k} p_1^{\theta_1} \dots p_k^{\theta_k}$ $\theta_j = 0, 1, 2, \dots, n; \sum_{j=1}^k \theta_j = n$	$E(\theta_j) = np_j$ $\text{var}(\theta_j) = np_j(1-p_j)$ $\text{cov}(\theta_i, \theta_j) = -np_i p_j$
$p(\theta) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^\theta$ $\theta = 0, 1, 2, \dots$	$E(\theta) = \frac{\alpha}{\beta}$ $\text{var}(\theta) = \frac{\alpha}{\beta^2}(\beta+1)$
$p(\theta) = \frac{\Gamma(n+1)}{\Gamma(\theta+1)\Gamma(n-\theta+1)} \frac{\Gamma(a+\theta)\Gamma(n+b-\theta)}{\Gamma(a+b+n)} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ $\theta = 0, 1, 2, \dots, n$	$E(\theta) = n \frac{\alpha}{\alpha+\beta}$ $\text{var}(\theta) = n \frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$



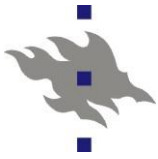
Finding answers to practical problems ...

- The Bayesian approach can be used for finding direct answers to questions such as: "Given the existing background knowledge and the evidence provided by the data, what is the probability that Treatment 1 is better than Treatment 2?"
- The answer is often given by evaluating a posterior probability of the form $P(\theta_1 > \theta_2 \mid \text{data}, I)$, where θ_1 and θ_2 represent systematic treatment effects, and the data have been collected from an experiment designed and carried out for such a purpose.
- The probability is computed by an integration of the posterior density over the set $\{(\theta_1, \theta_2): \theta_1 > \theta_2\}$.



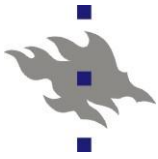
Finding answers to practical problems ...

- The same device can also be used for computing posterior probabilities for 'null hypotheses', of the form $P(\theta = \theta_0 | \text{data}, I)$.
- This is what many people – erroneously – believe the (frequentist) p-values to be. (Thus they are being 'Bayesians' – because they assign probabilities to parameter values - but do not usually themselves realize this.)
- Likewise, one can consider posterior probabilities of the form $P(c_1 < \theta < c_2 | \text{data}, I)$, where the constants c_1 and c_2 may – or may not – be computed from the observed data values. Again, this is how many people (incorrectly) interpret the meaning of their computed (frequentist) confidence intervals.



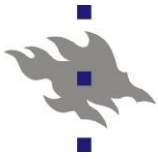
Finding answers to practical problems ...

- Answers formulated in terms of probabilities assigned to model parameters can be difficult to understand. This is because the meaning of such parameters is often quite abstract.
- Therefore it may be a good idea to summarize the results from the statistical analysis in **predictive distributions** of the form $P(X_{i+1} \in A / X_1, X_2, \dots, X_i, I)$, where X_1, X_2, \dots, X_i are considered as 'data' and X_{i+1} is a (perhaps only hypothetical) future response variable that was to be predicted.
- Think about weather prediction!



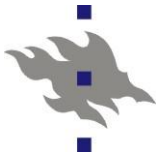
Finding answers to practical problems ...

- The computation of predictive probabilities involves an integration with respect to the posterior. In practice this requires numerical Monte Carlo simulations, which however can be carried out jointly with the estimation of the model parameters ('data augmentation').



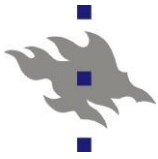
Notes on statistical modeling

- The 'Balls in a box' -example had the advantage that the 'parameter' K had an obvious concrete meaning.
- Therefore also the prior and posterior probabilities assigned for different alternatives $\{K = k\}$ could be understood in an intuitive way.



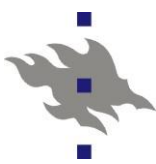
Notes on statistical modeling

- The situation is rather different if we think of commonly used parametric distributions such as, for example, the normal distribution $N(\mu, \sigma^2)$, where the interpretation of the parameters μ and σ^2 is provided by a reference to an infinite population.
- Such populations do not exist in reality, so we really cannot sample from them!



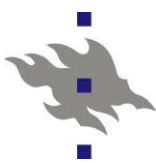
Notes on statistical modeling

- Neither do statistical models 'generate data' (except in computer simulations)!
- Rather, models are rough descriptions of the considered problems, formulated in the technical terms offered by probability calculus, which then allow for an inductive way of learning from data.
- Here is another way of looking at the situation ...



Exchangeability and de Finetti's representation theorem

- In frequentist statistical inference it is common to assume that the observations made from different individuals are 'independent and identically distributed', abbreviated as i.i.d.
- 'Identically distributed', for a Bayesian, means that his / her prior knowledge (before making an observation) is the same on all individuals. (This can also be a convention which is made for the sake of modeling.)



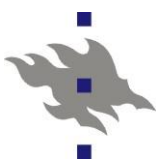
Exchangeability and de Finetti's representation theorem

- Recall from before: the observations may well be **physically independent**, but **not logically independent** - as otherwise there would not be any possibility of learning across the individuals. Statistics would be impossible!



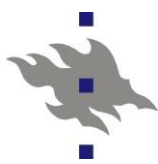
Exchangeability and de Finetti's representation theorem

- This status of information is in Bayesian inference expressed by the following **exchangeability postulate**: the joint probability $P(X_1, X_2, \dots, X_i | I)$ remains the same for all permutations of the variables (X_1, X_2, \dots, X_i) .
- Clearly then $P(X_i | I) = P(X_j | I)$ for $i \neq j$, but it does **not** say that, for example, $P(X_i, X_j | I) = P(X_j | I) P(X_i | I)$.
- Think about **shaking a drawing pin in a glass jar**: Let $X_1 = 1$ if the pin lands 'on its back', and $X_1 = 0$ if it lands 'sideways'. Repeat the experiment i times! Would you say that the sequence X_1, X_2, \dots, X_i is exchangeable?



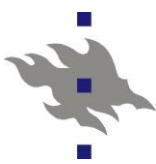
Exchangeability and de Finetti's representation theorem

- Think about shaking a drawing pin in a glass jar: Let $X_1 = 1$ if the pin lands 'on its back', and $X_1 = 0$ if it lands 'sideways'. Repeat the experiment i times! Would you say that the sequence X_1, X_2, \dots, X_i is exchangeable?
- Yes! And, in principle, the experiment could be carried out any number of times, a situation called 'infinite exchangeability'.



Exchangeability and de Finetti's representation theorem

- A frequentist statistical model for describing this situation would be: 'i.i.d. Bernoulli experiments, with an unknown probability θ for success'.



Exchangeability and de Finetti's representation theorem

- The Bayesian counterpart of this is the integral expression:

$$P(X_1, X_2, \dots, X_i | I) \\ = \int \theta^{\#\{\text{times lands on its back in } i \text{ trials}\}} (1 - \theta)^{\#\{\text{times lands sideways in } i \text{ trials}\}} p(\theta) d\theta.$$

- This result, due to de Finetti, looks like we had assumed independent Bernoulli trials with $\theta =$ 'probability of success', and had then taken an expectation with respect to a density $p(\theta)$.



Exchangeability and de Finetti's representation theorem

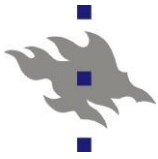
- Formally, it corresponds exactly to the result

$$P(X_1, X_2, \dots, X_i | I)$$

$$= \sum_k P(K = k | I) \left\{ \left(\frac{k}{N}\right)^{\#\{\text{white balls in } i \text{ draws}\}} \left[1 - \left(\frac{k}{N}\right)\right]^{\#\{\text{yellow balls in } i \text{ draws}\}} \right\},$$

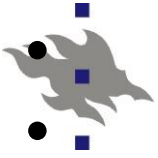
which we had derived in the 'Balls in a box' -example, by first specifying probabilities $P(\text{'the colour is white'} | K, I) = K/N$ and then assuming conditional independence given K .

- It is important to note that, if the 'infinite exchangeability' property is postulated, then the 'prior' $p(\theta)$, and the probabilities $P(K = k | I)$ in the 'Balls in a box' -example, are uniquely determined by the joint distributions $P(X_1, X_2, \dots, X_i | I)$, $i \geq 1$. Looking from this perspective, the existence of a prior – a red herring for many frequentists - should not be a problem.



Notes on statistical modeling

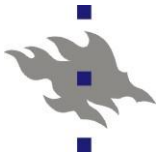
- **The choice of the statistical model**, i.e., of both the prior $P(\theta|I)$ and the likelihood $P(X|\theta, I)$, is a decision which is based on the considered problem context, and often in practice also on convenience or convention.
- As such, it is subject to debate! Models are probabilistic expressions arising from your background knowledge and the scientific assumptions that you make.
- Different assumptions naturally lead to different results.
- You should explain your choices! (The rest is just probability calculus, often combined with approximate numerical evaluation of the probabilities).



Notes on statistical modeling

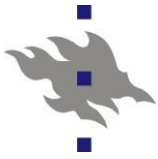
"All models are wrong but some are useful"

George Box (1919-2013)



Take home points ...

- Bayesian methods seem to be natural and useful particularly in areas where frequency interpretation of probability seems artificial.
- They offer greater flexibility in the modeling, in part, because of the possibility to incorporate existing prior knowledge into the model in an explicit way, but also because of the less stringent requirements for parameter identifiability.
- An additional bonus is that the methods are firmly anchored in the principles and rules of probability calculus.



Take home points ...

- Bayesian statistics is fun ... try it out!
- But remember: A Bayesian model is a formal expression of your thoughts. So you need to think carefully ...
- A wonderful book:
Sharon Bertsch McGrayne: "The Theory that Would Not Die",
Yale University Press 2011