Contents

Chapter 1

Introduction to the Bayesian framework

Remark 1. The topics of this block are contained in Chapters 2 and 3 of *Lambert B. (2018) A Student's Guide to Bayesian Statistics, Sage.*
Another reference is *Hoff, P. D. (2009). A first course in Bayesian statistical methods. Springer Science & Business Media. ISBN: 978-0-387-92299-7.* For this part section chapter 1 (1.1 - 1.4) and 2(2.1 - 2.6) while for remarks on important statistical distributions see pages 253-258.

1.1 Introduction to Bayesian reasoning

1.1.1 The concept of event

Definition 1.1.1 (Event)**.** An event is a logical entity which can be either true (T) or false (F).

Important remark 1. We have that:

- the event is something physical, observable, stated by a proposition
- in an experimental situation, after the experiment, it must be possible to verify whether the event has been T or F

Example 1.1.1 (A negative example)**.** The proportion of heads when tossing a coin is not an event.

1.1.2 Probability of an event

Remark 2. Two main idea/interpretation of probability are available.

Definition 1.1.2 (**Objective probability** (*logical* or *frequentist*))**.** Probability is a physical property of the event.

Important remark 2 (Critique). The definition is linked tot he concept of repeatable events but *repeatable events do non exist*: only past experiences under similar conditions do exist.

Definition 1.1.3 (**Subjective probability**)**.** Probability is the measure of plausibility that an individual assigns to an (uncertain) event.
The probability of an event E, for a certain individual (in a certain moment) is the price $P(E) = p$ that he considers right to pay to participate at a bet where he will win 1 if E occurs or 0 if it doesn't.

Remark 3. In this definition, probability:

- is not a physical property of the event, rather a formalization of the beliefs (and information) that the individual possesses about the event;
- can be different for different individuals (thus the term "subjective");

Remark 4. It is important to state that also what is *not observable* may receive a probability

Remark 5. The classical statistician does not accept subjective probability, since the last is not related to the concept of frequency.

1.1.3 Coherence of subjective probability

Definition 1.1.4 (Coherence of subjective probability)**.** A probability assessment about the n events $E_1, E_2, \ldots, E_n$ is said to be *coherent* if no combination of bets on these events allows a sure win (independently on the events E_i, $i = 1, \ldots, n$ that actually occur).

NB: These n events are alternative I think, think a partition ...

Remark 6. Necessary and sufficient condition for coherence of the subjective probability is expressed by the following theorem.

Theorem 1.1.1. *A necessary and sufficient condition for $P(E)$ coherence is that $0 \leq P(E) \leq 1$. In particular, if $P(E) = 0$ the event is impossible, while if $P(E) = 1$ the event is said certain.*

Proof. Let $p = P(E)$ the *price* and let assume to bet S about the occurrence of E . When

NB: per il gioco dobbiamo pagare p per singolo euro di vincita, se scommettiamo su S di vincita dobbiamo pagare pS

- E occurs, the gain obtained by the bet is the wager/win minus the price for the wager itself:
$$W(E) = S - pS = S(1 - p) = S(1 - P(E))$$
- E does not occur, the gain is negative (just the price paid):
$$W(\bar{E}) = -pS = -P(E) \cdot S$$

How choosing p and S in order avoiding to obtain a sure win (positive earning in any case)?

- if $p < 0$ it would be enough to bet a positive wager $S > 0$ to guarantee a sure win.
- if $p > 1$ it would be enough to bet a negative wager $S < 0$ to guarantee the sure win.

Thus it follows that to avoid a sure win it must be that $0 \leq P(E) \leq 1$.
Furthermore:

- if the event E is certain, it's payoff is $W(E) = S(1-p)$: the only way to avoid a sure win is to set $W(E) = 0$, by fixing $P(E) = 1$.
- if E is impossible, $\bar{E}$ is certain, its payoff is $W(\bar{E}) = -pS$: in order to avoid sure wins it has to be $W(\bar{E}) = 0$, from which $p = P(E) = 0$ (case of impossible events).

□

Theorem 1.1.2. *The probability of the union of many events (incompatible when considered in couples) is the sum of their probabilities.*

Proof. Omitted □

1.1.4 The axiomatic Kolmogorov framework

Remark 7. Allows any computations regarding probabilities, exploiting the analogy between probability and measure, and between mathematical expectation and integration according to Lebesgue.
It needs at least the definition of an algebra. or better, of a σ-algebra

Definition 1.1.5 (Algebra)**.** A class A of subsets of Ω is an algebra if:

1. $\Omega \in A$;
2. if $E_i \in A$ then $\bar{E}_i \in A$;
3. *finite additivity*: $\cup_{i=0}^{n} E_i \in A$ for events that are incompatible two by two

Definition 1.1.6 (σ-algebra)**.** Has the same two first properties, but the third consists of complete rather than finite additivity: $\cup_{i=0}^{\infty} E_i \in A$.

Remark 8. Complete additivity cannot be derived from finite additivity with the application of a limit.

1.1.5 Differences between classical/Bayesian statistics

Differences are resumed in table 1.1

1.1.5.1 Classical statistical inference

Inference is based on the distribution of a statistic, that varies in the set of possible sampling; inference relies on the idea the experiment may be repeated in order to obtain such distribution.
A critique to this reasoning is that decisions are taken on the basis of something that never will be observed. (Only one sample, that produces only one data set, is indeed achieved).

1.1.5.2 Likelihood based inference

The likelihood function provides all information contained in the sample and, for that specific sample, it measures the plausibility of the various alternatives for expressing how the phenomenon is.

	Classical statistics	**Bayesian statistics**
Experiment assumptions	Independence	Exchangeable series
Interpretation of probability	Relative frequency: can be applied to events that can be repeated	Degree of credibility: can be applied to unique events and series of events.
Statistical inference based on ...	Sampling distribution: a sampling space has to be specified.	Final/posterior distribution: the initial distribution has to be assigned.
Parameter estimation	Needs a theory of estimation	Needs description/synthesis of the final/posterior distribution.
Role of personal evaluations	The choice of the experiment is needed, as well as the choice of the procedures to adopt. Personal evaluations remain external (they are not quantified: the problem *appears* as dealt in an objective way.)	All knowledge can be formally incorporated in the initial distribution.

Table 1.1: Differences in bayesian vs frequentist inference

Example 1.1.2. The likelihood of the binomial distribution does not contain the binomial coefficient. Such likelihood is the same of the Bernoulli distribution. The binomial coefficient is not a part of the likelihood, since it does no contain the parameter.

1.1.5.3 Bayesian paradigm

Peculiarity of Bayesian inference: the link between likelihood and final distribution, starting from the initial/prior probability distribution.
The *main assumption* is that prior probabilities can be assigned, starting from an initial probability distribution.

1.1.5.4 Critical observations on classical statistics

- Several techniques of classical statistics do not respect the likelihood principle, that states that two experiments give the same information if the corresponding likelihood functions are inductively equivalent, i.e. differ for a multiplicative constant.
- The estimate may depend on how the experiment is developed

1.1.6 Probability distributions vs likelihoods

Example 1.1.3 (Binomial Case: Difference between Probability Distributions and Likelihoods)**.** Consider the throw of 10 coins with a given probability of head θ. Table 1.2 describes the probability of all possible outcomes (obtain

from 0 to 10 heads) using a binomial distribution for some different values of the parameter θ:

- each line contains the probability distribution of the outcomes from 0 to 10 for different values of the parameter (the values of each line sum to 1)
- each column likelihood (it is a function, not a probability distribution) of some values of the parameters for the possible different results (the sum of the probabilities of each column is not 1)

	Y=0	Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	Sum
0	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
0.1	0.349	0.387	0.194	0.057	0.011	0.001	0.000	0.000	0.000	0.000	0.000	1.000
0.2	0.107	0.268	0.302	0.201	0.088	0.026	0.006	0.001	0.000	0.000	0.000	1.000
0.3	0.028	0.121	0.233	0.267	0.200	0.103	0.037	0.009	0.001	0.000	0.000	1.000
0.4	0.006	0.040	0.121	0.215	0.251	0.201	0.111	0.042	0.011	0.002	0.000	1.000
0.5	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001	1.000
0.6	0.000	0.002	0.011	0.042	0.111	0.201	0.251	0.215	0.121	0.040	0.006	1.000
0.7	0.000	0.000	0.001	0.009	0.037	0.103	0.200	0.267	0.233	0.121	0.028	1.000
0.8	0.000	0.000	0.000	0.001	0.006	0.026	0.088	0.201	0.302	0.268	0.107	1.000
0.9	0.000	0.000	0.000	0.000	0.000	0.001	0.011	0.057	0.194	0.387	0.349	1.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000
Sum	1.491	0.829	0.906	0.910	0.909	0.909	0.909	0.910	0.906	0.829	1.491	

Table 1.2: Probability vs likelihood

Important remark 3. The examples concerning events are very simple/intuitive. When probabilities of events are substituted by the probabilities of random variables and the notion of statistical model is introduced, things get a little more complicated.

1.2 Bayes theorem for events

Theorem 1.2.1 (Compound probability theorem)**.** *The conditional probability of one event E_1 given another E_2 can be written as*

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1)P(E_2|E_1)}{P(E_2)}$$

Remark 9. In the case of independence, the relationship becomes even simpler.

Remark 10. Bayes theorem starts from the compound probability theorem by changing the denominator.

Theorem 1.2.2 (Bayes theorem)**.** *Considering a finite partition $\{H_i\}$ $i = 1, \ldots, I$ of the certain event, and an event B with probability $P(B) > 0$.*
The probabilities of B conditional on the H_i are known: $P(B|H_i)$, $i = 1, \ldots, I$.
Then

$$P(H_i|B) = \frac{P(H_i)P(B|H_i)}{P(B)} = \frac{P(H_i)P(B|H_i)}{\sum_{i=1}^{I} P(H_i)P(B|H_i)}.$$

	D	$\bar{D}$
T	0.95	0.02
$\bar{T}$	0.05	0.98
	1	1

Table 1.3: Experiment and conditional distributions

1.2.1 Examples on (discrete) prior for events

1.2.1.1 Rare disease test (asymmetric/informative prior)

Remark 11. If an individual is positive to the medical test for a disease, which is the probability that she/he is sick? Much of the answer depends on the contribution of the prior/*non modifiable* situation

Example 1.2.1 (Rare disease)**.** The event D is suffering from a rare disease, with $P(D) = 0.001$ (so $P(\bar{D}) = 0.999$). The prevalence of the disease $P(D)$ is our prior/starting point, the *state of the world*: it cannot be changed (neither $P(\bar{D})$).
The event T is being positive at a medical test for that disease. Technology can provide probabilities of T that are conditional on being sick or not where hold:

$$P(T|D) + P(\bar{T}|D) = 1$$
$$P(T|\bar{D}) + P(\bar{T}|\bar{D}) = 1$$

An experiment T is performed conditional on D and results are organized in tables like 1.3 where two distributions are available (conditional on the two states of the world) in columns.
Now some naming:

- $P(\bar{T}|D)$ is the probability of false negatives: the disease is present but the test is negative (the individuals are lost for following tests). It is the first type error $\alpha = 0.05$.
- $P(T|D)$ is the *sensitivity* $(1 - \alpha) = 0.95$
- $P(T|\bar{D})$ is the probability of false positives: the result of the test is positive even if the individual has not the disease (second type error) $\beta = 0.02$.
- $P(\bar{T}|\bar{D})$ is the *specificity* (power) $(1 - \beta) = 0.98$.

The quantity $P(D|T)$ (probability of being diseased given a positive test) is computed using Bayes theorem as:

$$P(D|T) = \frac{P(D)P(T|D)}{P(T)}$$

In order to compute the denominator $P(T)$ at (i.e. the probability of being positive to the test, irrespective of the state of the world) the weighted sum of the probabilities of being sick for all possible states of the world has to be computed (weighs are $P(D)$ and $P(\bar{D})$ respectively). In this way the possible

	D	$\bar{D}$
T	0.99	0.005
$\bar{T}$	0.01	0.995
	1	1

Table 1.4: Techology improvement

states of the world have been *integrated out* of the formula.

$$\begin{aligned} P(T) &= P(D)P(T|D) + P(\bar{D})P(T|\bar{D}) \\ &= 0.001 \cdot 0.95 + 0.999 \cdot 0.02 \\ &= 0.02093 \end{aligned}$$

We thus have that

$$P(\bar{T}) = 1 - P(T) = 1 - 0.02093 = 0.9791$$

and

$$\begin{aligned} P(D|T) &= \frac{P(D)P(T|D)}{P(T)} = \frac{0.001 \cdot 0.95}{0.02093} = 0.04539 \\ P(\bar{D}|T) &= \frac{P(\bar{D})P(T|\bar{D})}{P(T)} = \frac{0.999 \cdot 0.02}{0.02093} = 0.9546 \end{aligned}$$

Note that $0.04539 + 0.9546 = 1$, coherently.

Example 1.2.2 (Techology improvements)**.** If $P(A) = 0.001$ and $P(\bar{A}) = 0.999$ remain unchanged, while technology improves as in the conditional distribution of table 1.4, the results become

$$\begin{aligned} P(D|T) &= 0.1654 \\ P(\bar{D}|T) &= 0.8346 \end{aligned}$$

So probability of being diseased given a positive test is increased due to a greater confidence in the testing system (among the positives

Example 1.2.3 (Less rare disease)**.** If disease is less rare, i.e. $P(A) = 0.01$

- under the first hypothesis on technological development we have $P(A|B) = 0.324$ and $P(\bar{A}|B) = 0.676$,
- under the second hypothesis on technological development $P(A|B) = 0.666$ and $P(\bar{A}|B) = 0.333$.

So in order to obtain $P(A|B) > 0.5$, the disease cannot be too rare and technological development must be very high.

1.2.1.2 DNA test for a crime (ignorance prior))

Example 1.2.4 (Crime and genetic tests)**.** A crime has been committed and there's a suspected: the event C is "the suspected committed the crime". Initially

	C	$\bar{C}$
Compatible DNA	0.999	0.02
Not compatible DNA	0.001	0.98
	1	1

Table 1.5: genetic test performance

no one knows whether the suspected committed the crime or not so our prior (*ignorance prior*) is

$$P(C) = P(\bar{C}) = 0.5$$

A genetic test is available and in situation like this its performance are reported in table 1.5.
What is needed is $P(C|\text{compatible DNA})$ which can be obtained via the Bayes theorem as

$$P(C|\text{compatible DNA}) = 0.9804$$

In other words, if the DNA of the suspected is compatible, the probability that the suspected committed the crime strongly increases compared to $P(C) = 0.5$.

1.2.2 Prior odd ratios, Bayes factor

Some definitions using the notation of D (disease) and T (test)

$$\text{Prior odds ratio in favor} = \frac{P(D)}{1 - P(D)} = \frac{P(D)}{P(\bar{D})}$$
$$\text{Prior odds ratio against} = \frac{P(\bar{D})}{P(D)}$$

We're interested in:

$$\text{Posterior odds ratio in favor } = \frac{P(D|T)}{P(\bar{D}|T)}$$

to compute it note first that

$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D) + P(\bar{D})P(T|\bar{D})}$$
$$P(\bar{D}|T) = \frac{P(\bar{D})P(T|\bar{D})}{P(D)P(T|D) + P(\bar{D})P(T|\bar{D})}$$

Since the denominators are the same their computation is not needed and thus

$$\text{Posterior odds ratio in favor} = \frac{P(D|T)}{P(\bar{D}|T)} = \frac{P(D)P(T|D)}{P(\bar{D})P(T|\bar{D})} = \frac{P(D)}{P(\bar{D})}r$$

The ratio:

$$r = \frac{P(T|D)}{P(T|\bar{D})}$$

is known as **Bayes factor**; since we are speaking of events, it depends on data and not on priors.
So posterior odds ratio in favor can be written as prior odds ratio in favor (which contains info on the prior only) times the Bayes factor (which contains info on the data only).
Note that if $P(D) = 0.5$ (*ignorance prior*), the prior odds ratio in favor is 1 and all the decision depends on the Bayes factor/data.

Example 1.2.5 (Rare disease (continued))**.** If $P(D) = 0.001$:

$$\text{Prior odds ratio in favor} = \frac{P(D)}{P(\bar{D})} = \frac{0.001}{0.999} = 0.001001$$
$$\text{Prior odds ratio against} = \frac{P(\bar{D})}{P(D)} = \frac{0.999}{0.001} = 999$$

The Bayes factor is

$$r = \frac{P(T|D)}{P(T|\bar{D})} = \frac{0.95}{0.02} = 47.5$$

Here the Bayes factor assumes a high > 1 value, the hypothesis of disease is seconded by/after conducing the experiment; the value of the bayes factor is something similar to hypothesis testing based on only the sample as frequentist do[1].
The posterior odds ratio is

$$\frac{P(D|T)}{P(\bar{D}|T)} = \frac{0.0045}{0.955} = 0.04712$$

Example 1.2.6 (Crime and DNA test (continued))**.** We have

$$\begin{aligned}\text{Prior odds ratio in favor} &= \frac{P(C)}{P(\bar{C})} = 1\\ \text{Posterior odds ratio in favor} &= \frac{P(C)}{P(\bar{C})} \cdot r = 1\frac{P(\text{compatible DNA}|C)}{P(\text{compatible DNA}|\bar{C})}\\ &= \frac{0.999}{0.002} = 50.02.\end{aligned}$$

1.3 The statistical model as the basic element for inference

1.3.1 Definition of a statistical model

Among the basic statistical models some examples can be considered.

1. Experiment with known probability of success: Bernoulli distribution.
2. Model for repeated measures: normal distribution.

[1] Actual testing is not used so much in Bayesian statistics because Bayesian says that everything is included in the posterior distribution.

3. Time of functioning of homogeneous apparatuses: exponential distribution.
4. Non parametric models.
5. Sampling without replacement.
6. Inverse sampling.

The case of *inverse sampling* deserves some comments. It illustrates how the way of conducting the experiment may determine the statistical distribution to consider.
The proportion of a characteristic in a population can be managed via the Binomial distribution (that models X as the number of successes in n trials); also X can be the number of failures before obtaining n successes leading to the following distribution (negative binomial)

$$p(x|n,\theta) = \binom{n+x-1}{x}\theta^n(1-\theta)^x$$

Support of X: set of natural numbers that are $\geq x$. $\mathbb{N} = \{x, x+1, \ldots\}$, with $0 \leq \theta \leq 1$.

Now it turns out that n is not necessarily an integer (in negative binomial can be a real number);

- if that is the case the distribution is known as Pascal distribution (special, most famous, case of negative binomial)
- if furthermore $n = 1$ then $p(x|\theta) = \theta(1-\theta)^x$, and we have a Geometric distribution.

Remark 12. Alternative ways of writing a binomial coefficient can be found in the literature, since

$$\binom{n+x-1}{x} = \binom{n-1}{n-x-1}.$$

1.4 Probability refresher

Remark 13. Idea è mettere qui in un unico posto tutti i richiami di probabilità da usare nel seguito.

1.4.1 χ^2 as Gamma distribution

Remark 14. χ^2 distribution is a special case of Gamma; in this Section we see how one can pass from a $X \sim \chi^2_\nu$ to a $Gamma(\alpha, \beta)$.

Important remark 4 (First parametrization). In case of positive support random variable, we say that X is distributed as χ^2 with ν degrees of freedom $X \sim \chi^2_\nu$ if:

$$p(X|\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}\exp\left(-\frac{1}{2}x\right)$$

where $E(X|\nu) = \nu$ and $V(X|\nu) = 2\nu$.
If we create a new variable by applying the transformation

$$Y = \frac{X}{S} = S^{-1}X$$

then $X = SY$ and $\frac{\partial X}{\partial Y} = S$ (Jacobian of the transformation). Thus

$$Y \sim S^{-1}\chi^2_\nu$$

with

$$E(Y|\nu) = S^{-1}\nu$$
$$V(Y|\nu) = S^{-2}2\nu$$

For the density of Y (we substitute X with SY and multiply by the Jacobian):

$$\begin{aligned} p(Y|\nu) &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} (Sy)^{\nu/2-1} \exp\left(-\frac{1}{2}Sy\right) S \\ &= \frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)} (y)^{\nu/2-1} \exp\left(-\frac{1}{2}Sy\right) \end{aligned}$$

In order to recognize it as a Gamma distribution with parameters α, λ, let's set

$$S = 2\lambda \implies \lambda = S/2$$
$$\alpha = \nu/2 \implies \nu = 2\alpha$$

obtaining

$$\begin{aligned} p(Y|\alpha,\lambda) &= \frac{(2\lambda)^\alpha}{2^\alpha\Gamma(\alpha)} (y)^{\alpha-1} \exp\left(-\frac{1}{2}2\lambda y\right) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\lambda y) \end{aligned}$$

which is the first parametrization with, thus

$$E(Y|\alpha,\lambda) = \alpha/\lambda$$
$$V(Y|\alpha,\lambda) = \alpha/\lambda^2$$

Important remark 5 (Second parametrization). If we set $\beta = 1/\lambda$ we find the well-known two-parameters Gamma distribution $Gamma(\alpha, \beta)$

$$p(Y|\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right)$$

with moments

$$E(Y|\alpha,\beta) = \alpha\beta$$
$$V(Y|\alpha,\beta) = \alpha\beta^2$$

Easily, via a simple substitution, one retrieves the moments of Y defined as a χ^2:

$$E\left(Y|\frac{\nu}{2},\frac{2}{S}\right)=\frac{\nu}{S}$$
$$V\left(Y|\frac{\nu}{2},\frac{2}{S}\right)=\frac{2\nu}{S^2}$$

The transformed variable Y coming from X is distributed as $Y\sim S^{-1}\chi^2_\nu$, in this case, since $S^{-1}=\beta/2$ and $\nu=2\alpha$ we get

$$Gamma(\alpha,\beta)=\frac{\beta}{2}\chi^2_{2\alpha}$$

We can find the moments of the χ^2 with S^{-1} and ν and those of the Gamma with α and β, obtaining the same results.

Important remark 6 (One-parameter Gamma distribution). This special case $Y\sim Ga(\alpha)$ is derived from the second parametrization, having set $\beta=\lambda=1$:

$$p(Y|\alpha)=\frac{1}{\Gamma(\alpha)}y^{\alpha-1}\exp(-y)$$

that is also $Gamma(\alpha)=\frac{1}{2}\chi^2_{2\alpha}$.

1.4.2 Inverse χ^2 distribution

TODO: to check yet

If $X\sim\chi^2_\nu$ e $Y\sim S^{-1}\chi^2_\nu$

$$p(X|\nu)=\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}\exp\left(-\frac{1}{2}x\right),\qquad X>0$$
$$p(Y|\nu)=\frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}y^{\nu/2-1}\exp\left(-\frac{1}{2}Sy\right),\qquad Y>0.$$

The inverse χ^2 is such that $\frac{1}{X}\sim\chi^2_\nu$ or in other words $X\sim\chi^{-2}_\nu$

$$p(X|\nu)=\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{-\nu/2-1}\exp\left(-\frac{1}{2}x^{-1}\right),\qquad X>0$$

but also such that $\frac{1}{Y}\sim S^{-1}\chi^2_\nu$ or in other words $Y\sim S\chi_\nu^{-2}$

$$p(Y|\nu)=\frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}y^{-\nu/2-1}\exp\left(-\frac{1}{2}\frac{S}{y}\right),\qquad Y>0.$$

Important remark 7. Remember that in the density of $\chi^{-2}_{\nu-1}$, Y has exponent $-\frac{\nu-1}{2}-1=-\frac{\nu+1}{2}$, so we can write

$$p(Y|\nu)=\frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}y^{-(\nu+1)/2}\exp\left(-\frac{1}{2}\frac{S}{y}\right),\qquad Y>0.$$

that is a $\chi^{-2}_{\nu-1}$.

Chapter 2

From prior to posterior distribution

TODO - before vs after in exchangeability and introduction of choose in binary model (sufficient stat) pag 35 - kernel density e posterior density pag 38

Remark 15. The topics of this block are contained in Chapters 4, 5,6, 7 and 8 of Lambert. Other references in Hoff, chapters 3 and 5[1]

Remark 16. In this section we move from events (and their probability) to random variables (and their distribution function); the aim is to adapt all the machinery of bayesian stuff to do inference.
We need to adapt bayes theorem, the usage of prior for parameters of interest and likelihood to arrive at a posterior distribution of the parameter of interest. I have an idea of the state of the world (prior). Classical statistics is based on the fact that something fixed is in the population and how to use data to estimate that; the parameter is unknown but fixed. In Bayesian stats there's uncertainty on the state of the world/parameter, so we can assume a probability

[1]Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* Springer Science & Business Media. ISBN: 978-0-387-92299-7
The notes illustrate the passage from the prior distribution of the parameter to the posterior. This argument is developed in **Chapter 3** of the textbook, where several further topics are introduced, that will be deepened later in these notes, after the present block.
In this block of notes, we focus on the cases of **a)** discrete prior and likelihood **b)** continuous prior and discrete likelihood, where the discrete likelihood is the binomial, the continuous prior is the Beta. See how the textbook sketches the topic for the binomial likelihood in *Section 3.1 - The binomial model.*
The relevance of predictive distributions is illustrated in *Section 2.3* of the textbook.
Also the normal univariate model is illustrated in this block of notes. In the textbook this topic appears in **Chapter 5**, where the case of normal continuous prior and normal continuous likelihood is developed (continuous normal posterior): *Section 5.1 - The normal model* and *Section 5.2 - Inference for the mean, conditional on the variance.*
Hoff's **Chapter 3** deals with *conjugacy* in *Section 3.1* (page 38), the whole *Section 3.3 - Exponential families and conjugate priors.*
The link between the distribution χ^2 and the two parameters Gamma, together with the case of normal gamma prior and discrete Poisson likelihood (continuous gamma posterior) are developed in *Section 3.2 - The Poisson model.*
The topic of credibility intervals is developed in the whole *Section 3.1.2 - Confidence regions* and is not developed in this block.
Important statistical distributions. For remarks on important statistical distributions see pages 253-258.

of states of the world, the prior.
After fixing the prior something is done (experiment) to see if our idea changes. However in passing from probability of events to statistical distributions something occurs/changes.

2.1 Bayes theorem for random variables

Important remark 8. We can express the passage from the prior to the posterior as:

$$h(\theta|x) = \frac{g(\theta) \cdot p(x|\theta)}{p(x)} = \frac{g(\theta) \cdot L(\theta; x)}{p(x)} \propto g(\theta) \cdot L(\theta; x)$$

where

- $g(\theta)$ is the prior probability distribution for a parameter of interest
- $p(x|\theta) = L(\theta; x)$ is the likelihood of the current sample given the value asumed by θ
- the denominator $p(x)$ is a constant that only depends on the data (it's averaged on all the possible value of the parameter θ, which are at least two[2])

Important remark 9. In the following we specify this expression for the discrete and continuous case.

2.1.1 Discrete parameter and discrete data

Here θ has p possible values, $\theta_1, \ldots, \theta_p$, each with a certain prior probability that sums to the unit:

$$\sum_{i=1}^{p} g(\theta_i) = 1$$

For a sample of size n, we have:

$$\begin{aligned} h(\theta_i|x_1, \ldots, x_n) &= \frac{g(\theta_i) \cdot L(\theta_i; x_1, \ldots, x_n)}{\sum_{i=1}^{p} g(\theta_i) \cdot L(\theta_i; x_1, \ldots, x_n)} \\ &\propto g(\theta_i) \cdot L(\theta_i; x_1, \ldots, x_n) \\ &\propto g(\theta_i) \cdot L(\theta_i; x_1) \cdot \ldots \cdot L(\theta_i; x_n) \end{aligned}$$

where in the last passage the likelihood has been rewritten as a product (of individual likelihood/densities), since we assume independence of the observation (actually we do not need observation to be independent, they could be correlated, but we do not deal with this case).
The posterior distribution is a probability distribution:

$$\sum_{i}^{p} h(\theta_i|x_1, \ldots, x_n) = 1.$$

[2]If the prior was Dirac there would not be the needs to perform an experiment

Important remark 10. We do not develop the case of discrete parameter and continuous data. This can be used when prior are expressed by experts' considerations, or to choosing/compare among a discrete set of models; each model may receive an a priori probability and can be supported by an experiment.

2.1.2 Continuous parameter and discrete data

Parameter θ follows a (prior) probability distribution $g(\theta)$ which integrates to 1. For samples of size n, we can write

$$\begin{aligned} h(\theta|x_1,\dots,x_n) &= \frac{g(\theta)\cdot L(\theta;x_1,\dots,x_n)}{\int g(\tilde{\theta})\cdot L(\tilde{\theta};x_1,\dots,x_n)d\tilde{\theta}} \\ &\propto g(\theta)L(\theta;x_1,\dots,x_n) \end{aligned}$$

Also in this case the likelihood can be written in a simpler way under the independence condition.

2.2 Exchangeability

In Bayesian statistics the concept of exchangeability (of a set of random variables) become important

NB prof: sezione completamente introdotta, sintesi di Hoff 2.7 e 2.8

Definition 2.2.1 (Exchangeability)**.** Let $Y_1,\dots,Y_n$ be a sequence of random variablew and $p(y_1,\dots,p_n)$ its joint density.
We say $Y_1,\dots,Y_n$ are exchangeable if $p(y_1,\dots,y_n) = p(y_{\pi_1},\dots,y_{\pi_n})$ for all permutations π of $\{1,\dots,n\}$.

Remark 17. Roughly speaking, $Y_1,\dots,Y_n$ are exchangeable if the subscript label convey no information about the outcomes.

Remark 18. What is the relationship between exchangeability and iid? It can be proved that

Theorem 2.2.1 (DeFinetti)**.** *The random variable $Y_1,\dots,Y_n$ are exchangeable for all n if and only if they depend on a common unknown parameter having distribution $\theta \sim p(\theta)$, but conditionally of its assumed variable are iid, that is*

$$\begin{cases} \theta \sim p(\theta) \\ Y_1,\dots,Y_n|\theta \text{ are iid} \end{cases} \iff Y_1,\dots,Y_n \text{ are exchangeable for all } n$$

Proof. Omitted □

Important remark 11. If exchangeability holds we can assume, conditionally on the parameter defining the random variable distributions, that these random variable are iid.
In the application of bayes theorem this helps in writing likelihood which can be written as product of individual density (using the common parameter), that is assuming (conditional) independence.

Important remark 12. When is the condition $Y_1,\dots,Y_n$ are exchangeable for all n reasonable?
For this condition to hold we must have *exchangeability* and *repeatability*:

- exchangeability will hold if the labels convey no informations (eg not a time)
- situations in which repeatability is reasonable include the following:
  - $Y_1, \ldots, Y_n$ are outcomes of a repeatable experiment
  - $Y_1, \ldots, Y_n$ are sampled from a finite population with relacement
  - $Y_1, \ldots, Y_n$ are sampled from an infinite population without relacement

Thus in the classical case, if $Y_1, \ldots, Y_n$ are exchangeable and sampled from a finite population of size $N > n$ without replacement, then they can be modeled as approximately being iid

Remark 19. Starting from the next session we see some one-parameter models; these areclass of sampling distributions that is indexed by a single unknown parameter such as binomial, normal and poisson models

2.3 Inference for a proportion

Remark 20. Here we see the how to apply the bayes theorem with random variable: we have a prior distribution for the parameter of interest (proportion) and we conduct an experiment; finally we obtain a posterior distribution for the parameter.
We tackle the case using different priors for the parameter (discrete or continuous) and considering an experiment which produces dichotomic data

2.3.1 Discrete prior

Important remark 13. In both case we start from a prior of four possible proprotions, with uniform probability:

$$\theta = (0.2; 0.4; 0.6; 0.8)$$
$$g(\theta) = 1/4 = 0.25$$

2.3.1.1 Sample of $n = 1$

Example 2.3.1 (One replication of the experiment)**.** The probability distribution for this unique observation, that is also the likelihood for the parameter, is the Bernoulli

$$p(X|\theta) = \theta^x (1-\theta)^{1-x}$$

If

- $X = 1$ is observed then $p(X = 1|\theta_j) = \theta_j$, for each possible value of θ.
  The passages we do to modify our prior opinion about any possible value of the parameter after observing $X = 1$ is resumed in table 2.1.
  So it turns out that a single positive answer makes 4 times more plausible the that the population proportion is 0.8 rather than 0.2.

θ_j	$g(\theta_j)$	$p(X=1\|\theta_j)$	$g(\theta_j)p(X=1\|\theta_j)$	$h(\theta\|X=1)$
0.2	0.25	0.2	0.05	0.10
0.4	0.25	0.4	0.1	0.20
0.6	0.25	0.6	0.15	0.30
0.8	0.25	0.8	0.20	0.40
Sum	1		$p(x)=0.5$	1

Table 2.1: Experiment with single unit extraction with $X = 1$.

θ_j	$g(\theta_j)$	$p(X=0\|\theta_j)$	$g(\theta_j)p(X=0\|\theta_j)$	$h(\theta\|X=0)$
0.2	0.25	0.8	0.2	0.40
0.4	0.25	0.6	0.15	0.30
0.6	0.25	0.4	0.1	0.20
0.8	0.25	0.2	0.05	0.10
Sum	1		$p(x)=0.5$	1

Table 2.2: Experiment with single unit extraction with $X = 0$.

- $X = 0$ is observed then $p(X = 0|\theta_j) = 1 - \theta_j$.
  The passages we do to modify our prior opinion about any possible value of the parameter after observing $X = 0$ is resumed in table 2.2.
  Thus a single negative answer makes 4 times less plausible the statement that the population proportion is 0.8 rather than 0.2.

2.3.1.2 Sample of $n = 20$

Example 2.3.2 (n replications of the experiment)**.** Instead of performing only one trial, we have $n = 20$. The probability distribution of the number r of successes among n observations is the binomial:

$$p(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$$

In case we observe $r = 15$ successes out of $n = 20$ trials, the likelihood/data generation distribution is:

$$L(\theta_j; 15, 20) = \binom{20}{15}\theta_j^{15}(1-\theta_j)^5, \quad j = 1, 2, 3, 4$$

The passages we do to modify our prior opinion about any possible value of the parameter after observing $r = 15$ successes out of $n = 20$ trials is resumed in table **??** (where the two central columns were multiplied by 10^{-7} in order to be readable, all are very low probability values).
Note that the likelihood could not to contain the binomial coefficient (since it is simplified in the division going from column 3 to 4, being both at the numerator (column 3) and denominator (sum of column 3)) .
Again the last division of column 4 to compute $h(\theta|r,n) = L(\theta; r, n)g(\theta)$ for each θ_j make a normalization so that the posterior sums to the unit and is a proper probability distribution.
Again the posterior (conditional on the experiment) for the parameter is proportional to the product of the likelihood by the prior.

NB: prof calls likelihood without binomial coefficient "proper" likelihood since the binomial coefficient does not contains the parameter θ_j

θ_j	$g(\theta_j)$	$L(\theta_j; 15, 20) \times 10^{-7}$	$g(\theta_j)L(\theta_j; 15, 20) \times 10^{-7}$	$h(\theta_j \vert 15, 20)$
0.2	0.25	0.00	0.000	0.000
0.4	0.25	0.83	0.201	0.005
0.6	0.25	48.10	12.025	0.298
0.8	0.25	112.60	28.150	0.697
Sum	1		40.376(∗)	1

Table 2.3: Experiment with $n = 20$

Comparison with pure ML estimation For comparison's sake, let's look at the ML estimator of the proportion (using only the sample, not the prior). Here we could ignore the binomial coefficient (as done by prof) in the optimization (since it's just a constant which does not depend on θ). The derivation goes like:

NB prof: nelle note originali vi è un typo algebrico all'ultimo passaggio

$$L(\theta; r, n) = \binom{n}{r} \theta^r (1-\theta)^{n-r}$$

$$\log L(\theta; r, n) = \log \binom{n}{r} + r \log \theta + (n-r) \log(1-\theta)$$

$$\frac{\partial \log L(\theta; r, n)}{\partial \theta} = r\frac{1}{\theta} + (n-r)\frac{1}{1-\theta}(-1) = \frac{r}{\theta} - \frac{n-r}{1-\theta} = \frac{r(1-\theta) - (n-r)\theta}{\theta}$$
$$= \frac{r - r\theta - n\theta + r\theta}{\theta(1-\theta)} = \frac{r - n\theta}{\theta(1-\theta)}$$

Thus by putting $\frac{\partial \log L}{\partial \theta} = 0$ we derive the ML estimator, which as we know is:

$$\hat{\theta} = \frac{r}{n}$$

Applied to our sample is $\hat{\theta} = \frac{15}{20} = 0.75$ (the result is somewhat between the parameter having the posterior higher probability).
This value, a point estimate, differs from what will be derived when passing for priors to posteriors.

```
# riproduzione della tabella, ignorando il 10^{-7}
theta <- seq(0.2, 0.8, 0.2)
prior <- rep(0.25, 4)
lik   <- sapply(theta, function(t) dbinom(x = 15, size = 20, prob = t))
num   <- prior * lik
den   <- sum(num)
posterior <- num/den
## sum(posterior) ## ==1

round(cbind(theta, prior, lik, num, posterior), digits = 3)

##      theta prior   lik   num posterior
## [1,]   0.2  0.25 0.000 0.000     0.000
## [2,]   0.4  0.25 0.001 0.000     0.005
## [3,]   0.6  0.25 0.075 0.019     0.298
## [4,]   0.8  0.25 0.175 0.044     0.697
```

```
## probabilmente non torna nelle colonne centrali perché', a parte l'esponente,
## ha ignorato il coefficiente binomiale nei conti, sebbene lo presenti nella
## formula

## check with expected value of posterior distribution
sum(theta * posterior) ## not exactly the same as ML estimate, btw

## [1] 0.7383344
```

2.3.2 Continuous prior

Remark 21. The prior knowledge for the proportion θ of the population can be expressed as a continuous distribution using the Beta.

2.3.2.1 Beta distribution reminder

Definition 2.3.1. Characterized by two parameters, $a > 0$ and $b > 0$, defined by:

$$p(X|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} I_{(0,1)}(x)$$

and contains the Beta function

$$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

which contains the Gamma function that, for a positive integers n, is defined as

$$\Gamma(n) = (n-1)!$$

Important remark 14. The first two moments of the Beta distribution are

$$E(X|a,b) = \frac{a}{(a+b)}$$
$$V(X|a,b) = \frac{ab}{(a+b+1)(a+b)^2}$$

Remark 22. The variance can be written differently, by noticing that:

$$\frac{ab}{(a+b+1)(a+b)^2} = \frac{a}{a+b}\frac{b}{a+b}\frac{1}{a+b+1} = \frac{a}{a+b}\left(1-\frac{a}{a+b}\right)\frac{1}{a+b+1}$$

NB prof: qui nell'originale due volte a al numeratore

and thus obtaining

$$V(X|a,b) = \frac{E(X|a,b)\,(1-E(X|a,b))}{(a+b+1)}. \tag{2.1}$$

Example 2.3.3. When $a = b = 1$, the $Beta(1,1)$ coincides with a $U(0,1)$

2.3.2.2 Uniform prior

Remark 23. We start with a $U(0,1)$ prior, that is also $Beta(1,1)$.[3]

Let us assume a uniform prior in the interval (0,1), i.e. $U(0,1)$:

$$g(\theta|0,1) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

for which the moments are (by adapting the Beta formulas with $a = b = 1$):

$$E(\theta|0,1) = 1/2 = 0.5 \qquad V(\theta|0,1) = 1/12 = 0.083$$

As before in the experiment we observe $r = 15$ successes out of $n = 20$ trials.

The expression of the likelihood is also the same (below the binomial coefficient remains), with continuos domain from 0 to 1:

NB prof: qui dovrebbe essere = invece di $\propto$ se si tiene il coefficiente binomiale?

$$L(\theta; 15, 20) = \binom{20}{15}\theta^{15}(1-\theta)^5$$

The posterior distribution of θ is defined in this case by having an integral over all its possible values at the denominator

$$h(\theta|r,n) = \frac{g(\theta)L(\theta;r,n)}{\int_0^1 g(\theta)L(\theta;r,n)d\theta} = \frac{g(\theta|0,1)L(\theta;15,20)}{\int_0^1 g(\theta|0,1)L(\theta;15,20)d\theta}$$

Now considering the case of $U(0,1)$ prior (for which the density is $g(\theta) = 1, \forall\theta \in [0,1]$, experiment with $r = 15$ and $n = 20$, the posterior becomes:

$$\begin{aligned} h(\theta|15,20) &= \frac{1 \cdot \binom{20}{15}\theta^{15}(1-\theta)^5}{\int_0^1 1 \cdot \binom{20}{15}\theta^{15}(1-\theta)^5 d\theta} = \frac{\theta^{15}(1-\theta)^5}{\int_0^1 \theta^{15}(1-\theta)^5 d\theta} \\ &= \frac{\theta^{16-1}(1-\theta)^{6-1}}{\int_0^1 \theta^{16-1}(1-\theta)^{6-1}d\theta} = \frac{\theta^{16-1}(1-\theta)^{6-1}}{B(16,6)} \\ &= Beta(16,6) \end{aligned}$$

Virtual and actual sample So, it turns out that if the prior is a $Beta(1,1)$ ($U(0,1)$) and in the experiment we have 15 successes and 5 failures, the posterior becomes a $Beta(16,6)$. This gives an interpretation to the parameter a, b of the distribution which can be seen as sample size (before is virtual, after experiment is virtual + actual) of the units having failures and successes (tab 2.4).
To express a prior $U(0,1)$ information, that is a $Beta(1,1)$, 2 virtual cases are enough (1 success and 1 failure); enough to model a distribution with that expected value and variance. (The concept of virtual sample will be summarized also below.)

Prior vs Posterior Now let's compare moments and shapes of the prior and posterior distributions to appreciate the knowledge benefit of the experiment: before it we knew nothing (we knew θ can be in 0,1 but no more than that). In:

[3] The example that is developed below is analogous to the *happiness data* example of Hoff, *Chapter 3 - One-parameter models*, *Section 3.1 - The binomial model.*

	Prior	Experiment	Posterior
Successes	1	15	16
Failures	1	5	6
Total	2	20	22

Table 2.4: Virtual and actual sample for this case

	Prior $Beta(1,1)$ $(U(0,1))$	Posterior $Beta(16,6)$
$E(\theta)$	$E(\theta\|1,1) = 1/2 = 0.5$	$E(\theta\|16,6) = \frac{a}{a+b} = \frac{16}{22} = 0.7272$
$V(\theta)$	$V(\theta\|1,1) = 1/12 = 0.083$	$V(\theta\|16,6) = \frac{ab}{(a+b+1)(a+b)^2} = \frac{16 \cdot 6}{(16+6+1)(16+6)^2} = 0.008624$

Table 2.5: Prior vs posterior moments in uniform(0,1)-binomial case

- table 2.5 the moments of the two distribution are compared and an improvement can be observed: the posterior variance is approximately 1/10 of the prior variance (this is due to the consideration of both the prior and the experiment information)
- fig 2.1 we can plot the distribution to shows the passage from the prior to the posterior through the experiment

Conjugacy introduction

Remark 24. [From BDA3] The property that the posterior distribution follows the same parametric form as the prior distribution (in this case the Beta) is called *conjugacy*.
Here we can say that the **beta prior** distribution is a **conjugate family** for the *binomial likelihood*.
The conjugate family is mathematically convenient in that the posterior distribution follows a known parametric form

2.3.2.3 Generic beta prior

A more generic prior $Beta(a,b)$ (with $a,b > 0$ to be choosen to have a certain expected value and variance of θ distribution) is expressed, as usual as

NB prof: qui cambio notazione per uniformarla a prima, g prior h posterior

$$\begin{aligned} g(\theta|a,b) &= \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} I_{(0,1)}(\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

The binomial likelihood is:

$$\begin{aligned} p(r|n,\theta) = L(\theta;r,n) &= \binom{n}{r} \theta^r (1-\theta)^{n-r} \\ &\propto \theta^r (1-\theta)^{n-r} \end{aligned}$$

The posterior distribution is thus

NB prof: qui qualche esplicitazione in più introdotta

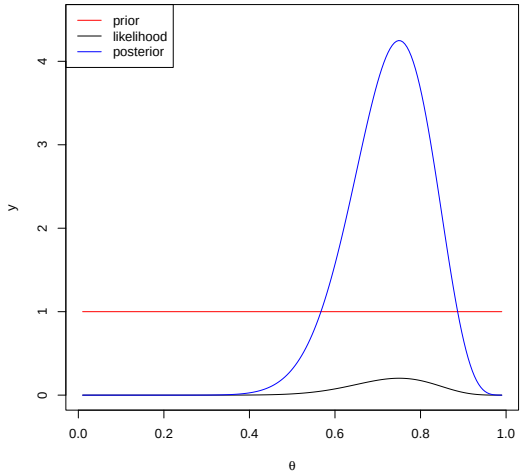

Figure 2.1: Uniform prior and binomial likelihood case

$$\begin{aligned} h(\theta|a,b,r,n) &= \frac{g(\theta|a,b)\cdot L(\theta;r,n)}{\int_0^1 g(\theta|a,b)\cdot L(\theta;r,n)d\theta} \\ &\propto g(\theta|a,b)L(\theta;r,n) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\cdot\theta^r(1-\theta)^{n-r} \\ &= \theta^{a+r-1}(1-\theta)^{b+(n-r)-1} \end{aligned}$$

Remark 25. The last term is the *kernel* of a $Beta(a+r, b+n-r)$ density (that is its numerator, ignoring the constant denominator $B(a+r, b+n-r)$).

Remark 26. [Miei ragionamenti] Abbiamo trovato che il kernel della posteriori è quello di una beta con parametri $a+r$, $b+n-r$;

- sappiamo che differisce dalla distribuzione della posteriori per una costante,
- sappiamo però che la posteriori è una distribuzione con integrale a 1

quindi la costante deve essere la costante di una beta con tali parametri e dunque la posteriori è effettivamente una beta con tali parametri

Remark 27. [From wikipedia] In statistics, especially in Bayesian statistics, the kernel of a probability density function (pdf) or probability mass function (pmf) is the form of the pdf or pmf in which any factors that are not functions of any of the variables in the domain are omitted.[1] Note that such factors may well be functions of the parameters of the pdf or pmf. These factors form part of the normalization factor of the probability distribution, and are unnecessary in many situations. For example, in pseudo-random number sampling, most sampling algorithms ignore the normalization factor. In addition, in Bayesian analysis of conjugate prior distributions, the normalization factors are generally ignored during the calculations, and only the kernel considered. At the end, the form of the kernel is examined, and if it matches a known distribution, the

normalization factor can be reinstated. Otherwise, it may be unnecessary (for example, if the distribution only needs to be sampled from).

Important remark 15. So we have seen the generalized rules for cooking up a beta prior based on a and b, and a binomial likelihood based on r, n (where the special case of uniform was tackled before), ending up in a beta posterior with parameters depeding on prior and likelihood.
The next step is just changing the prior to accomodate different hypotheses on previous knowledge.

2.3.2.4 Beta with prespecified expected value/variance

Consider another prior, based on conjectures on the expected value and the variance

$$\begin{cases} E(\theta|a,b) = 0.4 \\ V(\theta|a,b) = 0.01 \end{cases}$$

To obtain a, b we set a system of equation[4] and exploit the alternative definition of variance of the beta distribution (eq 2.1)

$$\begin{cases} E(\theta|a,b) = \frac{a}{a+b} = 0.4 \\ V(\theta|a,b) = \frac{E(\theta|a,b)(1-E(\theta|a,b))}{a+b+1} = 0.01 \end{cases} \quad \begin{cases} \frac{a}{a+b} = 0.4 \\ \frac{0.4 \cdot 0.6}{a+b+1} = 0.01 \end{cases} \quad \dots \begin{cases} a = 9.2 \\ b = 13.8 \end{cases}$$

So in order to express those prior information (in terms mean and variance) we need $a = 9.2$ successes and $b = 13.8$ failures in the prior "virtual sample" and thus the prior information corresponds to 23 virtual cases.
The experiment is kept the same as before: 20 trials and 15 successes, as well as the binomial likelihood.
Thus:

- the complete sample has 43 cases (tab 2.6)
- the posterior distribution is a Beta, with parameters

  $$a' = a + r = 9.2 + 15 = 24.2$$
  $$b' = b + n - r = 13.8 + 20 - 15 = 18.8$$

  and moments

  $$E(\theta|a',b') = E(\theta|a,b,n,r) = \frac{a'}{a'+b'} = \frac{a+r}{a+r+b+n-r} = 0.562$$
  $$V(\theta|a',b') = V(\theta|a,b,n,r) = \frac{(a' \cdot b')}{(a'+b'+1) \cdot (a'+b')^2} = 0.0056$$

- the distributions plot (fig 2.2) shows the passage from the prior to the posterior through the experiment and that the variability of the posterior is smaller that in the first case.

[4] Full development:

$$\begin{cases} \frac{a}{a+b} = 0.4 \\ \frac{0.4 \cdot 0.6}{a+b+1} = 0.01 \end{cases} \quad \begin{cases} a = 0.4a + 0.4b \\ a + b + 1 = 24 \end{cases} \quad \begin{cases} b = \frac{3}{2}a \\ a = 23 - b \end{cases} \quad \begin{cases} b = \frac{3}{2}(23 - b) \\ a = 23 - b \end{cases} \quad \begin{cases} b = \frac{69}{5} = 13.8 \\ a = 23 - 13.8 = 9.2 \end{cases}$$

	Prior	Experiment	Posterior
Successes	9.2	15	24.2
Failures	13.8	5	18.8
Total	23	20	43

Table 2.6: Sample sizes of prior (virtual) and experiment (actual)

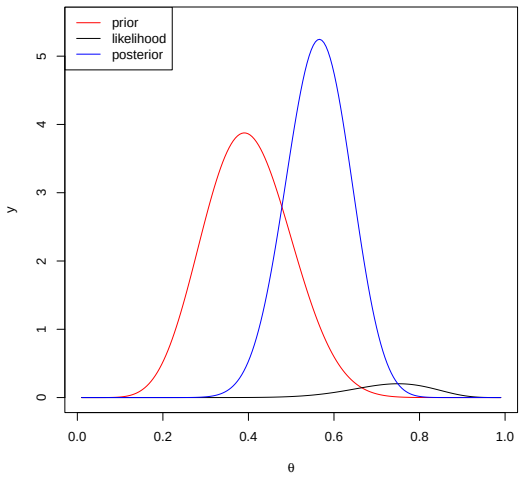

Figure 2.2: General beta-binomial example

	Prior	Experiment	Posterior
Successes	124.5	15	139.5
Failures	124.5	5	129.5
Total	249	20	269

Table 2.7: Virtual sample with precise/indifferent prior

2.3.2.5 Precise/informative, indifference, beta prior

Considering another prior, based on the following conjectures on the expected value and the (small) variance

$$\begin{cases} E(\theta|a,b) = 0.5 \\ V(\theta|a,b) = 0.001 \end{cases}$$

These hypotheses yields to the following parameters for the prior

$$\begin{aligned} a &= 124.5 \\ b &= 124.5 \end{aligned}$$

Thus the prior expected value 0.5 (and variance 0.001) is equivalent to 124.5 virtual successes out of 249 virtual cases. Adopting the same experiment/likelihood:

- the complete sample becomes composed of 269 cases (2.7)
- parameters of the posterior are

  $$\begin{aligned} a' &= 124.5 + 15 = 139.5 \\ b' &= 124.5 + 20 - 15 = 129.5 \end{aligned}$$

  so it is a $Beta(139.5; 129.5)$ with moments

  $$\begin{aligned} E(\theta|a',b') &= E(\theta|a,b,n,r) = \frac{139.5}{269} = 0.518 \\ V(\theta|a',b') &= V(\theta|a,b,n,r) = 0.000925 \end{aligned}$$

  To note that the ratio (posterior variance/prior variance) is near to 1, since the experiment adds very few cases to the virtual ones:

  $$\frac{V(\theta|a,b,n,r)}{V(\theta|a,b)} = \frac{0.000925}{0.001} = 0.99.$$

- finally, the plot (fig 2.3) shows the passage from the prior to the posterior through the experiment.

2.3.2.6 Virtual sample sum up

Important remark 16. In essence[5] , in the previous examples:

[5]This topic is also dealt in Hoff, Section 3.3.1, pages 38-39 for the binomial model, and later, in Section 3.2. for the Poisson model.

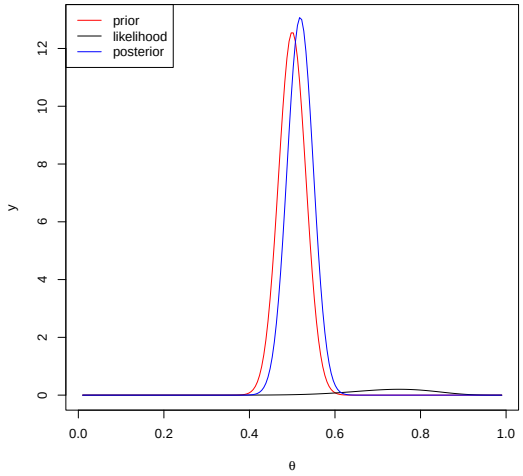

Figure 2.3: Experiment with precise uninformative prior

1. for a conjecture on the 0.5 expected value, without considering its variability, a prior of 1 case out of 2 is sufficient (tab 2.4)
2. for a conjecture on specific expected value and variances, a prior of 23 virtual cases is needed (tab 2.6)
3. a very precise conjecture on an indifference expected value needs a prior of many virtual cases (tab 2.7)

2.3.2.7 Expected values of posterior, prior and likelihood

Through the Bayes theorem, one can pass from a prior $Beta(a, b)$ to a posterior $Beta(a', b')$ with $a' = a + r$ and $b' = b + n - r$. Therefore some algebrical manipulations can highlight one fact

$$\begin{aligned} E(\theta|a', b') &= \frac{a'}{a' + b'} = \frac{a + r}{a + r + b + n - r} = \frac{a + r}{a + b + n} = \frac{a}{a + b + n} + \frac{r}{a + b + n} \\ &= \frac{a}{a + b + n}\frac{a + b}{a + b} + \frac{r}{a + b + n}\frac{n}{n} = \frac{a + b}{a + b + n}\frac{a}{a + b} + \frac{n}{a + b + n}\frac{r}{n} \\ &= \frac{a + b}{a + b + n}E(\theta|a, b) + \frac{n}{a + b + n}\bar{X} \end{aligned}$$

The last passage show how the posterior distribution expected value can be seen as a weighted mean of the prior $E(\theta|a, b)$ and the sample mean from the experiment $\bar{X}$, with weights proportional to the virtual sample and sample size respectively.

2.4 Inference for a mean

NB prof: Sezione interamente rivista nella notazione perché θ e ϕ le confondo e soprattutto di ϕ secondo me non c'è bisogno, sostituibile con σ^2.
In sintesi è un gran mischione di notazioni tra appunti e hoff che però mi risulta personalmente funzionale.

Here we consider the case of a continuous random variable for which our parameter of interest is the mean (assuming known/given variance of the variable),

with the following assumptions. For:

- the **prior**, we assume that the population mean could be approximatively normally distributed/described, with parameter θ_0 and τ_0^2 parameters.

$$\theta \sim N(\theta_0, \tau_0^2)$$
$$g(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left[-\frac{1}{2\tau_0^2}(\theta - \theta_0)^2\right] \propto \exp\left[-\frac{1}{2\tau_0^2}(\theta - \theta_0)^2\right]$$

  Note that θ_0 and τ_0^2 are *not* θ and σ^2, the mean and variance of the continuous variable itself, the first of which we're interested in

- the **likelihood** we assume that we have sampled $n > 1$ units coming from independent normals distribution with variance σ^2 known and mean θ unknown:

$$\left\{X_1, X_2, \ldots, X_n \mid \theta, \sigma^2\right\} \sim i.i.d.\ N(\theta, \sigma^2)$$

  Thus the likelihood becomes:

$$\begin{aligned} p(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right] \end{aligned}$$

- the **posterior** can be derived by writing only one exponent, computing the squares, transferring the terms not containing θ to the left side of the equation $\propto$:

$$\begin{aligned} h(\theta \mid x_1, \ldots, x_n) &\propto g(\theta)\, p(x_1, \ldots, x_n \mid \theta) \\ &= \exp\left[-\frac{1}{2\tau_0^2}(\theta - \theta_0)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right] \\ &= \exp\left\{-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta^2 + \theta_0^2 - 2\theta\theta_0)^2 + \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i^2 + n\theta^2 - 2\theta\sum_{i=1}^{n} x_i\right)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{\theta^2}{\tau_0^2} + \frac{\theta_0^2}{\tau_0^2} - \frac{2\theta\theta_0}{\tau_0^2} + \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2} + \frac{n\theta^2}{\sigma^2} - \frac{2\theta\sum_{i=1}^{n} x_i}{\sigma^2}\right]\right\} \end{aligned}$$

  Now at this point, remembering it's a function of θ we gather terms with θ^2, θ and all the remaining stuff which will be avoided by proportionality

$$\begin{aligned} &= \exp\left\{-\frac{1}{2}\left[\theta^2 \underbrace{\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)}_{a} - 2\theta \underbrace{\left(\frac{\theta_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}\right)}_{b} + \underbrace{\frac{\theta_0^2}{\tau_0^2} + \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2}}_{c}\right]\right\} \\ &= \exp\left[-\frac{1}{2}\big(a\theta^2 - 2b\theta\big)\right] \end{aligned}$$

Now, following Hoff, let's see if $h(\theta \mid x_1, \ldots, x_n)$ takes the form of a normal density

$$\begin{aligned}
h(\theta \mid x_1, \ldots, x_n) &\propto \exp\left[-\frac{1}{2}\big(a\theta^2 - 2b\theta\big)\right] \\
&= \exp\left[-\frac{1}{2}a\left(\theta^2 - \frac{2b\theta}{a}\right)\right] \\
&= \exp\left[-\frac{1}{2}a\left(\theta^2 - \frac{2b\theta}{a} + \frac{b^2}{a^2} - \frac{b^2}{a^2}\right)\right] \\
&= \exp\left[-\frac{1}{2}a\left(\theta^2 - \frac{2b\theta}{a} - \frac{b^2}{a^2}\right) + \frac{1}{2}\frac{b^2}{a}\right] \\
&\propto \exp\left[-\frac{1}{2}a\left(\theta - \frac{b}{a}\right)^2\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{\theta - \frac{b}{a}}{1/\sqrt{a}}\right)^2\right]
\end{aligned}$$

This function is the kernel of a normal with mean b/a and $1/\sqrt{a}$ standard deviation, so $h(\theta \mid x_1, \ldots, x_n)$ being a probability distribution it will be normal.
We name the mean and variance of posterior density as θ_1 and τ_1^2, so after the experiment

$$\theta | x_1, \ldots, x_n \sim N(\theta_1, \tau_1^2)$$

where:

$$\begin{aligned}
\theta_1 &= \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\theta_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \\
\tau_1^2 &= \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}
\end{aligned}$$

Important remark 17. Some remarks:

1. the inverse of variance is often referred as *precision* of a random variable (that is how close we are to the centre). Let the following be the prior, sampling and posterior precisions:

$$\begin{aligned}
&\frac{1}{\tau_0^2} = \tilde{\tau}_0^2 && \text{prior precision} \\
&\frac{1}{\sigma^2} = \tilde{\sigma}^2 && \text{sampling precision} \\
&\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} = \tilde{\tau}_1^2 && \text{posterior precision}
\end{aligned} \tag{2.2}$$

It is convenient to think about precision as the quantify of information/-precision on an additive scale. For this normal-normal model equation 2.2 yields:

$$\tilde{\tau}_1^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$$

and posterior information/precision = prior info + data info.

2. we can "state" a weight as ratio between prior and posterior precision, and then develop it; we have:

$$\omega = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} = \frac{\frac{1}{\tau_0^2}}{\frac{\sigma^2+n\tau_0^2}{\tau_0^2\sigma^2}} = \frac{1}{\tau_0^2}\cdot\frac{\tau_0^2\sigma^2}{\sigma^2+n\tau_0^2} = \frac{\sigma^2}{\sigma^2+n\tau_0^2}$$

The remaining weight (to sum up to 1)

$$1-\omega = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} = \frac{n\tau_0^2}{\sigma^2+n\tau_0^2}$$

3. the posterior variance can be rewritten as

$$\begin{aligned}\tau_1^2 &= \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} = \frac{1}{\frac{\sigma^2+n\tau_0^2}{\tau_0^2\sigma^2}} = \frac{\tau_0^2\sigma^2}{\sigma^2+n\tau_0^2} = \sigma^2\cdot\frac{\tau_0^2}{\sigma^2+n\tau_0^2}\\ &= \frac{\sigma^2}{n}\cdot\frac{n\tau_0^2}{\sigma^2+n\tau_0^2} = \frac{\sigma^2}{n}(1-\omega)\end{aligned}$$

So the posterior variance is smaller than the one of the ML estimator for the mean ($\frac{\sigma^2}{n}$)

4. regarding the posterior expected value, it can be rewritten as a weighted mean of prior expectation and the sample mean, with the weights above

$$\theta_1 = \frac{\frac{1}{\tau_0^2}\theta_0+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} = \theta_0\underbrace{\left(\frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\right)}_{\omega} + \bar{x}\underbrace{\left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\right)}_{1-\omega}$$

5. Thus finally:

$$\theta \mid x_1,\ldots,x_n \sim\ N\left(\omega\theta_0+(1-\omega)\bar{x}, \frac{\sigma^2}{n}(1-\omega)\right)$$

Important remark 18 (About the variance of the prior and sample sizes). If either $\tau_0^2 \to \infty$ or $n \to \infty$, then in both cases

$$\omega = \frac{\sigma^2}{\sigma^2+n\tau_0^2} \to 0$$

In this cases, therefore after the experiment:

$$\theta \mid x_1,\ldots,x_n \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

The posterior will no longer depend on the prior (the parameters of the prior disappear from the posterior) and moments of the posterior coincide with the ML estimates.

Example 2.4.1. Suppose that $\tau_0^2 < \sigma^2$, and especially $\tau_0^2 = \frac{\sigma^2}{m}$ with $m > 1$. In this case we have that

- the prior distribution of the mean is

$$\theta \sim N\left(\theta_0, \frac{\sigma^2}{m}\right)$$

Adopting the sample mean distribution, m can be seen as the number of observation which contributed to the prior distribution definition, the so-called *virtual sample*.
The virtual sample size can also be seen as

$$m = \frac{\sigma^2}{\tau_0^2}$$

i.e., as the ratio of the prior precision to the likelihood *precision*;

- the weights for the posterior moments become

$$\omega = \frac{\frac{m}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{m}{\sigma^2}} = \frac{m}{m+n}$$
$$1 - \omega = \frac{n}{m+n}$$

- expected value, variance and distribution simplifies to

$$\theta_1 = \theta_0 \frac{m}{m+n} + \bar{x}\frac{n}{m+n}$$
$$\tau_1 = \frac{\sigma^2}{n}\frac{n}{m+n} = \frac{\sigma^2}{m+n}$$
$$\theta \mid x_1, \ldots, x_n \sim N\left(\theta_0 \frac{m}{m+n} + \bar{x}\frac{n}{m+n}, \frac{\sigma^2}{m+n}\right)$$

2.5 Inference for a count

Remark 28. Some measurements (eg number of friends) have values that are whole numbers. For these phenomenon the simplest probability model of the measurement is the Poisson model.
In a bayesian context thus, the likelihood depends only on one parameter of the Poisson distribution, the mean $\lambda > 0$. Now we switch to the common notation to θ to mean λ as parameter of interest.
A prior distribution for θ that is in some sense natural is the Gamma distribution.

Important remark 19 (Gamma-Poisson model). We assume:

- a Poisson data generating process. Conditionally on unknown mean θ each exchangeable/independent rv is distributed according to

$$p(X|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

and thus the **likelihood** for a sample of size n

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

- a Gamma **prior** distribution for mean θ. Has positive only support (which is what we want) $\theta > 0$ and depends on two shape parameters, positive as well; two parametrization are used the following is preferred.
  The density depends on two parameters α, λ (ie $G(\alpha, \lambda)$)

$$p(\theta|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\lambda\theta)$$

  In this parametrization

$$E(\theta) = \frac{\alpha}{\lambda}$$
$$V(\theta) = \frac{\alpha}{\lambda^2}$$

- the **posterior** $p(\theta|\alpha, ..., x)$ is a Gamma as well and can be obtained according

$$\begin{aligned} p(\theta|\alpha, \lambda, x) &\propto e^{-n\theta}\theta^{\sum_{i=1}^n x_i} \cdot \theta^{\alpha-1}e^{-\lambda\theta} \\ &= e^{-\theta(n+\lambda)} \cdot \theta^{\alpha+\sum_{i=1}^n x_i - 1} \end{aligned}$$

  we recognize the kernel of a gamma, that is $G(\alpha', \lambda')$, where

$$\alpha' = \alpha + \sum_{i=1}^n x_i$$
$$\lambda' = \lambda + n$$

  for which

$$E(\theta|\alpha, \lambda, x) = \frac{\alpha'}{\lambda'} = \frac{\alpha + \sum_{i=1}^n x_i}{\lambda + n}$$
$$V(\theta|\alpha, \lambda, x) = \frac{\alpha'}{\lambda'^2} = \frac{\alpha + \sum_{i=1}^n x_i}{(\lambda + n)^2}$$

Remark 29. Looking at the posterior expectation we can interpret the parameters of prior and likelihood:

- λ is interpreted as number of prior observations (while n as sample size of the experiment)
- α is interpreted as sum of counts from the λ prior observations (while $\sum_{i=1}^n x_i$ is for the experiment)

Thus the posterior expected value is a weighted mean between prior mean and experiment sample mean (with weights based on number of observations).

Example 2.5.1. [Birth rate] A Survey was conducted to estimate the mean number of children women without (θ_1) and with (θ_2) bachelor degree.
Let's assume that the prior for that mean are distributed both according

$$\theta_1, \theta_2 \sim Ga(\alpha = 2, \lambda = 1)$$

with a common expected value of 2 children per woman.
The survey gathered data on number of children in the two groups ($n_1 = 111$ women without degree, $n_2 = 44$ women with), and the mean of children per woman was higher in the without bachelor degree mothers:

$$n_1 = 111, \sum_{i=1}^{n_1} Y_{i,1} = 217 \implies \bar{Y}_1 = 1.95$$
$$n_2 = 44, \sum_{i=1}^{n_2} Y_{i,2} = 66 \implies \bar{Y}_1 = 1.50$$

Assuming a Poisson model is appropriate to describe/synthetize the empirical distribution, a posterior distribution for the two parameters are easily two gammas

$$\theta_1 | \left\{ n_1 = 111, \sum_{i=1}^{n_1} Y_{i,1} = 217 \right\} \sim Ga(2 + 217, 1 + 111) = Ga(219, 112)$$
$$\theta_2 | \left\{ n_2 = 44, \sum_{i=1}^{n_2} Y_{i,2} = 66 \right\} \sim Ga(2 + 66, 1 + 44) = Ga(68, 45)$$

The posterior means for θ_1, θ_2 becomes $219/112 = 1.95$ and $68/45 = 1.51$, so we moved way far from the initial 2 mean value for the women with degree.

Remark 30 (Alternative parametrization of the Gamma). Using the second parametrization:

- for the prior density we define $\beta = \frac{1}{\lambda}$ and thus the density becomes (eg $G(\alpha, \beta)$):

$$p(\theta|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \theta^{\alpha-1} \exp\left(-\frac{\theta}{\beta}\right)$$

  For this parametrization

$$E(\theta) = \alpha\beta$$
$$V(\theta) = \alpha\beta^2$$

- the posterior using the second parametrization

$$\begin{aligned} p(\theta|\alpha, \beta, x) &\propto e^{-n\theta} \theta^{\sum_{i=1}^n x_i} \cdot \theta^{\alpha-1} e^{-\frac{\theta}{\beta}} \\ &= e^{-\theta(n+\frac{1}{\beta})} \cdot \theta^{\alpha+\sum_{i=1}^n x_i - 1} \end{aligned}$$

  If we substitute

$$\beta' = \left(n + \frac{1}{\beta}\right)^{-1} = \left(\frac{\beta n + 1}{\beta}\right)^{-1} = \frac{\beta}{\beta n + 1}$$

  we recognize the kernel of a Gamma, $G(\alpha', \beta')$ where

$$\alpha' = \alpha + \sum_{i=1}^n x_i$$
$$\beta' = \frac{\beta}{\beta n + 1}$$

for which

$$E(\theta|\alpha,\beta,x) = \left(\alpha + \sum_{i=1}^{n} x_i\right)\left(\frac{\beta}{\beta n+1}\right)$$
$$V(\theta|\alpha,\beta,x) = \left(\alpha + \sum_{i=1}^{n} x_i\right)\left(\frac{\beta}{\beta n+1}\right)^2$$

Important remark 20. Looking, for instance, at the expectations, the two parametrizations are equivalent since

$$\frac{\beta}{\beta n+1} = \frac{1}{n+\lambda}.$$

2.6 Natural conjugate distributions

Remark 31. We have seen, among other, that beta prior distribution and binomial sampling model lead to a beta posterior distribution.
To reflect this we say that *the class of beta priors is conjugate for the binomial sampling model.*
It is desirable to have a prior such that the posterior has a tractable form and is algebraically convenient/known.

Definition 2.6.1 (Conjugacy)**.** A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(x|\theta)$ if the posterior is in the same class of distributions

$$p(\theta) \in \mathcal{P} \implies p(\theta|x) \in \mathcal{P}$$

Remark 32. In other words, *conjugacy* is the property that the posterior distribution follows the same parametric form as the prior distribution.

2.6.1 Sufficient statistics

Remark 33. Sufficiency is a property of a statistic/function T (e.g. the sum of cases) computed on a sample dataset $(x_1, \ldots, x_n)$, in relation to a parametric model.
Informally speaking, a sufficient statistic contains all of the information that the dataset provides about the model parameters.

Remark 34. The following theorem provides a characterization of a sufficient statistic

Theorem 2.6.1 (Fisher-Neyman factorization theorem)**.** *A statistic t is sufficient for θ given the sample x if and only if there are functions f and g such that:*

$$p(x|\theta) = g(x)f(t|\theta)$$

Remark 35. Often we have $g(x) = 1$, so only $f(t|\theta)$ remains.

Important remark 21 (Sufficient statistic). If $t = T(x)$ is a sufficient statistic for the sample x:

$$p(\theta|x) = p(\theta|t)$$

In the bayesian framework being sufficient implies that

$$p(\theta|x) = p(\theta|t) \propto p(\theta)p(t|\theta)$$

Remark 36. Only likelihoods that admit sufficient statistics are considered.

2.6.2 One-parameter exponential family

Definition 2.6.2. A density belongs to the one-parameter exponential family if:

- for one observation, if it can be expressed in the form:

$$p(x|\theta) = L(\theta; x) = g(x)h(\theta)\exp\left[t(x)\Psi(\theta)\right]$$

- for n independent observations, if the likelihood of the sample $p(x|\theta)$ can be expressed as:

$$L(\theta; x) \propto h(\theta)^n \exp\left[\sum t(x_i)\Psi(\theta)\right]$$

  where $g(x)$ can be omitted since it is not a function of the parameter, and $\sum t(x_i)$ is a sufficient statistic for θ.

2.6.2.1 Examples of distributions belonging to this family

Remark 37. The following examples show how different distributions belong to the exponential family.

Example 2.6.1 (Normal with known variance)**.**

$$\begin{aligned} p(x \mid \theta) &= (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\phi}(x-\theta)^2\right\} \\ &= \underbrace{(2\pi\phi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^2\frac{1}{\phi}\right)}_{g(x)} \underbrace{\exp\left(-\frac{\theta^2}{2\phi}\right)}_{h(\theta)} \underbrace{\exp\left(\frac{x\theta}{\phi}\right)}_{\exp[t(x)\Psi(\theta)]} \end{aligned}$$

Example 2.6.2 (Normal with known mean)**.**

$$p(x \mid \phi) = \underbrace{(2\pi)^{-\frac{1}{2}}}_{g(x)} \underbrace{\phi^{-\frac{1}{2}}}_{h(\phi)} \underbrace{\exp\left\{-\frac{1}{2\phi}(x-\theta)^2\right\}}_{\exp[t(x)\Psi(\phi)]}$$

Example 2.6.3 (Poisson)**.**

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

note that $\ln(\lambda^x) = x\ln\lambda$ thus $\lambda^x = \exp(x\ln x)$, thus

$$p(x \mid \lambda) = \underbrace{\frac{1}{x!}}_{g(x)} \underbrace{\exp(-\lambda)}_{h(\lambda)} \underbrace{\exp\left[x\ln\lambda\right]}_{\exp[t(x)\Psi(\lambda)]}$$

Example 2.6.4 (Binomial)**.**

$$\begin{aligned} p(x \mid \pi) &= \binom{n}{x} \pi^x (1-\pi)^{n-x} \\ &= \binom{n}{x} (1-\pi)^n \pi^x (1-\pi)^{-x} \end{aligned}$$

considering that

$$\ln\left(\pi^x (1-\pi)^{-x}\right) = x \ln \frac{\pi}{1-\pi}$$

then

$$\pi^x (1-\pi)^{-x} = \exp\left[x \ln \frac{\pi}{1-\pi}\right]$$

so that, finally

$$p(x \mid \pi) = \underbrace{\binom{n}{x}}_{g(x)} \underbrace{(1-\pi)^n}_{h(\pi)} \underbrace{\exp\left[x \ln \frac{\pi}{1-\pi}\right]}_{\exp[t(x)\Psi(\pi)]}$$

Example 2.6.5 (Exponential)**.**

$$p(x \mid \theta) = \theta \exp(-\theta x) = \underbrace{\theta}_{h(\theta)} \underbrace{\exp(-\theta x)}_{\exp[t(x)\Psi(\theta)]} \underbrace{1}_{g(x)}$$

2.6.2.2 Relationship between conjugacy and exponential family

Remark 38. There's a relation between exponential family and conjugacy

Important remark 22. If the experiment produces data belonging to a distribution of the exponential family, having thus likelihood

$$L(\theta; x) \propto h(\theta)^n \exp\left[\sum t(x_i)\Psi(\theta)\right]$$

the conjugate prior $\mathcal{P}$ belongs to the family with density

$$p(\theta) \propto h(\theta) \exp\left\{\tau \Psi(\theta)\right\}$$

where τ stresses the fact that no observation related to the experiment can be considered.

Example 2.6.6. Some examples of experiments with likelihood belonging to the one parameter exponential family (and depend on the sufficient statistic for the parameter) are reported in table 2.8

2.6.3 Two-parameters exponential family

Definition 2.6.3. A probability density belongs to the two parameters exponential family:

- for one observation, if it can be expressed in the form:

$$p(x|\theta, \varphi) = L(\theta, \varphi; x) = g(x)h(\theta, \varphi) \exp\left[t(x)\Psi(\theta, \varphi) + u(x)\chi(\theta, \varphi)\right],$$

Likelihood	Conjugate Prior	Case/Naming
Binomial	Beta	Beta-Binomial
Normal (known variance)	Normal	Normal-Normal
Normal (known mean)	Inverse-gamma	Normal-Inverse-gamma
Poisson	Gamma	Gamma-Poisson

Table 2.8: Likelihood and conjugate priors: notable examples.

- for n independent observations if the likelihood of the sample $p(x|\theta,\varphi)$ can be expressed as

$$L(\theta,\varphi;x) \propto h(\theta,\varphi)^n \exp\left[\sum t(x_i)\Psi(\theta,\varphi) + \sum u(x_i)\chi(\theta,\varphi)\right],$$

  where $g(x)$ is not considered since it is not a function of the parameter.

Remark 39. The relation between exponential family and conjugacy becomes the following

Important remark 23. In this case the family of conjugate densities has the form:

$$p(\theta,\varphi) \propto h(\theta,\varphi)^n \exp\left[\tau\Psi(\theta,\varphi) + \nu\chi(\theta,\varphi)\right]$$

Important remark 24. Here $(\sum t(x_i), \sum u(x_i))$ is a sufficient statistic for the bidimensional vector (θ,φ) given x.