9 Bayesian inference

9.1 Subjective probability

This is probability regarded as *degree of belief*.

A subjective probability of an event $A$ is assessed as $p$ if you are prepared to stake £$pM$ to win £$M$ and equally prepared to accept a stake of £$pM$ to win £$M$.

In other words ...

... the bet is fair and you are assumed to behave rationally.

9.1.1 Kolmogorov's axioms

*How does subjective probability fit in with the fundamental axioms?*

Let $\mathcal{A}$ be the set of all subsets of a countable sample space $\Omega$. Then

(i) $P(A) \geq 0$ for every $A \in \mathcal{A}$;

(ii) $P(\Omega) = 1$;
(iii) If \( \{ A_\lambda : \lambda \in \Lambda \} \) is a countable set of mutually exclusive events belonging to \( \mathcal{A} \), then
\[
P \left( \bigcup_{\lambda \in \Lambda} A_\lambda \right) = \sum_{\lambda \in \Lambda} P(A_\lambda).
\]

Obviously the subjective interpretation has no difficulty in conforming with (i) and (ii). (iii) is slightly less obvious.

Suppose we have 2 events \( A \) and \( B \) such that \( A \cap B = \emptyset \). Consider a stake of \( \mathcal{L}p_A M \) to win \( \mathcal{L}M \) if \( A \) occurs and a stake \( \mathcal{L}p_B M \) to win \( \mathcal{L}M \) if \( B \) occurs. The total stake for bets on \( A \) or \( B \) occurring is \( \mathcal{L}p_A M + \mathcal{L}p_B M \) to win \( \mathcal{L}M \) if \( A \) or \( B \) occurs. Thus we have \( \mathcal{L}(p_A + p_B) M \) to win \( \mathcal{L}M \) and so
\[
P(A \cup B) = P(A) + P(B).
\]

**9.1.2 Conditional probability**

Define \( p_B \), \( p_{AB} \), \( p_A | B \) such that:
- \( \mathcal{L}p_B M \) is the fair stake for \( \mathcal{L}M \) if \( B \) occurs;
- \( \mathcal{L}p_{AB} M \) is the fair stake for \( \mathcal{L}M \) if \( A \) and \( B \) occur;
- \( \mathcal{L}p_A | B \) is the fair stake for \( \mathcal{L}M \) if \( A \) occurs given \( B \) has occurred — otherwise the bet is off.

Consider the 3 alternative outcomes \( A \) only, \( B \) only, \( A \) and \( B \). If \( G_1, G_2, G_3 \) are the gains, then
\[
\begin{align*}
A : & \quad -p_B M_2 - p_{AB} M_3 = G_1 \\
B : & \quad -p_{A|B} M_1 + (1 - p_B) M_2 - p_{AB} M_3 = G_2 \\
A \cap B : & \quad (1 - p_{A|B}) M_1 + (1 - p_B) M_2 + (1 - p_{AB}) M_3 = G_3
\end{align*}
\]

The principle of rationality does not allow the possibility of setting the stake to obtain sure profit. Thus
\[
\begin{vmatrix}
0 & -p_B & -p_{AB} \\
-p_{A|B} & 1 - p_B & -p_{AB} \\
1 - p_{A|B} & 1 - p_B & 1 - p_{AB}
\end{vmatrix} = 0
\]

\[\Rightarrow \quad p_{AB} - p_B p_{A|B} = 0\]

or
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}.
\]
9.2 Estimation formulated as a decision problem

\( \mathcal{A} \) The action space, which is the set of all possible actions \( a \in \mathcal{A} \) available.

\( \Theta \) The parameter space, consisting of all possible “states of nature”, only one of which occurs or will occur (this “true” state being unknown).

\( L \) The loss function, having domain \( \Theta \times \mathcal{A} \) (the set of all possible consequences \( (\theta, a) \)) and codomain \( \mathbb{R} \).

\( \mathbb{R}_X \) Containing all possible realisations \( x \in \mathbb{R}_X \) of a random variable \( X \) belonging to the family \( \{ f(x; \theta) : \theta \in \Theta \} \).

\( \mathcal{D} \) The decision space, consisting of all possible decisions \( \delta \in \mathcal{D} \), each decision function having domain \( \mathbb{R}_X \) and codomain \( \delta \).

9.2.1 The basic idea

- The true state of the world is unknown when the action is chosen.
- The loss is known which results from each of the possible consequences \( (\theta, a) \).
- Data are collected to provide information about the unknown \( \theta \).
- A choice of action is taken for each possible observation of \( X \).

Decision theory is concerned with the criteria for making a choice.

*The form of the loss function is crucial — it depends upon the nature of the problem.*

Example

You are a fashion buyer and must stock up with trendy dresses. The true number you will sell is \( \theta \) (unknown).
You stock up with \( a \) dresses. If you overstock you must sell off at your end-of-season sale and lose \( LA \) per dress. If you understock you lose the profit of \( LB \) per dress. Thus

\[
L(\theta, a) = \begin{cases} 
  A(a - \theta), & a \geq \theta, \\
  B(\theta - a), & a \leq \theta.
\end{cases}
\]

**Special forms of loss function**

If \( A = B \), you have *modular loss*,

\[
L(\theta, a) = A |\theta - a|
\]

To suit comparison with classical inference, *quadratic loss* is often used.

\[
L(\theta, a) = C(\theta - a)^2
\]

### 9.2.2 The risk function

This is a measure of the quality of a decision function. Suppose we choose a decision \( \delta(x) \) (i.e. choose \( a = \delta(x) \)).

The function \( R \) with domain \( \Theta \times \mathcal{D} \) and codomain \( \mathbb{R} \) defined by

\[
R(\theta, \delta) = \int_{\mathbb{R}^X} L(\theta, \delta(x)) f(x; \theta) dx
\]

or

\[
R(\theta, \delta) = \sum_{x \in \mathbb{R}^X} L(\theta, \delta(x)) p(x; \theta)
\]

is called the *risk function*.

We may write this as

\[
R(\theta, \delta) = E[L(\theta, \delta(x))].
\]

**Example**

Suppose \( X \) is a random sample from \( N(\theta, \theta) \) and suppose we assume quadratic loss.
Consider
\[ \delta_1(X) = \bar{X}, \quad \delta_2(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \]
\[ R(\theta, \delta_1) = E \left[ (\theta - \bar{X})^2 \right]. \]
We know that \( E(X) = \theta \), so
\[ R(\theta, \delta_1) = V(\bar{X}) = \frac{\theta}{n}. \]
\[ R(\theta, \delta_2) = E \left[ (\theta - \delta_2(X))^2 \right]. \]
But \( E[\delta_2(X)] = \theta \), so that
\[ R(\theta, \delta_2) = V[\delta_2(X)] \]
\[ = \frac{\theta^2}{(n-1)^2} V \left[ \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\theta} \right] \]
\[ = \frac{\theta^2}{(n-1)^2} \cdot 2(n-1) = \frac{2\theta^2}{n-1}. \]

Without knowing \( \theta \), how do we choose between \( \delta_1 \) and \( \delta_2 \)?

**Two sensible criteria**

1. Take precautions against the worst.
2. Take into account any further information you may have.
Criterion 1 leads to minimax. Choose the decision leading to the smallest maximum risk, i.e. choose $\delta^* \in D$ such that
\[
\sup_{\theta \in \Theta} \{ R(\theta, \delta^*) \} = \inf_{\delta \in D} \sup_{\theta \in \Theta} \{ R(\theta, \delta) \}.
\]

Minimax rules have disadvantages, the main one being that they treat Nature as an adversary. Minimax guards against the worst choice. The value of $\theta$ should not be regarded as a value deliberately chosen by Nature to make the risk large.

Fortunately in many situations they turn out to be reasonable in practice, if not in concept.

Criterion 2 involves introducing beliefs about $\theta$ into the analysis. If $\pi(\theta)$ represents our beliefs about $\theta$, then define Bayes risk $r(\pi, \delta)$ as
\[
\begin{align*}
 r(\pi, \delta) &= \left\{ \begin{array}{ll}
 \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta & \text{if } \theta \text{ is continuous}, \\
 \sum_{\theta \in \Theta} R(\theta, \delta) \pi(\theta) & \text{if } \theta \text{ is discrete}.
\end{array} \right.
\end{align*}
\]

Now choose $\delta^*$ such that
\[
r(\pi, \delta^*) = \inf_{\delta \in D} \{ r(\pi, \delta) \}.
\]

**Bayes Rule**

The *Bayes rule* with respect to a prior $\pi$ is the decision rule $\delta^*$ that minimises the Bayes risk among all possible decision rules.

Bayes' tomb
9.2.3 Decision terminology

\( \delta^\ast \) is called a *Bayes decision function*.

\( \delta(X) \) is a *Bayes estimator*.

\( \delta(x) \) is a *Bayes estimate*.

If a decision function \( \delta' \) is such that \( \exists \delta'' \) satisfying

\[
R(\theta, \delta'') \leq R(\theta, \delta') \quad \forall \theta \in \Theta
\]

and

\[
R(\theta, \delta'') < R(\theta, \delta') \quad \text{for some} \ \theta \in \Theta,
\]

then \( \delta' \) is *dominated* by \( \delta'' \) and \( \delta' \) is said to be *inadmissible*.

All other decisions are *admissible*.
9.3 Prior and posterior distributions

The prior distribution represents our initial belief in the parameter \( \theta \). This means that \( \theta \) is regarded as a random variable with p.d.f. \( \pi(\theta) \).

From Bayes Theorem,

\[
f(x \mid \theta)\pi(\theta) = \pi(\theta \mid x)f(x)
\]

or

\[
\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)d\theta}.
\]

\( \pi(\theta \mid x) \) is called the posterior p.d.f.

It represents our updated belief in \( \theta \), having taken account of the data.

We write the expression in the form

\[
\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)
\]

or

\[\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.\]

**Example**

Suppose \( X_1, X_2, \ldots, X_n \) is a random sample from a Poisson distribution, \( Pois(\theta) \), where \( \theta \) has a prior distribution

\[
\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta}}{\Gamma(\alpha)}, \quad \theta \geq 0, \alpha > 0.
\]

The posterior distribution is given by

\[
\pi(\theta \mid x) \propto \theta^{\sum x_i + \alpha - 1}e^{-\theta(n+1)} \prod x_i! \cdot \frac{\theta^{\alpha-1}e^{-\theta}}{\Gamma(\alpha)}
\]

which simplifies to

\[
\pi(\theta \mid x) \propto \theta^{\sum x_i + \alpha - 1}e^{-\theta(n+1)}, \quad \theta \geq 0.
\]

This is recognisable as a \( \Gamma \)-distribution and, by comparison with \( \pi(\theta) \), we can write

\[
\pi(\theta \mid x) = \frac{(n + 1)^{\sum x_i + \alpha} \theta^{\sum x_i + \alpha - 1}e^{-(n+1)\theta}}{\Gamma(\sum x_i + \alpha)}, \quad \theta \geq 0,
\]

which represents our posterior belief in \( \theta \).
9.4 Decisions

We have already defined the Bayes risk to be

\[ r(\pi, \delta) = \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta \]

where

\[ R(\theta, \delta) = \int_{\mathbb{R}^X} L(\theta, \delta(x)) f(x \mid \theta) dx \]

so that

\[ r(\pi, \delta) = \int_{\Theta} \int_{\mathbb{R}^X} L(\theta, \delta(x)) f(x \mid \theta) dx \pi(\theta) d\theta \]

\[ = \int_{\mathbb{R}^X} \int_{\Theta} L(\theta, \delta(x)) \pi(\theta \mid x) d\theta f(x) dx \]

For any given \( x \), minimising the Bayes risk is equivalent to minimising

\[ \int_{\Theta} L(\theta, \delta(x)) \pi(\theta \mid x) d\theta. \]

i.e. we minimise the posterior expected loss.

Example

Take the posterior distribution in the previous example and, for no special reason, assume quadratic loss

\[ L(\theta, \delta(x)) = |\theta - \delta(x)|^2. \]

Then we minimise

\[ \int_{\Theta} |\theta - \delta(x)|^2 \pi(\theta \mid x) d\theta \]

by differentiating with respect to \( \delta \) to obtain

\[ -2 \int_{\Theta} |\theta - \delta(x)| \pi(\theta \mid x) d\theta = 0 \]

\[ \Rightarrow \delta(x) = \int_{\Theta} \theta \pi(\theta \mid x) d\theta. \]
This is the mean of the posterior distribution.
\[
\delta(x) = \int_0^{\infty} \frac{(n+1)^{\sum x_i + \alpha} \theta^{\sum x_i + \alpha} e^{-(n+1)\theta}}{\Gamma(\sum x_i + \alpha)} d\theta
\]
\[
= \frac{(n+1)^{\sum x_i + \alpha} \Gamma(\sum x_i + \alpha + 1)}{\Gamma(\sum x_i + \alpha)(n+1)^{\sum x_i + \alpha + 1}}
\]
\[
= \frac{\sum x_i + \alpha}{n + 1}
\]

**Theorem:**
Suppose that \(\delta^\pi\) is a decision rule that is Bayes with respect to some prior distribution \(\pi\). If the risk function of \(\delta^\pi\) satisfies
\[
R(\theta, \delta^\pi) \leq r(\pi, \delta^\pi) \quad \text{for all } \theta \in \Theta,
\]
then \(\delta^\pi\) is a minimax decision rule.

**Proof:**
Suppose \(\delta^\pi\) is not minimax. Then there is a decision rule \(\delta'\) such that
\[
\sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta^\pi).
\]
For this decision rule we have, since, if \(f(x) \leq k\), then \(E(f(X)) \leq k\),
\[
r(\pi, \delta') \leq \sup_{\theta \in \Theta} R(\theta, \delta')
\]
\[
< \sup_{\theta \in \Theta} R(\theta, \delta^\pi)
\]
\[
\leq r(\pi, \delta^\pi),
\]
contradicting the statement that \(\delta^\pi\) is Bayes with respect to \(\pi\). Hence \(\delta^\pi\) is minimax.

**Example**
Let \(X_1, \ldots, X_n\) be a random sample from a Bernoulli distribution \(B(1, p)\). Let \(Y = \sum X_i\) and consider estimating \(p\) using the loss function
\[
L(p, a) = \frac{(p - a)^2}{p(1-p)},
\]
a loss which penalises mistakes more if \( p \) is near 0 or 1. The m.l.e. is \( \hat{p} = Y/n \) and

\[
R(p, \hat{p}) = E \left( \frac{(\hat{p} - p)^2}{p(1 - p)} \right) = \frac{1}{p(1 - p)} \cdot \frac{p(1 - p)}{n} = \frac{1}{n}
\]

The risk is constant and, therefore, minimax.

[\text{NB A decision rule with constant risk is called an equalizer rule}]\

We now show that \( \hat{p} \) is the Bayes estimator with respect to the \( U(0, 1) \) prior which will imply that \( \hat{p} \) is minimax.

\[
\pi(\theta \mid y) \propto p^y(1 - p)^{n-y}
\]

so we want the value of \( a \) which minimises

\[
\int_0^1 \frac{(p - a)^2}{p(1-p)} \cdot p^y(1-p)^{n-y} dp.
\]

Differentiation with respect to \( a \) yields

\[
a = \frac{\int_0^1 p^y(1-p)^{n-y-1} dp}{\int_0^1 p^{y-1}(1-p)^{n-y-1} dp}.
\]

Now a \( \beta \)-distribution with parameters \( \alpha, \beta \) has pdf

\[
f(x) = \frac{\Gamma(\alpha + \beta)x^{\alpha-1}(1 - x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, \quad x \in [0, 1],
\]

so

\[
a = \frac{\Gamma(y + 1)\Gamma(n - y)}{\Gamma(n + 1)\Gamma(x)\Gamma(n - x)}
\]

and, using the identity \( \Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \),

\[
a = \frac{y}{n}, \quad i.e. \  \hat{p} = Y/n.
\]
9.5 Bayes estimates

(i) The mean of \( \pi(\theta \mid x) \) is the Bayes estimate with respect to quadratic loss.

**Proof**

Choose \( \delta(x) \) to minimise

\[
\int_\Theta (\theta - \delta(x))^2 \pi(\theta \mid x) d\theta
\]

by differentiating with respect to \( \delta \) and equating to zero. Then

\[
\delta(x) = \int_\Theta \theta \pi(\theta \mid x) d\theta
\]

since \( \int_\Theta \pi(\theta \mid x) d\theta = 1 \).

(ii) The median of \( \pi(\theta \mid x) \) is the Bayes estimate with respect to modular loss (i.e. absolute value loss).

**Proof**

Choose \( \delta(x) \) to minimise

\[
\int_\Theta |\theta - \delta(x)| \pi(\theta \mid x) d\theta
\]

\[
= \int_{-\infty}^{\delta(x)} (\delta(x) - \theta) \pi(\theta \mid x) d\theta + \int_{\delta(x)}^{\infty} (\theta - \delta(x)) \pi(\theta \mid x) d\theta
\]

Differentiate with respect to \( \delta \) and equate to zero.

\[
\int_{-\infty}^{\delta(x)} \pi(\theta \mid x) d\theta = \int_{\delta(x)}^{\infty} \pi(\theta \mid x) d\theta
\]

\[
\Rightarrow 2 \int_{-\infty}^{\delta(x)} \pi(\theta \mid x) d\theta = \int_{-\infty}^{\infty} \pi(\theta \mid x) d\theta = 1
\]

so that

\[
\int_{-\infty}^{\delta(x)} \pi(\theta \mid x) d\theta = \frac{1}{2},
\]

and \( \delta(x) \) is the median of \( \pi(\theta \mid x) \).
(iii) The mode of \( \pi(\theta \mid x) \) is the Bayes estimate with respect to zero-one loss (i.e. loss of zero for a correct estimate and loss of one for any incorrect estimate).

**Proof**

The loss function has the form

\[
L(\theta, \delta(x)) = \begin{cases} 
1, & \text{if } \delta(x) \neq \theta, \\
0, & \text{if } \delta(x) = \theta.
\end{cases}
\]

This is only a useful idea for a discrete distribution.

\[
\sum_{\theta \in \Theta} L(\theta, \delta(x)) \pi(\theta \mid x) = \sum_{\theta \in \Theta \setminus \{\theta^*\}} \pi(\theta \mid x) = 1 - \pi(\theta^* \mid x)
\]

where \( \delta(x) = \theta^* \).

Minimisation means maximising \( \pi(\theta^* \mid x) \), so \( \delta(x) \) is taken to be the ‘most likely’ value – the mode.

Remember that \( \pi(\theta \mid x) \) represents our most up-to-date belief in \( \theta \) and is all that is needed for inference.
9.6 Credible intervals

Definition
The interval \((\theta_1, \theta_2)\) is a \(100(1 - \alpha)\%\) credible interval for \(\theta\) if
\[
\int_{\theta_1}^{\theta_2} \pi(\theta \mid x) \, d\theta = 1 - \alpha, \quad 0 \leq \alpha \leq 1.
\]

Both \((\theta_1, \theta_2)\) and \((\theta'_1, \theta'_2)\) are \(100(1 - \alpha)\%\) credible intervals.

Definition
The interval \((\theta_1, \theta_2)\) is a \(100(1 - \alpha)\%\) highest posterior density interval if

(i) \((\theta_1, \theta_2)\) is a \(100(1 - \alpha)\%\) credible interval;

(ii) \(\forall \theta' \in (\theta_1, \theta_2)\) and \(\theta'' \notin (\theta_1, \theta_2)\),
\[
\pi(\theta' \mid x) \geq \pi(\theta'' \mid x)
\]

Obviously this defines the shortest possible \(100(1 - \alpha)\%\) credible interval.
9.7 Features of Bayesian models

(i) If there is a sufficient statistic, say $T$, the posterior distribution will depend upon the data only through $T$.

(ii) There is a natural parametric family of priors such that the posterior distributions also belong to this family. Such priors are called \textit{conjugate priors}.

\textbf{Example: Binomial likelihood, beta prior}

A binomial likelihood has the form

$$p(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, \ldots, n,$$

and a beta prior is

$$\pi(\theta) = \frac{\theta^{a-1}(1 - \theta)^{b-1}}{B(a, b)}, \quad \theta \in [0, 1],$$

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.$$

Posterior $\propto$ Likelihood $\times$ Prior,

so that

$$\pi(\theta \mid x) \propto \theta^{x+a-1}(1 - \theta)^{n-x+b-1}$$

and comparison with the prior leads to

$$\pi(\theta \mid x) = \frac{\theta^{x+a-1}(1 - \theta)^{n-x+b-1}}{B(x + a, n - x + b)}, \quad \theta \in [0, 1],$$

which is also a beta distribution.
9.8 Hypothesis testing

The Bayesian only has a concept of a hypothesis test in certain very specific situations. This is because of the alien idea of having to regard a parameter as a fixed point. However the concept has some validity in an example such as the following.

Example:

Water samples are taken from a reservoir and checked for bacterial content. These provide information about \( \theta \), the proportion of infected samples. If it is believed that bacterial density is such that more than 1% of samples are infected, there is a legal obligation to change the chlorination procedure.

Bayesian procedure is

(i) use the data to obtain the posterior distribution of \( \theta \);

(ii) calculate \( P(\theta > 0.01) \).

EC rules are formulated in statistical terms and, in such cases, allow for a 5% significance level. This could be interpreted in a Bayesian sense as allowing for a wrong decision with probability no more than 0.05, so order increased chlorination if \( P(\theta > 0.01) > 0.05 \).

Where for some reason (e.g. legal reason) a fixed parameter value is of interest because it defines a threshold, then a Bayesian concept of a test is possible. It is usually performed either by calculating a credible bound or a credible interval and looking at it in relation to that threshold value or values, or by calculating posterior probabilities based on the threshold or thresholds.
9.9 Inferences for normal distributions

9.9.1 Unknown mean, known variance

Consider data from $N(\theta, \sigma^2)$ and a prior $N(\tau, \kappa^2)$.

Posterior $\propto$ Likelihood $\times$ Prior.

so, since

$$f(x \mid \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \theta)^2\right)$$

and

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\kappa^2}} \exp\left(-\frac{1}{2\kappa^2}(\theta - \tau)^2\right), \quad \theta \in \mathbb{R},$$

$$\pi(\theta \mid x, \sigma^2, \tau, \kappa^2) \propto \exp\left(-\frac{n}{2\sigma^2}(\theta^2 - 2\theta\bar{x}) - \frac{1}{2\kappa^2}(\theta^2 - 2\theta\tau)\right).$$

But

$$\frac{n}{2\sigma^2}(\theta^2 - 2\theta\bar{x}) + \frac{1}{2\kappa^2}(\theta^2 - 2\theta\tau)$$

$$= \frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\kappa^2}\right) \theta^2 - \left(\frac{n\bar{x}}{\sigma^2} + \frac{\tau}{\kappa^2}\right) \theta + \text{garbage}$$

$$= \frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\kappa^2}\right) \left(\theta - \frac{n\bar{x}/\sigma^2 + \tau/\kappa^2}{n/\sigma^2 + 1/\kappa^2}\right)^2$$

so the posterior p.d.f. is proportional to

$$\exp\left[-\frac{1}{2(n/\sigma^2 + 1/\kappa^2)^{-1}} \left(\theta - \frac{n\bar{x}/\sigma^2 + \tau/\kappa^2}{n/\sigma^2 + 1/\kappa^2}\right)^2\right]$$

This means that

$$\theta \mid x, \sigma^2, \tau, \kappa^2 \sim N\left(\frac{n\bar{x}/\sigma^2 + \tau/\kappa^2}{n/\sigma^2 + 1/\kappa^2}, (n/\sigma^2 + 1/\kappa^2)^{-1}\right).$$

The form is a weighted average.

The prior mean $\tau$ has weight $1/\kappa^2$.

this is $1/($prior variance),...
and the observed sample mean has weight $n / \sigma^2$: 
this is $1 / (\text{variance of sample mean})$.

This is reasonable because

$$\lim_{n \to \infty} \left( \frac{n \bar{x} / \sigma^2 + \tau / \kappa^2}{n / \sigma^2 + 1 / \kappa^2} \right) = \bar{x}$$

and

$$\lim_{\kappa^2 \to \infty} \left( n \bar{x} / \sigma^2 + \tau / \kappa^2 \right) = \bar{x}.$$ 

The posterior mean tends to $\bar{x}$ as the amount of data swamps the prior or as prior beliefs become very vague.

### 9.9.2 Unknown variance, known mean

We need to assign a prior p.d.f. $\pi(\sigma^2)$.
The $\Gamma$ family of p.d.f.'s provides a flexible set of shapes over $[0, \infty)$. We take

$$\tau = \frac{1}{\sigma^2} \sim \Gamma \left( \frac{\lambda}{2}, \frac{\nu \lambda}{2} \right)$$

where $\lambda$ and $\nu$ are chosen to provide suitable location and shape.

In other words, the prior p.d.f. of $\tau = 1 / \sigma^2$ is chosen to be of the form

$$\pi(\tau) = \frac{(\nu \lambda)^{\nu/2} \tau^{\nu/2-1}}{2^{\nu/2} \Gamma(\nu/2)} \exp \left( -\frac{\nu \lambda \tau}{2} \right), \quad \tau \geq 0.$$ 

or

$$\pi(\sigma^2) = \frac{(\nu \lambda)^{\nu/2} (\sigma^2)^{\nu/2-1}}{2^{\nu/2} \Gamma(\nu/2)} \exp \left( -\frac{\nu \lambda \sigma^2}{2\sigma^2} \right).$$

$\tau = 1 / \sigma^2$ is called the **precision**.
The posterior p.d.f. is given by

$$\pi(\sigma^2 | \mathbf{x}, \theta) \propto f(\mathbf{x} | \theta, \sigma^2) \pi(\sigma^2),$$

and the right-hand side as a function of $\sigma^2$, is proportional to

$$(\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum (x_i - \theta)^2 \right) \times (\sigma^2)^{-\nu/2-1} \exp \left( -\frac{\nu \lambda}{2\sigma^2} \right).$$
Thus \( \pi(\sigma^2 \mid x, \theta) \) is proportional to

\[
(\sigma^2)^{-(\nu+n)/2-1} \exp \left( -\frac{1}{2\sigma^2} \left[ \nu \lambda + \sum (x_i - \theta)^2 \right] \right)
\]

This is the same as the prior except that \( \nu \) is replaced by \( \nu + n \) and \( \lambda \) is replaced by

\[
\frac{\nu \lambda + \sum (x_i - \theta)^2}{\nu + n}.
\]

Consider the random variable

\[
W = \frac{\nu \lambda + \sum (x_i - \theta)^2}{\nu + n}.
\]

Clearly

\[
f_W(w) = C w^{(\nu+n)/2-1} e^{-w/2}, \quad w \geq 0.
\]

This is a \( \chi^2 \)-distribution with \( \nu + n \) degrees of freedom. In other words,

\[
\frac{\nu \lambda + \sum (x_i - \theta)^2}{\sigma^2} \sim \chi^2(\nu + n).
\]

Again this agrees with intuition. As \( n \to \infty \) we approach the classical estimate, and also as \( \nu \to 0 \) which corresponds to vague prior knowledge.

### 9.9.3 Mean and variance unknown

We now have to assign a joint prior p.d.f. \( \pi(\theta, \sigma^2) \) to obtain a joint posterior p.d.f.

\[
\pi(\theta, \sigma^2 \mid x) \propto f(x \mid \theta, \sigma^2) \pi(\theta, \sigma^2).
\]

We shall look at the special case of a non-informative prior. From theoretical considerations beyond the scope of this course, a non-informative prior may be approximated by

\[
\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}.
\]
Note that this is not to be thought of as an expression of prior beliefs but rather as an approximate prior which allows the posterior to be dominated by the data.

\[
\pi(\theta, \sigma^2 | x) \propto (\sigma^2)^{-n/2 - 1} \exp \left( -\frac{1}{2\sigma^2} \sum (x_i - \theta)^2 \right).
\]

The right-hand side may be written in the form

\[
(\sigma^2)^{-n/2 - 1} \exp \left( -\frac{1}{2\sigma^2} \left[ \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2 \right] \right).
\]

We can use this form to integrate out, in turn, \( \theta \) and \( \sigma^2 \).

Using

\[
\int_{-\infty}^{\infty} \exp \left( -\frac{n}{2\sigma^2}(\theta - \bar{x})^2 \right) d\theta = \sqrt{\frac{2\pi \sigma^2}{n}}
\]

we find

\[
\pi(\sigma^2 | x) \propto (\sigma^2)^{-n/2 - 1} \exp \left( -\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2 \right)
\]

and, putting \( W = \sum (x_i - \bar{x})^2 / \sigma^2 \),

\[
f_W(w) = C w^{(n-1)/2} e^{-w/2}, \quad w \geq 0,
\]

so we see that

\[
\frac{\sum (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n - 1).
\]

To integrate out \( \sigma^2 \), make the substitution \( \sigma^2 = \tau^{-1} \). Then \( \pi(\theta | x) \) is proportional to

\[
\int_0^\infty \tau^{n/2 - 1} \exp \left( -\frac{\tau}{2} \left[ \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2 \right] \right) d\tau.
\]

From the definition of the \( \Gamma \)-function,

\[
\int_0^\infty y^{\alpha - 1} e^{-\beta y} dy = \frac{\Gamma(\alpha)}{\beta^{\alpha}},
\]

and we see, since only the terms in \( \theta \) are relevant,

\[
\pi(\theta | x) \propto \left( \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2 \right)^{-n/2}.
\]
or

$$\pi(\theta | x) \propto \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

where

$$t = \frac{\sqrt{n}(\theta - \bar{x})}{\sqrt{\sum(x_i - \bar{x})^2/(n-1)}}.$$