Bayesian Inference

William Kraft

Farum Math Camp 2024

Contents

0.1 What is probability?

Definition 0.1 (Outcome)**.** An outcome is a possible result form an experiment. They are mutually exclusive, and contain all relevant data about the result.

Definition 0.2 (Sample Space)**.** The sample space, denoted Ω, of an experiment is the set of all possible outcomes.

Definition 0.3 (Event)**.** An event is a subset of the sample space. We say it occurs if the outcome is an element in the event.

Definition 0.4.

- The complementary event, A^* of A is the complement of A
- The union of the events A and B is the set $A \cup B$
- The intersection of the events A and B is the set $A \cap B$
- The events A and B are mutually exclusive if they are disjoint, $A \cap B = \emptyset$

OBS! 0.5. *The terms from probability theory will be used interchangeably with their set theoretic counterparts.*

Definition 0.6 (Probability measure)**.** A probability measure $P(\cdot)$ is a function from an event to a real number which follows "Kolmogorov's System of Axioms":

Axiom 1: If A is any event, then $0 \leq P(A) \leq 1$

Axiom 2: If Ω is the entire sample space, then $P(\Omega) = 1$

Axiom 3: If $A, B, \ldots$ is a finite or infinite sequence of mutually exclusive events, then $P(A \cup B \ldots) = P(A) + P(B) + \ldots$

There are often multiple reasonable probability measures, but this allows us to easily work with any of them.

Example 0.7 (Frequency probability measure)**.** Given a repeatable experiment we define the frequency probability measure is the relative frequency of the event.

Example 0.8 (Other probability measure)**.** Often probability measures can be subjective. For example, what was the probability of a soccer match ending the way it did? We can not recreate it exactly, and so there is no way to use the frequency probability measure. Instead it would depend on who you asked. As long as their probability concurs with Kolmogorov's axioms, we can work with it!

Corollary 0.9 (Basic Rules)**.** From the axioms it can be proven that

- $P(A^*) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A) + P(B) = P(A \cup B) - P(A \cap B)$

Definition 0.10 (Discrete and continuous sample space)**.** If a sample space is either finite or countably infinite it is discrete. Otherwise, it is continuous.

Example 0.11. The different outcomes of a 6 sided dice are discrete while its exact landing spot is continuous. Note that the same experiment may have different sample spaces depending on what you are studying.

1 Probability Theory

1.1 Motivation

The world is full of systems where no model we can construct, either due to complexity or lack of data, could produce accurate results. Probability theory instead attempts to find patterns within the chaos, making us able to still study these random, stochastic, systems.

1.2 Conditional Probability

Definition 1.1 (Conditional probability)**.** The probability of A **given** (an outcome of) B is defined as

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

OBS! 1.2. *This definition of conditional probability is quite natural as it preserves relative probability, and gives the new sample space probability 1.*

OBS! 1.3. *The probability measure $P(\cdot|B)$ also satisfies Kolmogorov's axioms.*

Example 1.4. The probability of having rolled a 5 on a 6 sided die **given** that the result was odd, is $\frac{1}{3}$.

Theorem 1.5 (Total Probability Theorem)**.** Let $B_1, \ldots, B_n$ be mutually exclusive events and their union be the sample space. If A is an event, then

$$P(A) = P(\cup_{k=1}^{n} A \cap B_k) = \sum_{k=1}^{n} P(A \cap B_k) = \sum_{k=1}^{n} P(A|B_k) \cdot P(B_k)$$

Theorem 1.6 (Bayes' Theorem)**.** Let $A_1, \ldots, A_n$ be mutually exclusive events and their union be the sample space. If B is an event, then

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{k=1}^{n} P(B|A_k) \cdot P(A_k)}$$

Definition 1.7 (Independent events)**.** The events A and B are independent, $A \perp B$, if

$$P(A \cap B) = P(A)P(B).$$

OBS! 1.8. *This does not imply, for example, that they are independent given C.*

1.3 Random variables

Definition 1.9 (Random variables)**.** A random variable (rv) is a function from our sample space to $\mathbb{R}$.

Example 1.10. Consider the experiment of throwing a dart at a dartboard. Let X be the number of centimeters from the bullseye. Then X is a random variable. We could have many random variables stemming from the same experiment.

Example 1.11. Let X be the price of a match (in £), and Y be the price of a can of gasoline. Then the cost of both is $X + Y$. If Z is the number of times you do this purchase, the total cost is $(X + Y) \cdot Z$.

OBS! 1.12. *We often write the shorthand of P(condition) to denote $P(A)$ where A is the set of all outcomes satisfying the condition. For example $P(X = 1)$ where X is the result of a 6-sided die.*

Definition 1.13 (Outcome of a random variable)**.** An outcome of a random variable is a value obtained by running the experiment. Multiple outcomes of a known random variable are independent.

Example 1.14. Let X be the number of leafs on a clover when you pick it up. If i pick up a lucky clover (4 leafs), it means the outcome of picking that clover was $x = 4$.

1.4 Distributions

Definition 1.15. The distribution function (cumulative distribution function) of a random variable X is defined as

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

Note that X is the random variable, while x is an argument to the function.

Corollary 1.16. A distribution function $F_X(x)$ of a random variable X has that

$$F_X(X) \to \begin{cases} 0 \\ 1 \end{cases} \quad \text{when } x \to \begin{cases} -\infty \\ \infty \end{cases} \tag{1}$$

$$F_X(x) \text{ is a non-decreasing function of } x \tag{2}$$

$$F_X(x) \text{ is continuous from the right for any} x \tag{3}$$

Theorem 1.17.

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Definition 1.18 (Probability function)**.** For a discrete random variable X (from a discrete sample space), the probability function is

$$f_X(x) = P(X = x)$$

Theorem 1.19. The sum

$$\sum_{x=-\infty}^{\infty} f_X(x) = 1$$

Definition 1.20 (Density function)**.** Let X be a continuous random variable. If the distribution function can be written as

$$F_X(x) = \int_{-\infty}^{x} f_x(t)dt,$$

then $f_X(x)$ is called the density function (probability density function), and X is a continuous random variable.

OBS! 1.21. *We will often be able to work similarly with continuous and discrete random variables, with the only mayor difference being integration vs summation. As we will focus more on continuous functions in the following lectures, I might exclude the discrete case at times.*

Definition 1.22 (Distribution)**.** A distribution function corresponds to a unique distribution. Two stochastic variables have the same distribution, if they have the same distribution functions.

OBS! 1.23. *This gives us a way to look at the properties of categorising random variables without looking at their corresponding sample spaces.*

Example 1.24. A continuous random variable X has a uniform distribution on the interval (a, b) (written $X \sim U(a, b)$) if

$$F_X(x) = \begin{cases} 0, x < a \\ 1, x > b \\ \frac{x-a}{b-a}, \text{ otherwise} \end{cases}$$

OBS! 1.25. *Both the probability function and density function, also uniquely correspond to a distribution function. We may even say it has the distribution of such a function.*

Example 1.26. A discrete random variable $X \sim \text{Bin}(n, p)$ if

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $0 \leq x \leq n$.

Theorem 1.27. Let X, Y be discrete random variables with $Y = g(X)$ and g invertible. Then

$$f_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = f_X(g^{-1}(y))$$

If g is not invertible, then

$$f_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x \in A} P(X = x) = \sum_{x \in A} f_X(x)$$

where $A = \{x | g(x) = y\}$.

Theorem 1.28. Let X, Y be random variables with $Y = g(X)$ and g strictly increasing. Then

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

If g strictly decreasing

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Theorem 1.29. Let X, Y be continuous random variables with $Y = g(X)$, g continuous and bijective. Then

$$f_Y(y) = F'_Y(y) = \begin{cases} F'_X(g^{-1}(y)) \cdot g^{-1}{}'(y) \\ -F'_X(g^{-1}(y)) \cdot g^{-1}{}'(y) \end{cases} = f_X(g^{-1}(y)) \cdot g^{-1}{}'(y).$$

1.5 Multivariate distributions

OBS! 1.30. *We may without problem choose two or more random variables from a single experiment.*

Definition 1.31. Given the random variables $X_1, \ldots, X_n$, we can define:

- The distribution function
$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1 \wedge \cdots \wedge X_n \leq x_n).$$
- The probability function (given X_i discrete)
$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(x_1, \ldots, x_n).$$
- The density function, f which satisfies: (given X_i continuous)
$$\int_{\mathbb{R}\times\cdots\times\mathbb{R}} \cdots \int F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

Theorem 1.32. Let $\vec{X}, \vec{Y}$ be vectors of random variables with $\vec{Y} = g(\vec{X})$ and g bijective, continuous. Then

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |\det J|$$

where J is the Jacobian of g^{-1}. (Proven using variable substitution for integrals)

Definition 1.33 (Marginal distribution)**.** Let X, Y be two random variables with the distribution function $F_{X,Y}(X, Y)$. Then the marginal distribution of X is

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

With a similar spirit, we can define the marginal probability function as

$$f_X(x) = \sum_{y=-\infty}^{\infty} P_{X,Y}(x, y)$$

and the marginal density function

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Example 1.34. Let X, Y be the results of two 6-sided dice. As they are independent, the probability function would be

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(x) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

where $1 \leq X, Y \leq 6$. The marginal distribution of X would be

$$f_X(x) = \sum_{y=1}^{6} f_{X,Y}(x,y) = \frac{1}{6}$$

Definition 1.35 (Conditional distribution)**.** The probability or distribution function of $X|Y$ is

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

where $f_Y(y)$ is the marginal density function of $f_{X,Y}(x,y)$.

OBS! 1.36. *Let $\Delta y > 0$ and $I = [y, y + \Delta Y]$*

$$\begin{aligned} P(X < x|Y \in I) &= \frac{P(X < x \wedge Y \in I)}{P(Y \in I)} \\ &= \frac{\int_y^{y+\Delta y} \int_{-\infty}^{x} f_{X,Y}(X,Y)dxdy}{\int_y^{y+\Delta y} f_Y(y)dy}. \end{aligned}$$

If we let $\Delta y \to 0$, using the mean value theorem for integrals we get

$$\begin{aligned} P(X < x|Y \in I) &= \frac{\Delta y \int_{-\infty}^{x} f_{X,Y}(X,Y)dx}{\Delta y f_Y(y)} \\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(X,Y)dx}{f_Y(y)} \end{aligned}$$

This would correspond to a distribution function of X given $Y = y$. Despite Y never being exactly Y, the same applies to the numerator, and the limit is defined. By integrating we would intstead get the density funciton of X given $Y = y$, as defined in the definition.

Example 1.37. Let X, Y be result of two independent dice, and $Z = X + Y$. When given the result of X, we have

$$f_{Z|X}(z|x) = f_Y(z - x).$$

1.6 Exercise Problems

Problem 1.38. The probability of having a certain disease is 0.01, the probability of getting a positive test result is 0.9 if you have the disease, and 0.01 if you dont. Given a positive test result, what is the probability of you having the disease?

Problem 1.39. Let X be a continuous variable which is uniformly distributed on (a, b). What is the distribution of $Y = \frac{X}{2} + 3$?

Problem 1.40. What is the probability function of the sum of one 6-sided die and one 8-sided die?

Problem 1.41. You want to burn up a few documents in your backyard. The time in minutes, T, it takes for all documents to have been burnt up is distributed as $T \sim Exp(r)$ where R is the burn rate of the gasoline you choose. As your time is

limited, the chance of you choosing the fast burning new gasoline with burn rate 10 is 30% and otherwise you choose the old with burn rate 2. What is the chance of you having burnt everything within 5 minutes?

Problem 1.42. A friend picked randomly between a 4, 6, and 8 sided die and only told you they rolled a 3. What is the probability they rolled the 4-sided die? If they roll again with the same die, what is the probability they roll a 5?

Problem 1.43. Given the distribution of $X|\Theta$ and Θ, what is the distribution of X? What is the distribution of $\Theta|X$?

2 Bayesian Inference

2.1 Situations

Using probability theory, we can calculate the chances of events given the distribution. While the type of distribution can often be inferred, finding the hidden parameter is a very challenging task.

2.2 Philosophy

Bayesian inference is based around the idea that one never encounters a situation without previous knowledge! For example, you may know, that the dice is likely fair, and otherwise, probably loaded in favor of higher numbers. This information is as valuable as any other, and should be combined with your new observations. Using this, we can calculate the chances of you rolling quintuple sixes... **given** your previous knowledge, and the fact you have already rolled quadruple sixes. This is achieved by not considering the probabilities as constants, but as random variables themselves.

2.3 Prior distribution

Definition 2.1. Let X be a random variable with a distribution depending on the outcome θ of the random variable Θ. The distribution of Θ is called the *prior* distribution, while the distribution of $X|\Theta$ is called the data distribution.

The prior distribution represents your prior knowledge of the parameter, while the data distribution represents the distribution of X given you know the outcome of Θ which the distribution depends on.

Example 2.2. Let the result of a coin flip (0 if tails, 1 if heads) be $X|\Theta \sim \text{Ber}(\theta)$ where θ is the chance of flipping heads. When seeing a new coin its properties affect the chances of flipping heads, and might be distributed as $\Theta \sim \text{Beta}(2, 2)$. The former is the data distribution, the latter the prior distribution.

OBS! 2.3. *Most further calculations will be done only in the continuous case, the discrete case is left as an exercise.*

Definition 2.4 (Prior predictive distribution)**.** The distribution of X is the called the prior predictive distribution. It is what we, before observing more data, would find to be the distribution of X. It is calculated from our prior distribution and data distribution, as the marginal distribution of the simultaneous distribution

$$f_{X,\Theta}(x, \theta) = f_\Theta(\theta) \cdot f_{X|\Theta}(x|\theta).$$

That is,

$$f_X(x) = \int_{-\infty}^{\infty} f_\Theta(\theta) \cdot f_{X|\Theta}(x|\theta) d\theta$$

2.4 Posterior distribution

The goal of Bayesian inference is creating the posterior distribution $\Theta|X = x$ where x is an outcome of the random variable X. That is, using the previous outcomes, we update our belief of what the distribution of Θ may be.

Definition 2.5 (Posterior distribution)**.** The posterior distribution $\Theta|x$ is defined as

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) \cdot f_\Theta(\theta)}{f_X(x)}$$

and is simply an application of Bayes' rule!!!

OBS! 2.6. *Once we have calculated the posterior distribution, we may calculate anything we would ever want to know about the distribution of X.*

OBS! 2.7. *The gamma function is defined as*

$$\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t} dt$$

and has the property $\Gamma(x) \cdot x = \Gamma(x+1)$

Example 2.8. Let $\Theta \sim \text{Gamma}(\alpha, \lambda)$ (the prior distribution) and $X \sim \text{Exp}(\theta)$ (the data distribution). The density functions become

$$\begin{aligned} f_\Theta(\theta) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta} \\ f_{X|\Theta}(x|\theta) &= \theta e^{-\theta x} \end{aligned}$$

where $\theta, x > 0$.

The prior predictive distribution becomes

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{x,\Theta}(x,\theta) d\theta \\ &= \int_0^\infty f_\Theta(\theta) f_{X|\Theta}(x|\Theta) d\theta \\ &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta} \cdot \theta e^{-\theta x} \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^\alpha e^{-\theta(x+\lambda)} d\theta \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} (\lambda + x)^{-\alpha-1} \int_0^\infty u^\alpha e^{-u} du \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} (\lambda + x)^{-\alpha-1} \Gamma(\alpha + 1) \\ &= \alpha \lambda^\alpha (\lambda + x)^{-\alpha-1} \end{aligned}$$

We may then calculate the posterior distribution

$$\begin{aligned} f_{\Theta|X}(x|\theta) &= \frac{f_{x,\Theta}(x,\theta)}{f_X(x)} \\ &= \frac{f_\Theta(\theta) f_{X|\Theta}(x|\Theta)}{f_X(x)} \\ &= \frac{\frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta} \cdot \theta e^{-\theta x}}{\alpha \lambda^\alpha (\lambda + x)^{-\alpha-1}} \\ &= \frac{(\lambda + x)^{\alpha+1}}{\Gamma(\alpha)\alpha} \theta^\alpha e^{\theta(\lambda+x)} \\ &= \frac{(\lambda + x)^{\alpha+1}}{\Gamma(\alpha + 1)} \theta^{\alpha+1-1} e^{\theta(\lambda+x)} \end{aligned}$$

Which we can easily recognise as the gamma distribution, giving us

$$\Theta|X \sim \text{Gamma}(\alpha + 1, \lambda + x)$$

Definition 2.9 (Proportionality)**.** A function $f(x)$ is proportional to a function $g(x)$ if there is a constant $c \neq 0$ such that for all x,

$$f(x) = c \cdot g(x).$$

We write this relationship

$$f(x) \propto g(x).$$

If they are dependent on multiple variables, the relevant one will be given by the context.

Example 2.10. Let $\Theta \sim \text{Beta}(\alpha, \beta)$ and $X|\Theta \sim \text{Bin}(n, \theta)$. The density functions will then be

$$f_\Theta(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}\theta^{\beta-1}$$
$$f_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

We notice that in respect to θ the posterior distribution is proportional to

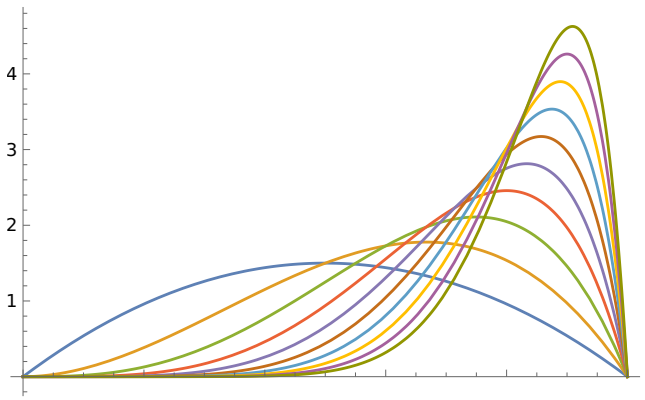
$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{f_{x,\Theta}(x,\theta)}{f_X(x)} \\ &= f_{x,\Theta}(x,\theta) \\ &= f_{x|\Theta}(x|\theta)\cdot f_\Theta(\theta) \\ &\propto \theta^x(1-\theta)^{n-x}\cdot\theta^{\alpha-1}\theta^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}. \end{aligned}$$

We notice that this is proportional to the beta distribution. As the integral of the function is 1, we can then scale the function to satisfy the criterion. We may otherwise notice this is a scaled version of the Beta distribution and scale it to match, giving us

$$\Theta|X \sim \text{Beta}(\alpha + x, \beta + n - x).$$

OBS! 2.11. *Notice that this rule may be used to update the distribution multiple times. These updating rules can often be found in formula sheets, but are derived as above.*

Example 2.12. Let $\Theta \sim \text{Beta}(2,2)$ and $X|\Theta \sim Ber(\theta)$. For example X could be 1 if we flip heads, and 1 if we flip heads. While flipping 5 heads in a row, and updating each time, we notice our distribution gets heavily skewed towards higher probabilities of heads.



If our prior distribution had been less varied, the results would have been showing later.

2.5 Exercise problems

Problem 2.13. Find the updating rule for $X|\Theta \sim \text{Po}(\theta)$ and $\Theta \sim \text{Gamma}(\alpha, \lambda)$ where

$$f_{X|\Theta}(x|\theta) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Problem 2.14. Find the updating rule for $\Theta \sim N(\mu, \sigma_1^2)$ and $X|\Theta \sim N(\theta, \sigma_2^2$ where the $N(\mu, \sigma$ is the normal distribution with density function

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma^2}}.$$

Find the updating rule!

Problem 2.15. Given the posterior distribution, how would you find an interval where Θ has p probability to be within it?

Problem 2.16. If you have the posterior distribution, and the data distribution. How could you calculate the posterior predictive distribution?

3 Applications

By now, we know the distribution of the parameter, and may calculate anything we want to know about the parameter or the distribution!

3.1 Posterior predictive distribution

Definition 3.1 (Posterior predictive distribution)**.** Let Y be a random variable which is identically distributed to X. Given the prior distribution, data distribution and outcomes of X, we may then calculate the simultaneous distribution of Y and Θ, and marginalise it to get the distribution of Y. The latter is called the posterior predictive distribution and is (in the continuous case) calculated as

$$f_{Y|X}(x|\theta) = \int_{-\infty}^{\infty} f_{\Theta|X}(\theta|x) \cdot f_{Y|\Theta}(y|\theta) d\theta.$$

OBS! 3.2. *The posterior predictive distribution is very similar to the prior predictive distribution, but instead of using the prior distribution, it uses the posterior distribution.*

3.2 Credible interval

Definition 3.3 (Credible interval)**.** A $1-\alpha$ credible interval of Θ is an interval I such that $P(\Theta \in I|X) = 1-\alpha$.

Example 3.4 (Continuous Case)**.** A $1-\alpha$ credible set A would have

$$\int_A f_{\Theta|X}(\theta|x) d\theta = 1-\alpha$$

OBS! 3.5. *We often choose a symmetric credible interval, where X is equally likely to be lower or higher than the interval, but may choose it as we please as long as the probability of X falling within it is p.*

OBS! 3.6. *Quite often numerical approximation of integrals is used to find credible intervals.*

3.3 Decisions

Definition 3.7 (Expected Value)**.** The expected value of function g of a continuous random variable X is

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

OBS! 3.8. *The expected value is linear.*

Definition 3.9 (Loss Function)**.** The loss function

$$L\left[\theta, a\right]$$

where is a possible outcome of Θ, and a is an action, is the "loss" which occurs if we take the action a when Θ has the outcome θ.

Definition 3.10 (Bayes Risk)**.** The Bayes risk,

$$E\left[L(\Theta, W(X))\right],$$

is the expected loss when using the decisions rule $W(x)$ to decide your action given observation

Theorem 3.11 (Minimising Bayes Risk)**.** In order to minimise the Bayes risk however, notice that the outcome x of X is already determined, and thus we simply minimize

$$E\left[L(\Theta, W(x))|X\right] = \int_{-\infty}^{\infty} L(\theta, W(x)) \cdot f_{\Theta|X}(\theta|x)d\theta.$$

Example 3.12. Let the action to be taken be, giving an estimate a of θ, and the loss function be $(\theta - a)^2$. We wish to minimize

$$E[L(\Theta, a)|X] = E[(\Theta - a)^2|X] = E[\Theta^2|X] - 2aE[\Theta|X] + a^2.$$

As we can only affect a, we wish to minimize

$$a^2 - 2aE[\Theta|X]$$

which is done by setting $a = E[\Theta|X]$.

3.4 Bayesian Hierarchy

Definition 3.13. If $X|\Theta$, $\Theta|\Phi$ and Φ have known distributions, we call this a Bayesian Hierarchy, where Φ is a super parameter.

Example 3.14. Let X be the result of a chess player during a single match, Θ be their ability depending on that day, and Φ be their skill. Then by seeing results from matches, we can update our belief of their ability that day, and in turn, update our belief of their skill.

3.5 Easier calculations

While Bayesian inference gives a good framework for statistics, it also requires a lot of integrals. All the given examples have been nice in this regard, but one easily encounters situation where the integrals are unfeasible, or even unsolvable. In these cases we may leave them, and when calculating the final values simply approximate them. There are however some methods to make our lives easier.

Definition 3.15 (Conjugating distributions)**.** We call a prior distribution and a data distribution conjugating, if the posterior distribution is in the same family (of the same form) as the prior distribution.

OBS! 3.16. *These are often the nicest cases, and if within reason, we prefer to choose these types of distributions.*

3.6 Exercise Problems

Problem 3.17. What would be the posterior distribution of Φ in the example about a Bayesian Hierarchy?

Problem 3.18. Let $X|\alpha, \beta$, what would be a conjugating distribution of α, β?