Bayesian Econometrics Summary - lectures 1 to 6 based on checklist for exam

Bayesian Econometrics (Vrije Universiteit Amsterdam)

Summary Bayesian Econometrics P2

Jan-Willem de Jong

December 6, 2021

Abstract

Summary for the course Bayesian Econometrics

1 Lecture 1

1.1 Name four advantages of the Bayesian approach over the frequentist approach

1. It is possible to make probability statements about parameters θ.
2. There is the possibility to include prior knowledge in a quite natural way.
3. We have valid results for finite samples, which we don't have with asymptotic results.
4. We have a natural way to take into account uncertainty on parameter values (& possibly on model selection) in forecasts and decisions.

1.2 What is Bayes' rule?

VERY IMPORTANT

$$\Pr(\theta|Y) = \frac{\Pr(\theta)\Pr(Y|\theta)}{\Pr(Y)}$$

for parameters θ and data Y.

1.3 What is a prior, likelihood and posterior?

- A **prior** probability of parameter θ (short: prior) is the probability of θ *before* observing the data Y. Mathematically, we denote it as $\Pr(\theta)$.
- The **likelihood** (function) is defined as $\Pr(Y|\theta)$ which is the probability of Y given θ.
- The **posterior** probability of a parameter θ (short: posterior) is the probability of θ *after* observing the data Y.

1.4 Application of Bayes' rule to compute the posterior probabilities in example 1 & 2

1.4.1 Example 1

Later or not, because just literally the slides.

1.4.2 Example 2

Later or not, because just literally the slides.

1.5 What are the differences between the frequentist and Bayesian approaches in examples 1 & 2?

In *example 1*, the classical/frequentist approach is to perform a two-sided test and compute a p-value which is the probability of the outcome that is "at least as extreme" under the null hypothesis H_0. Then, decide whether we reject or not at a certain significance level. Note that we could also perform a one-sided test. Therefore, the conclusion can differ between different statisticians, because different significance levels and types of tests can be used.
The Bayesian approach starts with calculating the prior distribution of θ and the likelihood function in case $Y = 0$. Besides, we also find the (prior) probability that $Y = 0$. The Bayes' rule yields the posterior distribution of θ, the distribution in which both the a priori ideas and information are incorporated. We draw our conclusion based on this posterior distribution.

In *example 2*, we take similar steps. We see that the Bayesian method also takes into account how extreme an observation is under the *alternative* hypothesis.

1.6 What are the (dis)advantages of the Bayesian approach in these examples 1 & 2?

In example 1:

- A different prior distribution of θ leads to a different answer. In other words, the conclusion from a Bayesian analysis is very much dependent on a priori assumptions. However, this is not different for the classical approach. We also use a priori assumptions there like if we should use a one- or two-sided test or which significance level we should use.
- In fact, one could say that allowing different a priori assumptions causes discrimination. This is the case, but the discrimination is at least very easy to spot instead of the kind of 'secret' discrimination that you could have with the frequentist approach where the discrimination is in the choice of the significance level for example.

Nog niet echt af, maar is een beetje vaag

1.7 What does the $\propto$ symbol mean? What is a kernel of the posterior density or posterior density kernel?

The $\propto$ symbol stands for "is proportional with" . That is, dividing the left side of the operator by the right side of the operator should give a constant factor that doesn't depend on any parameter θ anymore.
We define the **kernel** of the posterior probability density function $p(\theta|y)$ as $p(\theta)p(y|\theta)$.

1.8 What are informative and non-informative priors?

For this question, we take a closer look to the **prior density** $p(\theta)$. We distinguish two types of priors:

- **Informative prior**: probability density reflecting beliefs on θ before one observes the data y. Note: this can be based on other/older datasets.
- **Non-informative prior**: probability density reflecting that one has no beliefs about θ before one observes the data y. For example, uniform distribution on interval $[0, 1]$ for probability parameter θ.

1.9 Derivations in a model with a Bernoulli distribution under a uniform prior

Slide 31 - 32

2 Lecture 2

For this part of the lecture, it is important that you're also able to reproduce exercise 1 and 2 from the slides.

2.1 Model with normal distribution and known variance

Suppose we have a model with a normal distribution and known variance. Suppose that we have $n = 1$ observation $y \sim N(\mu, \sigma^2)$ with unknown μ and known σ^2, and that we specify the prior: $\mu \sim N(m_{prior}, v_{prior})$.

2.1.1 Derive the posterior density kernel

Note, on the exam it is probably asked in a slightly different way. There, you should probably show that the posterior density kernel is $p(\mu|y) \propto \exp\left(-\frac{(\mu-m_{prior})^2}{2v_{prior}} - \frac{(y-\mu)^2}{2\sigma^2}\right)$.
With the current model, we have a likelihood equal to:

$$p(y|\mu) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Using our prior, we get a prior distribution with pdf:

$$p(\mu) = (2\pi v_{prior})^{-1/2} \exp\left(-\frac{(\mu - m_{prior})^2}{2v_{prior}}\right)$$

Then, we define our *posterior density kernel* as:

$$\begin{aligned} p(\mu|y) &\propto p(\mu)p(y|\mu) \\ &= (2\pi v_{prior})^{-1/2} \exp\left(-\frac{(\mu - m_{prior})^2}{2v_{prior}}\right) \\ &\times (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\mu - m_{prior})^2}{2v_{prior}} - \frac{(y-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

This last step holds, because the multiplication of two exponential terms is one exponential term with the sum of exponents as exponent. We can use is proportional to, because the other terms $(2\pi v_{prior})^{-1/2}$ and $(2\pi\sigma^2)^{-1/2}$ are just constants.
The derivation on slides 20-22 is **not** exam material.

2.1.2 Interpret the mean and the precision of the posterior distribution

The derivation on slides 20-22 results implicitly in the following *mean of the posterior distribution*:

$$\begin{aligned} m_{posterior} &= v_{posterior}\left[\frac{m_{prior}}{v_{prior}} + \frac{y}{\sigma^2}\right] \\ &= \frac{\frac{1}{v_{prior}}}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2}} m_{prior} + \frac{\frac{1}{\sigma^2}}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2}} y \end{aligned}$$

We interpret it as follows. The posterior mean is a *weighted average* of *prior mean* and *observation* with weights proportional to the precision of the prior and the precision of the observation. If the precision of the observation is for example very high, then the term $\frac{1}{\sigma^2}$ will become relatively large and the $m_{posterior}$ will be mostly determined by y instead of m_{prior}.

Where does this $v_{posterior}$ come from? It is the variance of the posterior distribution:

$$v_{posterior} = \frac{1}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2}}$$

We interpret it however in terms of precision, where we define the precision as 1/variance, which gives the precision as:

$$\frac{1}{v_{posterior}} = \frac{1}{v_{prior}} + \frac{1}{\sigma^2}$$

The interpretation is very simple. We say that the precision of the posterior distribution is the sum of the precision of the prior and the precision of the data.

2.1.3 What does the prior $p(\mu) \propto 1$ mean?

What is the situation? We assume a finite mean μ here. This is actually what we get if we specify a **non-informative prior** and let the variance $v_{prior} \to \infty$, such that the precision $\frac{1}{v_{prior}} \to 0$. Assuming the distribution is still normal, it can be derived that the $m_{posterior} = y$ and $v_{posterior} = \sigma^2$ using the formulas for these quantities from subsection 2.1.

Interpretation and improper prior How can we **interpret** this prior distribution? One interpretation is to state that $p(\mu) = \frac{1}{2M}$ for $-M < \mu < M$ with $M \to \infty$. Then, we could use the Uniform the distribution on $[-M, M]$ or a $N(0, M)$ prior with very large M.
We say that the prior 'distribution' (it is actually not really a distribution anymore) with kernel $p(\mu) \propto 1$ is not a **proper** distribution. Therefore, we call this distribution **improper**.
We call a distribution **improper** if we can *not* simulate draws from this prior distribution. Why is this often (and also here) the case? The density does *not* integrate to 1!
Note however that in combination with the likelihood of this model, this prior *does* yield a *proper posterior distribution* from which we *can* simulate. It is of course always useful to check of the posterior is really *proper*, if we use an *improper* prior. In some case, it can be that an improper prior yields an improper posterior too.

Non-informative prior We recall the definition of a *non-informative prior* from section 1.8. The distribution is maybe proportional to a constant, but the mean μ of the distribution is still not known. Therefore, it can not be an informative prior.

2.2 Model with i.i.d. observations $y_j \sim N(\mu, \frac{1}{h})$ and *unknown* variance

We have a model with i.i.d. observations $y_j \sim N(\mu, \frac{1}{h})$, $j = 1, \ldots, n$ and *unknown* variance $\sigma^2 = \frac{1}{h}$ under non-informative prior $p(\mu, h) \propto \frac{1}{h}$.

2.2.1 Derivation of the likelihood

Note that we have

$$p(y_j|\theta) = \frac{\sqrt{h}}{\sqrt{2\pi}} \exp\left[\frac{h}{2}(y_j - \mu)^2\right]$$

We define $y = (y_1, y_2, \ldots, y_n)'$ which results in that the *likelihood* is given by:

$$\begin{aligned} p(y|\theta) &= \prod_{j=1}^{n} \frac{h^{1/2}}{(2\pi)^{1/2}} \exp\left[-\frac{h}{2}(y_j - \mu)^2\right] \\ &= \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] \end{aligned}$$

2.3 Derivation of the joint posterior density kernel of μ and h

We define $\theta = \{\mu, h\}$. We use Bayes' rule to find the kernel of the joint posterior density of θ as the product of prior and likelihood:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &= p(\mu, h)p(y|\mu, h) \\ &\propto h^{-1}\frac{h^{n/2}}{(2\pi)^{n/2}}\exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right] \\ &\propto h^{n/2-1}\exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right] \end{aligned}$$

We set $p(h)$ and $p(\mu)$ as so-called *flat priors*. That is:

$$p(\mu) \propto 1$$

for finite μ and

$$p(h) \propto \frac{1}{h}$$

Note, that $p(h)$ maybe doesn't look as a flat prior, but using $\log(h)$, you'll see that $p(\log(h)) \propto 1$ for finite $\log(h)$. The advantage of these flat priors is that the results do not depend on the scale of the data.

2.4 The conditional posterior density of h given μ and μ given h

For computing the posterior mean of μ using the joint posterior density kernel, we need Gibbs sampling. In Gibbs sampling, we simulate in an iterative way from the conditional posterior of μ given h and the conditional posterior of h given μ. Therefore, it is useful to derive these conditional distributions.
We start with the conditional posterior density of μ given h. Given $h = 1/\sigma^2$, we again have a *known* variance, so we are back in the situation of n i.i.d. observations $y_j \sim N(\mu, \sigma^2)$ with a known variance σ^2.
Using the flat prior $p(\mu) \propto 1$, this implies that the *conditional distribution of μ given h* is:

$$\mu|h, y \sim N(m_{posterior}, v_{posterior}) = N\left(\overline{y}, \frac{\sigma^2}{n}\right) = N\left(\overline{y}, \frac{1}{hn}\right)$$

Then, the *conditional posterior of h given μ* is:

$$\begin{aligned} p(h|\mu, y) &= \frac{p(\mu|h, y)}{p(\mu|y)} \\ &\propto p(\mu|h, y) \propto h^{n/2-1}h^{n/2-1}\exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right] \end{aligned}$$

One can recognize a Gamma distribution with parameters $a = n/2$ and $b = \left(\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right)^{-1}$

2.5 Steps of Gibbs sampling approach in this model

Divide a set of parameters θ into m subsets $\theta_1, \ldots, \theta_m$. Define

$$\theta_{-s} = \{\theta_1, \ldots, \theta_{s-1}, \theta_{s+1}, \ldots, \theta_m\}$$

as the $m-1$ subsets excluding the s-th subset. Then the Gibbs sampling algorithm is as follows:

for $i = 1, \ldots, n_{draws}$ **do**
  for $s = 1, \ldots, m$ **do**
    Obtain $\theta_s^{(i)}$ from conditional posterior density $p\left(\theta_s | \theta_{-s}^{(i-1)}, y\right)$, where $\theta_{-s}^{(i-1)} = \{\theta_1^{(i)}, \ldots, \theta_{s-1}^{(i)}, \theta_{s+1}^{(i-1)}, \ldots \theta_m^{(i-1)}\}$ denotes all subsets except θ_s at their *most recently simulated* values.
  end for
end for

The series of draws called a Gibbs sequence froms a Markov chain. In fact, Gibbs sampling is a **Markov Chain Monte Carlo** method.

2.6 What is a burn-in?

The distribution of Gibbs draws *converges* to the posterior distribution. Usually, we use a *burn-in*, which means that we discard the first part of the Gibbs sequence. Why? These draws may depend too much on the initial values $\theta^{(0)}$ of the Gibbs sequence.

3 Lecture 3

A part of the exam material is being able to reproduce exercise 3 and 4 on the slides of this lecture. Take a look at the slides for that.

3.1 Derivation of the likelihood and posterior density kernel in an ARCH(1) model

We define an ARCH(1) model with mean 0 and a normal distribution as

$$y_t | I_{t-1} \sim N(0, \sigma_t^2)$$

with conditional variance

$$\sigma_t^2 = \text{var}(y_t | I_{t-1}) = \alpha_0 + \alpha_1 y_{t-1}^2$$

with information set $I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$.
We also have a few restrictions on the parameters:

- $\alpha_0 > 0$. This is the constant term
- $0 \leq \alpha_1 \leq 1$. This is the effect of yesterday's squared return y_{t-1}^2 on today's return's variance $\text{var}(y_t | I_{t-1})$.

These restrictions ensure that the variance σ_t^2 is not negative.
In the ARCH(1) model, the unconditional variance is:

$$\text{var}(y_t) = \frac{\alpha_0}{1 - \alpha_1}$$

You don't need to know the derivation. Before we head to the derivation of the *likelihood* and *posterior density kernel*, we introduce the concept of variance targeting:

3.1.1 Variance targeting

Variance targeting means estimating the model so that the (estimated) unconditional variance is equal to the sample variance $s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \overline{y})^2$.
In the ARCH(1) model, this implies that we set the unconditional variance equal to the sample variance, which gives as a result:

$$\alpha_0 = s^2(1 - \alpha_1)$$

We get consequently an ARCH(1) model with

$$\begin{aligned} y_t &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \text{Var}(y_t | I_{t-1}) = s^2(1 - \alpha_1) + \alpha_1 y_{t-1}^2 \end{aligned}$$

3.1.2 Derivation of the likelihood

For the likelihood, we need the *conditional density of y_t given the information set*. We derive:

$$\begin{aligned} p(y_t|y_{t-1},\alpha_1) &= (2\pi\sigma_t^2)^{-1/2}\exp\left(-\frac{y_t^2}{2\sigma_t^2}\right) \\ &= (2\pi[\alpha_0+\alpha_1 y_{t-1}^2])^{-1/2}\exp\left(-\frac{y_t^2}{2[\alpha_0+\alpha_1 y_{t-1}^2]}\right) \\ &= (2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2])^{-1/2}\exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right) \end{aligned}$$

Using the derivation above, we derive the *likelihood conditional on 'fixed' first observation* y_1:

$$\begin{aligned} p(y_2,\ldots,y_n|\alpha_1) &= \prod_{t=2}^{n} p(y_t|y_{t-1},y_{t-2},\ldots,\alpha_1) \\ &= \prod_{t=2}^{n} p(y_t|y_{t-1},\alpha_1) \\ &= \prod_{t=2}^{n}\{(2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2])^{-1/2}\times \\ &\quad \exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right) \end{aligned}$$

3.1.3 Derivation of the posterity density kernel

The posterior $p(\alpha_1|y) \propto p(y|\alpha_1)p(\alpha_1)$. Recall that the latter is what we call the *posterior density kernel*. Then:

$$\begin{aligned} p(y|\alpha_1)p(\alpha_1) \propto \prod_{t=2}^{n}\{(2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2])^{-1/2}\times \\ \exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right) \end{aligned}$$

Recognize that the posterior density kernel is proportional to the likelihood we derived in section 3.1.2. Why is this the case here? That is because we forgot to mention our prior. We specify a non-informative uniform prior on $[0,1)$ for α_1:

$$p(\alpha_1) = \begin{cases} 1 \text{ if } 0 \leq \alpha_1 \leq 1 \\ 0 \text{ else} \end{cases}$$

Then, it makes completely sense that the posterior density kernel is proportional to the likelihood. Because we don't have a well-known posterior distribution here, we have to use a simulation method like the *Metropolis-Hastings method*.
In practice, we would work with the logarithm of the posterior kernel, because the posterior kernel is otherwise often too small or too large to be stored on a computer.

3.2 Random walk Metropolis(-Hastings) method

3.2.1 The steps

So, how does this simulation method works? Usually, we use the following steps:

for $i = 1,\ldots,n_{draws}$ **do**
Simulate candidate draw $\tilde{\theta}$ from candidate density $Q(.)$ with mean θ_{i-1}

Compute the acceptance probability as

$$\alpha = \min\left\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\right\} = \min\{\exp\left[\ln P(\tilde{\theta}) - \ln P(\theta_{i-1})\right], 1\}$$

with target density kernel $P(\theta)$
Simulate U from a uniform distribution on $[0, 1]$.
if $U \leq \alpha$ **then**
Accept: $\theta_i = \tilde{\theta}$
else
Reject: $\theta_i = \theta_{i-1}$
end if
end for

We take a closer look at the formula for the acceptance probability α. It only depends on the ratio $\frac{P(\tilde{\theta})}{P(\theta_{i-1})}$. Therefore, if $P(\tilde{\theta}) \geq P(\theta_{i-1})$, then we accept with probability 1. Else, we may reject $\tilde{\theta}$.
We evaluate α, using the log-prior and loglikelihood for numerical reasons.

3.2.2 How to evaluate whether a candidate distribution is 'good' or 'bad'?

First, we can take a look at the *trace plot*. We only observe the accepted and repeated draws. We check if the draws move through the parameter space 'fast enough'. For example, if we have 1000 draws and the first 500 draws would be between 0.5 and 1 and the last 500 draws would be between 0 and 0.5, then we would be in very bad situation. In fact, these thousand draws are then actually equivalent to two independent draws from the uniform distribution. In that case, we would need maybe thousands of observations before we can be satisfied with the trace plot.
Second, we can observe the acceptance percentage. We don't want a percentage close to 0%, but also not close to 100%, because that can be bad too. Why? It can reflect that the candidate steps are very small. If we have a very low variance for the candidate distribution, the steps are very small, and the steps for the target distribution can also be very small. Then, the ratio of both is close to 1, so we have a high acceptance percentage. In fact, if we choose our candidate steps to be very small, then we can always get an acceptance percentage close to 100%. But this does not mean that the method works well. Although there are almost no rejections, the steps are very small, and we will observe that in the traceplot. We'll see that we still move very slow through the parameter space.
Lastly, we look at the (first order) serial correlation of the sequence. Very simple, the lower the serial correlation, the better. The serial correlation is between 0 and 1 for these Markov Chains. Closer to 0 is nice. Close to 1 is bad.

4 Lecture 4

4.1 The Poisson regression model

We first recall the Poisson distribution with mean $E(y_j|\mu) = \mu$ with pdf:

$$p(y_j|\mu) = \frac{\mu^{y_j}\exp(-\mu)}{y_j!} y_j = 0, 1, 2, \ldots$$

We use the Poisson distribution in the Poisson regression model:

$$y_j \sim \text{Poisson}(\mu_j) \text{ with } \mu_j = \exp(\beta_0 + \beta_1 x_j)$$

The exponent ensures that $\mu_j > 0$

4.1.1 Derivation of the likelihood

We first take a look at the Poisson pdf with mean $\mu_j = \exp(\beta_0 + \beta_1 x_j)$:

$$\begin{aligned} p(y_j|\theta) &= \frac{\mu^{y_j} \exp(-\mu)}{y_j!} \\ &\propto \mu_j^{y_j} \exp(-\mu_j) \\ &= \exp(\beta_0 + \beta_1 x_j)^{y_j} \exp(-\exp(\beta_0 + \beta_1 x_j)) \end{aligned}$$

with $\theta = (\beta_0, \beta_1)'$, where everything is conditional on exogenous $x = (x_1, \ldots, x_n)'$. Then, the *likelihood* can be derived as:

$$\begin{aligned} p(y|\theta) = p(y_1, \ldots, y_n|\theta) &= \prod_{j=1}^{n} p(y_j|\theta) \\ &\propto \prod_{j=1}^{n} \exp(\beta_0 + \beta_1 x_j)^{y_j} \exp(-\exp(\beta_0 + \beta_1 x_j)) \end{aligned}$$

4.1.2 Derivation of the (joint) posterior density kernel of $\theta = (\beta_0, \beta_1)$

We assume that we specify a non-informative prior: $p(\theta) \propto 1$ for $\theta = (\beta_0, \beta_1)'$, with finite β_0, β_1. Then, the posterior density kernel can be derived as:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \prod_{j=1}^{n} \exp(\beta_0 + \beta_1 x_j)^{y_j} \exp(-\exp(\beta_0 + \beta_1 x_j))$$

where $p(\theta|y)$ denotes the *posterior distribution* once again, and $p(\theta)p(y|\theta)$ denotes the *posterior density kernel* itself. These two are equal in this case, because of the choice of the prior $p(\theta)$.

4.1.3 How the random walk Metropolis(-Hastings) method can be used to simulate from this posterior

We need the random walk Metropolis(-Hastings) method in this case, because the posterior in section 4.1.2 is not a well-known distribution. We use the logarithm of posterior density kernel in the method:

$$\ln P(\theta) = \sum_{j=1}^{n} \{y_j(\beta_0 + \beta_1 x_j) - \exp(\beta_0 + \beta_1 x_j)\}$$

The following algorithm can be used:

Choose feasible initial value θ_0, like $(\ln \overline{y}, 0)$ or the ML estimator.
for $i = 1, \ldots, n_{draws}$ **do**
  Simulate candidate draw $\tilde{\theta}$ from candidate density $Q(.)$ with mean θ_{i-1}.
  Compute the acceptance probability as:

$$\alpha = \min\left\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\right\} = \min\{\exp(\ln P(\tilde{\theta}) - \ln P(\theta_{i-1})), 1\}$$

  Simulate U from the $U(0,1)$ distribution.
  if $U \leq \alpha$ **then**
    Accept draw: $\theta_i = \tilde{\theta}$
  else
    Reject: $\theta_i = \theta_{i-1}$
  end if
end for

4.2 Independence chain Metropolis-Hastings method

4.2.1 The steps and the intuition of the formula of the acceptance probability

The idea of the *independence chain* is that the candidate draws are independent. The following algorithm could be used:

Choose feasible initial value θ_0.
for $i = 1, \ldots, n_{draws}$ **do**
Simulate candidate draw $\tilde{\theta}$ from (fixed) candidate density $Q(.)$
Compute acceptance probability

$$\begin{aligned}\alpha &= \min\left\{\frac{P(\tilde{\theta})/Q(\tilde{\theta})}{P(\theta_{i-1})/Q(\theta_{i-1})}, 1\right\} \\ &= \min\left\{\exp\left[\ln P(\tilde{\theta}) - \ln P(\theta_{i-1})\right]\frac{Q(\theta_{i-1}}{Q(\tilde{\theta})}, 1\right\}\end{aligned}$$

with target density kernel $P(\theta)$ and candidate density $Q(\theta)$
Simulate U from the $U(0,1)$ distribution.
if $U \leq \alpha$ **then**
Accept draw: $\theta_i = \tilde{\theta}$
else
Reject: $\theta_i = \theta_{i-1}$
end if
end for

Now the steps are known to us, we go the intuition. Note that the acceptance probability α depends on the ratio. $P(\tilde{\theta})/Q(\tilde{\theta})$ that indicates whether the posterior density kernel $P(\tilde{\theta}) = p(\tilde{\theta})p(y|\tilde{\theta})$ is low or high in point $\tilde{\theta}$, compared with candidate $Q(\tilde{\theta})$.

- If $P(\tilde{\theta})/Q(\tilde{\theta})$ relatively very high $\implies$ large probability of accepting and repeating $\tilde{\theta}$ (or values close to it) $\iff$ not enough values simulated from cand.dens. $Q(.)$.
- If $P(\tilde{\theta})/Q(\tilde{\theta})$ relatively very low $\implies$ small probability of accepting and repeating $\tilde{\theta}$ (or values close to it) $\iff$ too many values simulated from cand.dens. $Q(.)$..
- If $P(\tilde{\theta}) = 0 \implies P(\tilde{\theta})/Q(\tilde{\theta}) = 0 \implies$ No probability of accepting $\tilde{\theta} \iff$ impossible value of $\tilde{\theta}$ (lies outside range of parameter values for example).

4.2.2 How can we evaluate whether a candidate distribution is 'good' or 'bad'?

We have a perfect candidate density of the acceptance percentage is 100% in combination with no first-order serial correlation and the trace plot resembles the first two properties.
Very often, if the acceptance percentage is low, we see no 'dense' trace plot.
If we have a low acceptance percentage, but high serial correlation and a plot that takes all parameter values, but is not dense, then the candidate density is probably too wide (slide 21)
If the trace plot is dense, but only goes through a part of the parameter space (slide 22), then we have probably a too narrow candidate density. It can still be that we have an acceptance percentage close to 100% and no first-order serial correlation. In other words, the acceptance percentage does **not** give a warning signal if the candidate density is too narrow!
Then the question remains, if the acceptance percentage is no good indicator to check whether a candidate density is too narrow, how to check it in a different way? Try a candidate density with a larger variance and check whether the estimation results substantially change. If the current candidate density would already cover the parameter space enough, then the results shouldn't change so much.

4.2.3 The relationship between a and the acceptance percentage (slide 23)

We actually discussed quite much in section 4.2.2. In short, if a, the upper bound of the uniform candidate distribution on the interval $[0, a]$ is too high, i.e. the candidate distribution is too wide, then we observe a **low** acceptance percentage.

4.2.4 The relationship between a and the serial correlation $\text{corr}(\theta_i, \theta_{i-1})$ (slide 24)

Also this part was discussed already in section 4.2.2. To summarize, if a gives a too wide candidate density, we probably observe a poor (large) first-order serial correlation.

4.2.5 Why is a candidate distribution with a too small variance more harmful than a candidate distribution with a too large variance in the independence chain Metropolis-Hastings method?

This was again partly discussed in section 4.2.2 :). It boils down to the fact that the acceptance percentage can be very high and the first-order serial correlation very low (good results!), while the candidate distribution is too narrow. A too small variance is therefore more harmful, because it is more difficult to recognize it. We should use the trace plots too detect it or try a candidate density with a larger variance and check whether the estimation results substantially change.

5 Lecture 5

5.1 When considering two models M_1 and M_2, what are the posterior odds ratio, prior odds ratio and Bayes factor?

We define the *posterior odds ratio* as:

$$K_{1,2} \equiv \frac{\Pr(M_1|y)}{\Pr(M_2|y)} = \frac{\Pr(M_1)}{\Pr(M_2)} \times \frac{p(y|M_1)}{p(y|M_2)}$$

where $\frac{\Pr(M_1)}{\Pr(M_2)}$ is the *prior odds ratio* which reflects prior beliefs. Besides, $\frac{p(y|M_1)}{p(y|M_2)}$ is what we call the *Bayes factor* which reflects how likely the data are given the models. Very often, we assume models to be equally likely a prior, i.e. that the prior odds ratio is 1. You can find on slide 2 and 3 how to calculate the probabilities.

5.2 How is the posterior odds ratio related to the prior, odds ratio and Bayes factor? Interpret this formula.

Already discussed in section 5.1.

5.3 What is the formula for the marginal likelihood of a model?

The Bayes factor $B_{1,2} = \frac{p(y|M_1)}{p(y|M_2)}$ is the ratio of *marginal likelihoods*:

$$p(y|M_1) = \int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1$$
$$p(y|M_2) = \int p(y|\theta_2, M_2)p(\theta_2|M_2)d\theta_2$$

The first part of the product is the prior density. The last part is the likelihood. Note from the slide: the prior densities and likelihood functions **contain all scaling constants**. That is, priors and likelihood are typically **not** allowed to be merely kernels (proportional to the prior or likelihood)! The only exception is when we use the *same* kernel of the prior density in both models M_1 and M_2.

5.4 Explain why posterior model probabilities are very sensitive to the choice of the prior distribution of the parameters in the models and to the choice of the prior model probabilities

You can see it in the formula for the posterior odds ratio. It depends quite heavily on the prior probabilities. Very often, we assume that the prior odds ratio is equal to 1 to get easier results. However, this assumption really has an impact on the result of course. Second, it is discussed in more detail in section 5.7.2.

5.5 Why can't we use the improper prior $p(DoF) \propto 1$ for $DoF > 2$ for the degrees of freedom parameter DoF when estimating a model with a Student-t distribution?

When we would use this prior, then the posterior does not exist. The sequence of the random walk Metropolis-Hastings method will be a random walk, wandering through the tail of the large positive values of DoF. See slide 14.

5.6 How can the method of importance sampling be used to compute a marginal likelihood?

When using *importance sampling* to compute the marginal likelihood , we can rewrite the marginal likelihood given by:

$$\begin{aligned} p(y) &= \int p(\theta)p(y|\theta)d\theta \\ &= \int \frac{p(\theta)p(y|\theta)}{Q(\theta)}Q(\theta)d\theta \\ &= E_{\theta\sim Q(\theta)}\left(\frac{p(\theta)p(y|\theta)}{Q(\theta)}\right) \\ &\approx \frac{1}{n_{draws}}\sum_{i=1}^{n_{draws}}\frac{p(\theta_i)p(y|\theta_i)}{Q(\theta_i)} \end{aligned}$$

with draws θ_i and $i = 1, 2, \ldots, n_{draws}$ from the candidate density $Q(\theta)$. The idea of importance sampling is the step where we involve this candidate density which is in this case called the *importance density.*
A special case is when use an importance density equal to the prior density $\implies$ $Q(\theta) = p(\theta)$. Then, as a result:

$$p(y) \approx \sum_{i=1}^{n_{draws}} p(y|\theta_i)$$

Note that we evaluate the model quality here as the *mean* of the likelihood values instead of the *maximum* likelihood value in the frequentist approach.

5.7 Comparing a model with a Student-t distribution and a model with a generalized error distribution (GER)

We first the model with a *Student-t distribution.* We assume that we have n i.i.d. observations $y_j, (j = 1, 2, \ldots, n)$ from the Student-t(μ, σ^2, DoF) distribution with density:

$$p(x) = \frac{\Gamma(\frac{DoF+1}{2})}{\Gamma(\frac{DoF}{2})\sqrt{DoF\pi}} \times \frac{1}{\sigma} \times \left(1 + \frac{(x-\mu)^2}{DoF\sigma^2}\right)^{-\frac{DoF+1}{2}}$$

For finite x, positive σ and $DoF > 2$ with:

- mean $= \mu$
- variance $= \sigma^2 \frac{DoF}{DoF-2}$
- skewness $= 0$
- kurtosis $= \begin{cases} 3 + \frac{6}{DoF-4} & \text{if } DoF > 4 \\ \infty & \text{if } 2 < DoF \leq 4 \end{cases}$

The alternative model also has i.i.d. observations $y_j (j = 1, 2, \ldots, n)$ but they are now generated by the *Generalized Error Distribution (GED)* with density:

$$p(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left[\left(\frac{|x-\mu|}{\alpha}\right)^{\beta}\right]$$

with finite x, location parameter μ and positive scale parameters α and β with

- mean $= \mu$
- variance $= \alpha^2 \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}$
- skewness $= 0$
- kurtosis $= \frac{\Gamma(5/\beta)\Gamma(1/\beta)}{(\Gamma(3/\beta))^2}$

Note: $\beta = 2$ implies a $N\left(\mu, \frac{\alpha^2}{2}\right)$ distribution. A smaller beta implies a fatter tail and a larger beta implies a thinner tail.

5.7.1 Why would we be interested in computing the Bayes factor?

If we want to compute the *posterior model probabilities for these two models*, then we have to calculate the posterior odds ratio which involves calculating the Bayes factor. We could be interested in this, because we want to choose one of the two models or we maybe want to combine them.

5.7.2 Explain why the posterior model probabilities of the two models are very sensitive to the choice of the prior distribution of DoF and β in these models

As was said already in section 5.4, the sensitivity of the posterior model probabilities to the choice of the prior distribution will be discussed here in more detail. Suppose we draw DoF values from a uniform prior distribution for DoF on a certain interval, then the choice of the interval has a large impact on the marginal likelihood that follows form it. If we choose the interval to be like [4.1,50], then changing that interval to [4.1,1000] for example makes the marginal likelihood much smalle, because then there would be more mass in the prior distribution for higher values of DoF which corresponds to thin-tailed distributions and these do not match very well with the simulated dataset that was used in the slides. So, in general, the choice of your interval for an uniform prior can have a large impact on the results, because we put more mass on 'stupid' values of the DoF parameter if we choose the interval too large.
Therefore, if we want to compare two models like a model with a Student-t density and a model with a GED density, we should not choose a very narrow interval for the prior of the Student-t model for example, while choosing a very wide interval for the prior of the GED model. In that case, we would always choose for the model with the narrow prior distribution.

5.7.3 How can we use the kurtosis to compute the posterior model probabilities in a 'fair' way?

We use the kurtosis to compute the posterior model probabilities in a 'fair' way. How? If the interval/range of the kurtosis is the same for both models (so for Student-t and GED), then the prior distributions are equally 'precise' or equally 'stupid' if we choose very wide / fat-tailed prior distributions for both models.
On slide 27, we see two graphs. The top graph indicates the kurtosis of the Student-t distribution model given values of DoF in the range [4.1,50]. We see that the maximum kurtosis is 63 and the minimum kurtosis is appoximately 3.13. The graph at the bottom indicates the kurtosis of the GED model given values of β in the range [0.38,1.88]. The maximum and minimum kurtosis are again approximately 63 and 3.13 respecitvely. The shape of the graph is different of course, because the variable on the x-axis is different.

6 Lecture 6

6.1 Model with normal distribution (unknown variance) under non-informative prior $p(\mu, h) \propto \frac{1}{h}$ (for $h > 0$,0 else)

We consider the simple model with i.i.d. *normally* distributed observations:

$$y_j \sim N\left(\mu, \frac{1}{h}\right) \qquad\qquad j = 1, \ldots, n$$

Define $y = (y_1, y_2, \ldots, y_n)'$. The likelihood is given by

$$\begin{aligned} p(y|\theta) &= \prod_{j=1}^{n} \frac{h^{1/2}}{(2\pi)^{1/2}} \exp\left[-\frac{h}{2}(y_j - \mu)^2\right] \\ &= \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] \end{aligned}$$

We choose a typical *non-informative prior* in this model:

$$p(\theta) = p(\mu, h) \propto \frac{1}{h}$$

for finite μ and positive.
A kernel of the joint posterior density of μ and h is:

$$p(\mu, h|y) \propto h^{n/2-1} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right]$$

This kernel follows from multiplication the prior with the likelihood and omitting the term with π, because that is a constant.
Then, we have a marginal posterior density of μ:

$$p(\mu|y) = \int p(\mu, h|y)dh \propto \int h^{n/2-1} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] dh$$

6.1.1 Derivation of the kernel of the marginal posterior density $p(\mu|y)$

Now, we have all the information needed to start deriving some interesting results. We start with the derivation of the *kernel of the marginal posterior density* $p(\mu|y)$. Therefore, we have to evaluate the integral in the expression of the marginal posterior density of section 6.1. We recognize the conditional posterior density $p(h|mu, y)$ and the integral of that density should therefore be equal to 1. Moreover, recognize that the conditional posterior density of h given μ is the Gamma density:

$$p(h|\mu, y) = \frac{1}{\Gamma(a)b^a} h^{a-1} \exp\frac{-h}{b}$$

with $a = n/2$ and $b = \frac{1}{\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu)^2}$ gives:

$$\frac{\left[\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right]^{n/2}}{\Gamma(n/2)} h^{n/2-1} \exp\left(-\left[\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] h\right) dh = 1$$

It follows that:

$$\int h^{n/2-1} \exp\left(-\left[\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] h\right) dh = \frac{\Gamma(n/2)}{\left[\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right]^{n/2}}$$

Then, the marginal posterior $p(\mu|y)$ is:

$$\begin{aligned} p(\mu|y) &\propto \int h^{n/2-1} \exp\left(- \left[\frac{1}{2} \sum_{j=1}^{n} (y_j - \mu)^2 \right] h \right) dh \\ &= \frac{\Gamma(n/2)}{\left[\frac{1}{2} \sum_{j=1}^{n} (y_j - \mu)^2 \right]^{n/2}} \\ &\propto \left[\sum_{j=1}^{n} (y_j - \mu)^2 \right]^{-n/2} \end{aligned}$$

6.1.2 The marginal posterior distribution of μ

The derivation on slide 7 (which is **not** exam material) gives:

$$p(\mu|y) \propto \left[1 + \frac{1}{n-1} \frac{(\mu - \overline{y})}{s^2/n} \right]^{-(n-1+1)/2} \tag{1}$$

In equation (1), whcih is the marginal posterior density of μ, we recognize the *Student-t distribution*:

$$p(x) = \frac{\Gamma(\frac{DoF+1}{2})}{\Gamma(\frac{DoF}{2})\sqrt{DoF\pi}} \frac{1}{c} \times \left[1 + \frac{1}{DoF} \frac{(x-m)^2}{c^2} \right]^{-(DoF+1)/2}$$

with $x = \mu$ and:

- $m = \overline{y}$, the location parameter
- $c^2 = s^2/n$, the variance parameter
- $DoF = n - 1$, the degrees of freedom parameter

6.1.3 Symmetry and differences between the Bayesian and frequentist approaches (slide 10)

Note, in *Bayesian analysis*, we have:

$$\frac{\mu - \overline{y}}{\sqrt{s^2/n}} | y \sim t(0, 1, n-1)$$

while in *frequentist/classical inference* we have:

$$\frac{\overline{y} - \mu}{\sqrt{s^2/n}} | y \sim t(0, 1, n-1)$$

Do you see the symmetry? Only the numerator is swapped around for the classical approach in comparison to the Bayesian approach.
However, the interpretation is different:

- **Bayesian:** we consider μ as a random variable, while $\overline{y}, s^2$ are functions of the data. The last two are given after being observed.
- **Frequentist:** $\overline{y}$ and s^2 are still functions of the data, but are now considered to be random. In contrast, μ is a fixed, unknown parameter.

6.2 Normal linear regression model under non-informative prior $p(\beta, h) \propto \frac{1}{h}$ (for $h > 0$,0 else)

We consider the linear regression model with normally distributed errors under the standard assumptions of homoskedasticity and independent normally distributed errors:

$$\begin{aligned} y_j &= x_j'\beta + \varepsilon_j, \\ \varepsilon_j &\sim N\left(0, \frac{1}{h}\right) \\ j &= 1, \ldots, n \end{aligned}$$

with *parameter set* $\theta = \{\beta, h\}$, β the *vector of coefficients*, $h = 1/\text{var}(\varepsilon_j)$ the precision of ε_j, y_j (scalar) the endogenous dependent variable and x_j a vector of exogenous variables.
Here, we have:

$$p(y_j|\theta) = \frac{\sqrt{h}}{\sqrt{2\pi}} \exp\left[-\frac{h}{2}(y_j - x_j'\beta)^2\right]$$

Note, y_j also depends on x_j and not only on θ, but we ignore that in the notation for now.
Define $y = (y_1, y_2, \ldots, y_n)'$ and $X = (x_1, x_2, \ldots, x_n)'$. Then, the *likelihood* is given by:

$$\begin{aligned} p(y|\theta) &= \prod_{j=1}^{n} \frac{\sqrt{h}}{\sqrt{2\pi}} \exp\left[-\frac{h}{2}(y_j - x_j'\beta)^2\right] \\ &= \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - x_j'\beta)^2\right] \\ &= \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{h}{2}(y - X\beta)'(y - X\beta)\right] \end{aligned}$$

Besides, as stated in the title of this section, we choose as non-informative prior:

$$p(\theta) = p(\beta, h) \propto \frac{1}{h}$$

for all $\beta \in \mathbb{R}^k$ and $h > 0$.
Note that we have a *joint prior* here which is in fact the product of two independent prior densities:

$$p(\theta) = p(\beta, h) = \left(\prod_{i=1}^{k} p(\beta_i)\right) \times p(h)$$

with *flat* prior on β_i for all $i = 1, \ldots, k$:

$$p(\beta_i) \propto 1$$

for finite β_i and

$$p(h) \propto \frac{1}{h}$$

for $h > 0$.

6.2.1 Marginal posterior distribution of regression coefficient β_i (slide 21)

The derivations on slides 17-19 (which are **not** part of the exam) give the following *marginal posterior density of* β:

$$p(\beta|y) \propto \left[1 + \frac{1}{n-k}(\beta - \hat{\beta})'\left[\frac{1}{s^2}X'X\right](\beta - \hat{\beta})\right]^{-(n-k+k)/2}$$

In the latter, we recognize the density of a *k-dimensional Student-t distribution*:

$$p(x) = \frac{\Gamma(\frac{DoF+k}{2})}{\Gamma(\frac{DoF}{2})DoF^{k/2}\pi^{k/2}}|\Sigma|^{-1/2}\left[1 + \frac{1}{DoF}(x-m)'\Sigma^{-1}(x-m)\right]^{-(DoF+k)/2}$$

with $x = \beta$ and

- $m = \hat{\beta}$, the *location vector parameter*.
- $\Sigma = s^2(X'X)^{-1}$, the *variance-covariance matrix parameters*.
- $DoF = n - k$, the *degrees of freedom parameter*.

6.2.2 Symmetry and differences between the Bayesian and frequentist approaches (slide 21)

This is actually quite similar to what we did in the previous model. In *Bayesian analysis*, we have for element β_i:

$$\frac{\beta_i - \hat{\beta}_i}{\sqrt{[s^2(X'X)^{-1}]_{ii}}}|y \sim t(0, 1, n-k)$$

while for the *frequentist/classical inference* we have:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{[s^2(X'X)^{-1}]_{ii}}}|y \sim t(0, 1, n-k)$$

The only difference is again the numerator where the terms are swapped around. Again, we see some symmetry in these expression, but the *interpretation is different*:

- **Bayesian:** the β_i's are random, while the $\hat{\beta}_i$'s and the s^2 are given (after being observed). They are functions of the data.
- **Frequentist:** the $\hat{\beta}_i$'s and s^2 are still functions of the data, but are now random. The β parameter is fixed and unknown.

6.3 The Savage-Dickey density ratio

We are now concerned with a model with a normal distribution with unknown mean and variance and we would like to compute $\Pr(\mu = 0|y)$ and $\Pr(\mu \neq 0|y)$. We introduce two models:

- M_1: a model with i.i.d. $y_j \sim N\left(\mu, \frac{1}{h}\right)$ and $\mu = 0$ with $\theta_1 = \{h\}$

M_2: the same model as above, but with a 'free' $\mu \implies 100\%$ prob. of $\mu \neq 0$, and $\theta_2 = \{\mu, h\}$ It could be that we want to calculate the *Bayes factor*. Remember this topic from lecture 5. We evaluate the *marginal likelihoods* in the Bayes factor using *Importance Sampling*. However, **if** M_1 is a restricted version of M_2, we can use something else: the *Savage-Dickey density ratio (SDDR)*.
In this case, M_1 imposes the restriction $\mu = 0$ on M_2 with the same prior for other parameters. Then, the Bayes factor is:

$$B_{1,2} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{p(y|\mu = 0, M_2)}{p(y|M_2)}$$

From Bayes' rule in the unrestricted model M_2, we have:

$$p(\mu = 0|y, M_2) = \frac{p(\mu = 0|M_2)p(y|\mu = 0, M_2)}{p(y|M_2)}$$

Then, dividing both sides by the prior density at $\mu = 0$, i.e. $p(\mu = 0|M_2)$, yields:

$$\frac{p(\mu = 0|y, M_2)}{p(\mu = 0|M_2)} = \frac{p(y|\mu = 0, M_2)}{p(y|M_2)}$$

Note that the right side of the last equation is exactly the same as the Bayes factor. The left side of the last equation is the SDDR.
How can we **interpret** it? A rough interpretation could be:

- $B_{1,2} > 1$: data y **support** the restricted model with $\mu = 0$ if the posterior $p(\mu = 0|y, M_2) >$ prior $p(\mu = 0|M_2)$.
- Similarly, if $B_{1,2} < 1$: the data y **don't support** the restricted model with $\mu = 0$ if the posterior $p(\mu = 0|y, M_2) <$ prior $p(\mu = 0|M_2)$.

The idea is that if μ is truly equal to 0, then the posterior of μ should get more and more concentrated close around $\mu = 0$ if more data are observed. Similarly, if $\mu = a, a \neq 0$, then the posterior should get more and more concentrated around this non-zero a if more data are observed.

6.3.1 When it can be used?

In short, it can be used if one model is a restricted version of the other.

6.3.2 Why can it be useful?

It can tell us something about the restriction itself. We can sort of test whether the restriction is true or not, but I have to be very careful with the word 'test', because we are working with a Bayesian approach. It looks actually quite similar to doing a Likelihood Ratio test in the frequentist approach. The advantage here is that we can also something about the sign of the restricted parameter if it is found to be non-zero.

6.4 Two models M_1, M_2, where M_1 is a restricted version of M_2 with restriction $\mu = 0$

The models were already introduced in section 6.3.

6.4.1 The similarity between the Bayes factor and the Savage-Dickey density ratio

6.4.2 Computation of the Bayes factor, posterior odds and posterior model probabilities using the Savage-Dickey density ratio

6.4.3 Why the posterior model probabilities are very sensitive to the choice of the prior density of μ in model M_2

He doesn't specify this in the lecture. I think it has a similar reason in comparison to what was told in lecture 5.
An interesting question could be whether it is a problem that the posterior model probabilities are very sensitive to the choice of the prior. In fact, this is not necessarily the case. Think of it; a classical/frequentist statistician also makes prior decisions like choosing to test $\mu = 0$ against $mu \neq 0$ or $mu < 0$ or $mu > 0$ for example. For a Bayesian statistician, it matters what you mean by $\mu \neq 0$. Are we talking about $\mu \in [1, 2]$ or $\mu \in [1000, 2000]$ for example? It could even be that you prefer $\mu \in [1, 2]$ over $\mu = 0$, but you also prefer $\mu = 0$ over $\mu \in [1000, 2000]$.
The question whether to choose $\mu = 0$ is $\mu \neq 0$ is actually a little bit weird. Think of the example that Lennart give in the lecture:
Suppose someone asks you: "Do you like to live in a small appartment in Amsterdam Buitenveldert or somewhere else?" A frequentist may answer: "I reject the optimality of a small appartment in Amsterdam Buitenveldert." A Bayesian may answer: "What do you mean by somewhere else? A large appartment in Amsterdam Buitenveldert or a villla in Amsterdam South? Or do you mean being homeless in Zimbabwe or working in a labor camp in North Korea?"

6.5 The Jeffreys-Lindley-Bartlett paradox

The *Jeffrey-Lindley-Bartlettt paradox* can be explained as follows. Suppose data stem from model M_2, so the true $\mu = 1$ and a frequentist test and a Bayesian 95% posterior interval clearly **correctly** reject the false value $\mu = 0$. Then, a Bayesian using posterior model probabilities in combination with a non-informative prior (with large enough prior variance) will choose **false** model M_1, so with $\mu = 0$. The first thing that you think of that this is a mistake. In fact, it isn't! It is just a paradox and not a mistake, because of the Bayesian framework. If M_2 puts almost all of its probability mass at 'silly' values, then $\mu = 0$ is definitely a better choice than $\mu \neq 0$.

6.6 Example of a situation where the Jeffreys-Lindley-Bartlett paradox is observed (slide 34)

Suppose we have i.i.d. $y_j \sim N\left(\mu, \frac{1}{h}\right)$ with prior $p(h) \propto \frac{1}{h}$. The **true value** of μ is equal to 1. We have $n = 16$ observations, a sample mean $\overline{y} = 1$ and $s^2 = 1$.
We start with performing a test according to the *frequentist approach* with $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. The t-value $= 1/\sqrt{1/16} = 4$ with a p-value equal to 0.0012. We clearly reject H_0 at a 5% significance level.
For the *Bayesian approach*, we calculate a 95% posterior interval. The result is $[0.43, 1.57]$ under flat prior $p(\mu) \propto 1$. A different choice could be a non-informative prior where $\mu \sim N(0, v_{prior})$, with large v_{prior}. Clearly, the interval doesn't contain $\mu = 0$, so we reject $\mu = 0$. So far, so good.
However, if the Bayesian would use posterior model probabilities and a non-informative prior with large enough variance, then we have $\Pr(\mu = 0|y) \approx 1$, so we would accept $\mu = 0$.

6.7 Last example:

The last example concerns forecasting real US GDP growth using 95% prediction intervals and AR models. Important for the exam:

- Steps that are used for direct simulation from the posterior predictive density of y_{T+1} in case of an $AR(p)$ model and how you can compute a Bayesian 95% prediction interval in this case.
- How to perform Bayesian Model Averaging (BMA) for forecast combination if we consider the AR(1) and the AR(2) model and how to compute a Bayesian 95% prediction interval in this case.

The remaining parts of the example don't discuss know material, so just take a look at slides 35-49 of lecture 6 for this last example.