STAT 206B
Chapter 4: Bayesian Point Estimation
Chapter 5: Hypothesis Testing & Confidence Regions

Winter 2022

† Bayesian Inference

- The posterior distribution supposedly contains all the available information about θ.
- The *entire* posterior distribution $\pi(\theta \mid x)$ is the extensive summary of the information available on the parameter θ.
- A visual inspection of the graph of the posterior will often provide the best insight concerning θ (at least in low dimensions)
- More standard uses of the posterior are still helpful e.g. point estimation, interval estimation, testing, prediction...
- CR Chapter 4 and JB Chapter 4.3

† **Bayesian Point Estimation:** the simplest inferential use of the posterior distribution

- Report a point estimate for $h(\theta)$, with an associated measure of accuracy

  $\Rightarrow$ Find $\pi(h(\theta) \mid x)$ and then the *Bayes rule* d, i.e., a solution of

  $$\min \mathrm{E}^{\pi} \{L(\theta, d) \mid x\} \quad \text{for } d \in \mathcal{D} \text{ and } \theta \in \Theta.$$

  $\star\star$ Recall we found the Bayes actions under standard loss functions such as the quadratic loss, the absolute error loss and the 0-1 loss.

  $\star\star$ The mean and median of the posterior are frequently better estimates of θ than the mode (i.e., MAP).

† Estimation Error

- We evaluate the precision of $\delta^\pi(x)$
- For example, we may use the posterior squared error:

$$\mathrm{E}^\pi[(\delta^\pi(x) - h(\theta))^2 \mid x].$$

⋆⋆ If we use $\mathrm{E}^\pi[h(\theta) \mid x]$ as the estimate of $h(\theta)$, report $\sqrt{\mathrm{Var}^\pi(h(\theta) \mid x)}$ as the standard error (posterior standard deviation).

- **JB Example 1** (p136) Consider the situation wherein a child is given an intelligence test. Assume that the tet result X is $\mathsf{N}(\theta, 100)$, where θ is the true IQ (intelligence) level of the child, as measured by the test. Assume also that, in the population as a whole, θ is distributed according to a $\mathsf{N}(100, 225)$ distribution. Suppose that we observe a student who scores 115 on the test.

  $\star\star$ We can find

  $$\theta \mid x \sim \mathsf{N}((1/100+1/225)^{-1}(x/100+100/225), (1/100+1/225)^{-1}).$$

  $$\Rightarrow \mu^{\pi}(115) = 110.39 \text{ and } \sqrt{V^{\pi}(115)} = \sqrt{69.23} = 8.32.$$

- **JB Example 8**(p137) Assume $X \sim \mathrm{N}(\theta, \sigma^2)$ (σ^2 known) and the noninformative prior $\pi(\theta) = 1$ is used, then the posterior distribution of θ given x is $\mathrm{N}(x, \sigma^2)$. Hence the posterior mean is $\mu^{\pi}(x) = x$ and the posterior variance and standard deviation are σ^2 and σ, respectively.

  $\star\star$ The same as the usual classical estimate with standard error.

  $\star\star$ Their interpretations are different!

- Sampling Properties

  ⋆⋆ Sampling properties: behavior of an estimator under hypothetically repeatable surveys or experiments.

  ⋆⋆ Suppose $\theta_0 =$ the true value of the population mean.

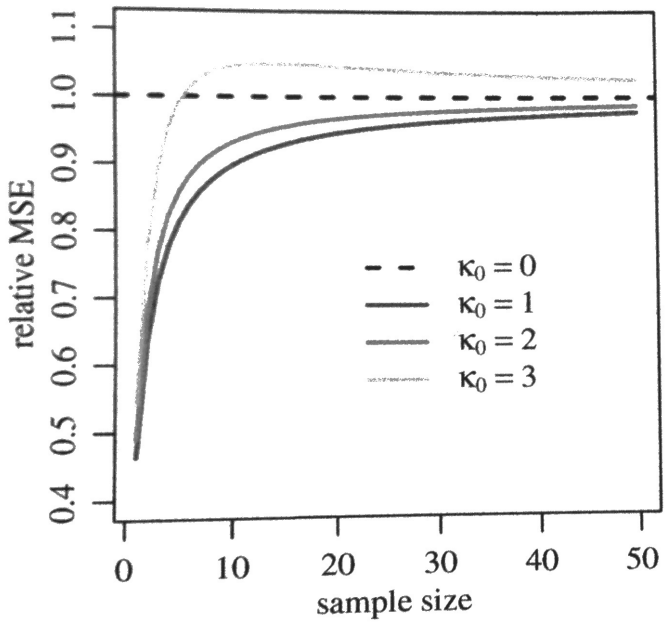
  ⋆⋆ To evaluate how close an estimator $\delta(x)$ is likely to be to θ_0, we use the mean square error(MSE)

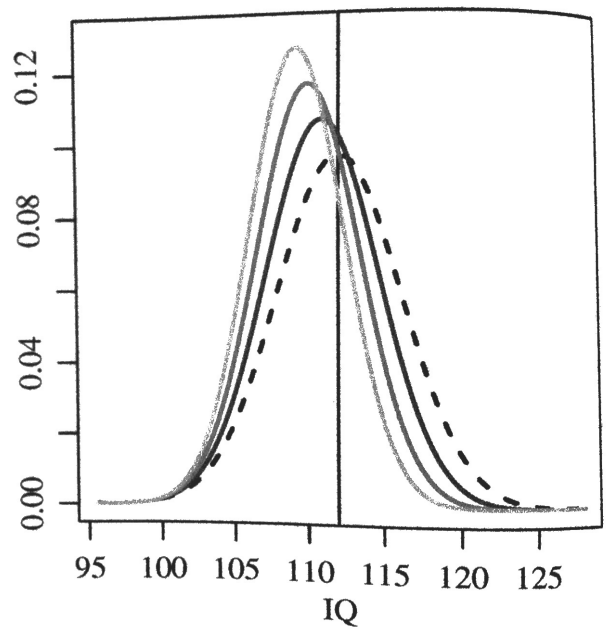
$$\begin{aligned} \mathrm{MSE}(\delta \mid \theta_0) &= \mathrm{E}\{(\delta - \theta_0)^2 \mid \theta_0\} \\ &= \mathrm{E}\{(\delta - m)^2 \mid \theta_0\} + \mathrm{E}\{(m - \theta_0)^2 \mid \theta_0\} \\ &= \mathrm{Var}(\delta \mid \theta_0) + \mathrm{Bias}^2(\delta \mid \theta_0), \end{aligned}$$

  where $m = \mathrm{E}(\delta \mid \theta_0)$

- **PH p82** Recall the IQ example (similar but different!).
  - ⋆⋆ $X \sim \mathrm{N}(100, 225)$ for the general population.
  - ⋆⋆ Suppose that we sample n individuals from a particular town and estimate θ, the town-specific mean IQ score based on the sample of size n.
  - ⋆⋆ In fact, people in the town are extremely exceptional so $\theta_0 = 112$ and $\sigma^2 = 169$.
  - ⋆⋆ Consider $x_i \mid \theta \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$, where $\sigma^2 = 169$ but θ is unknown
  - ⋆⋆ Assume $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$, where $\tau_0 = \sigma/\sqrt{\kappa_0}$
  - ⋆⋆ For Bayesian inference, we lack the information about the town a natural choice of $\mu_0 = 100$.

- **PH p82** Example: IQ Scores.
  - ⋆⋆ Let $\kappa_0 = \sigma^2/\tau^2$ and compare $\text{MSE}(\delta_n^\pi \mid \theta_0)$ and $\text{MSE}(\delta_n \mid \theta_0)$ by varying n and κ_0.
  - ⋆⋆ MSE errors and sampling distribution of different $\delta_n^\pi(x)$





- Comments on unbiasedness

  ⋆⋆ No Bayes estimate with respect to the squared error loss can be unbiased, except in a case when its Bayes' risk is 0 (that is, the perfect estimation is possible).

  $\Leftrightarrow$ If $\delta^{\pi}(x)$ is unbiased for θ, then $\delta^{\pi}(x)$ is not Bayes under the squared error loss unless its Bayes risk is zero.

  For your practice, show this.

  ⋆⋆ **No problem!** Even frequentist agree that insisting on unbiasedness can lead to bad estimators, and that in their quest to minimize the risk by trading off between variance and bias-squared a small dosage of bias can help.

† Interval Estimation (CR 5.5 and JB 4.3.2)

- $(1-\alpha)100\%$ confidence intervals (CI's)–Classical interval estimate

  ⋆⋆ Generate data from the assumed model many times and for each data set to exhibit the CI.

  ⋆⋆ Now, the proportion of CIs covering the unknown parameter "tends to" $1-\alpha$.

- We will construct $C_x \subset \Theta$ where θ should be with high probability.

  ⋆⋆ The distribution used to assess the credibility of an interval estimator is the posterior distribution.

† Credible Sets

- Credible Set: Assume the set C_x is a subset of Θ. Then C_x is a credible set with credibility $(1-\alpha)\cdot 100\%$ if

$$P^{\pi}(\theta \in C_x \mid x) = \mathsf{E}^{\pi}\{1(\theta \in C_x) \mid x\} = \int_{C_x} \pi(\theta \mid x) d\theta > 1-\alpha.$$

  ⋆⋆ If the posterior is discrete, then the integral becomes sum.

- Bayesian interpretation of a credible set C_x is natural: The probability of a parameter belonging to the set C_x is $1-\alpha$.

  ⋆⋆ The frequentist CI is random but our credible interval is fixed given data.

† Credible Sets (contd)

- For a given posterior function such set is not unique.

  ⋆⋆ *Q: How to choose one particular set?*

- For a given credibility level $(1-\alpha)100\%$, the shortest credible set is of interest.

- The size of a set is simply its total length if the parameter space Θ is one dimensional, total area, if Θ is two dimensional, and so on.

- To minimize the size, sets should correspond to highest posterior probability (density) areas.

† Credible Sets (contd)

- The $(1-\alpha)100\%$ HPD (high posterior density) credible set for parameter θ is a set C_x, subset of Θ of the form

$$C_x = \{\theta \in \Theta \mid \pi(\theta \mid x) \geq k_\alpha\},$$

where k_α is the largest constant for which

$$P^\pi(\theta \in C_x \mid x) \geq 1 - \alpha.$$

- Geometrically, if the posterior density is cut by a horizontal line at the height k_α, the set C is projection on the θ axis of the part of line inside the density, i.e., the part that lies below the density.

- See **Def 5.5.2**

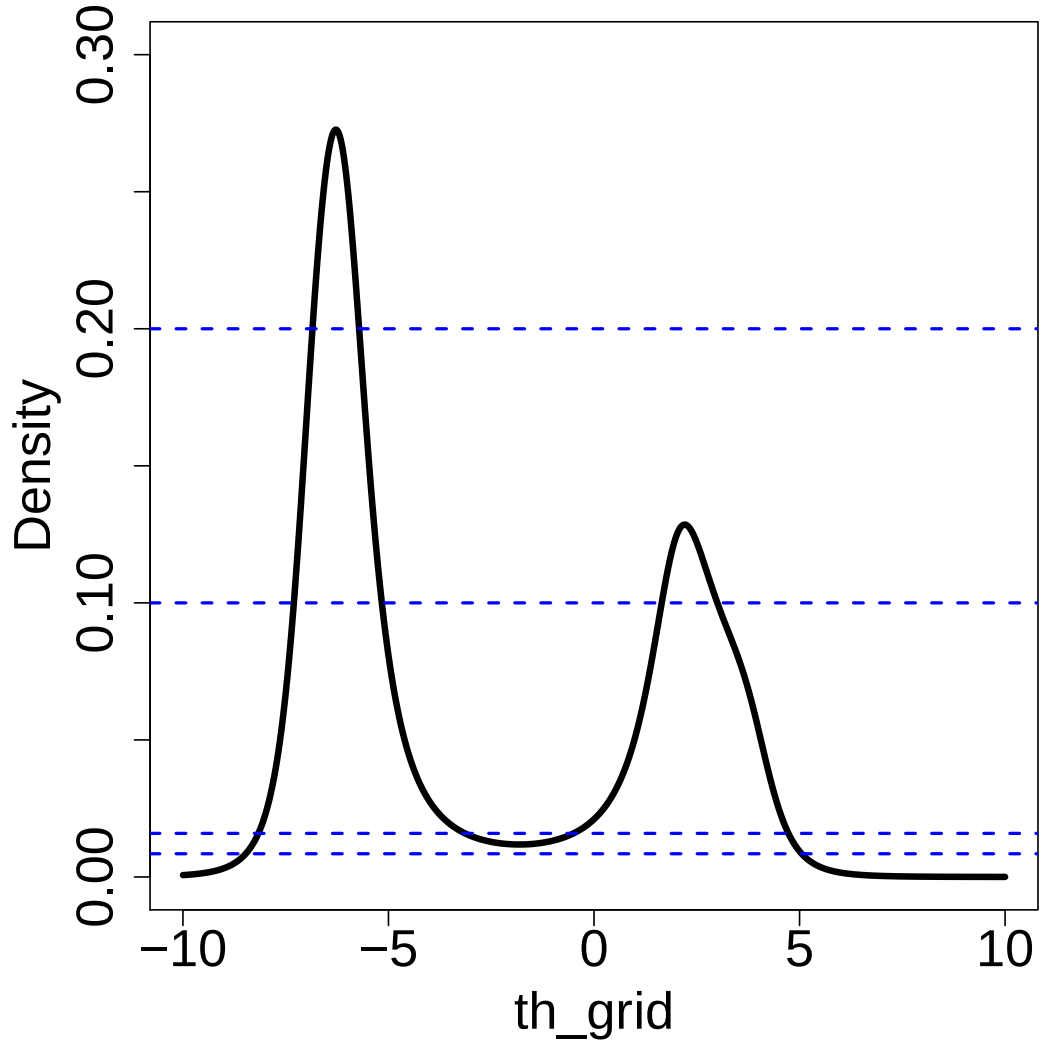
- **Example 5.5.3** Consider $x \sim \mathrm{N}(\theta, \sigma^2)$. Consider $\theta \sim \mathrm{N}(0, \tau^2)$. Find the $100(1-\alpha)\%$ HPD credible interval.

$\star\star$ Find the $100(1-\alpha)\%$ HPD credible interval with $\pi(\theta) \propto 1$.

$\star\star$ Note that we can use improper priors in this setting and do not encounter the same difficulties as when testing the point-null hypothesis.

- **JB Example 10** (p141, with a slight change) Assume that four observations, $x_i = 2, -7, 4, -6$, $i = 1, \ldots, 4$ are sampled from Cauchy $C(\theta, 1)$ distribution with parameter of interest θ ($f(x \mid \theta) = 1/\{\pi(1 + (x - \theta)^2)\}$. Consider the flat prior $\pi(\theta) = 1$. Sketch the posterior.

- **Example** (contd) The posterior is bimodal!



• **Example** (contd) Four horizontal lines at levels $k = 0.008475$, 0.0159, 0.1, and 0.2 are shown. These lines determine four credible sets,

- ⋆⋆ $k_{0.01} = 0.008475$: $[-8.498, 5.077]$ with $P^{\theta|X}(8.498 \leq \theta \leq 5.077) = 99\%$;
- ⋆⋆ $k_{0.05} = 0.0159$: $[-8.189, -3.022] \cup [-0.615, 4.755]$ with posterior credibility of 95%;
- ⋆⋆ $k = 0.1$: $[-7.328, -5.124] \cup [1.591, 3.120]$ with posterior credibility of 64.2%;
- ⋆⋆ $k = 0.2$: $[-6.893, -5.667]$ with posterior credibility of 31.3%.

- **Example** (contd)
  - ⋆⋆ Observe for $\alpha = 0.05$ and 0.1, the credible intervals consist of two separate intervals.
  - ⋆⋆ This may indicate that the prior is not agreeing with the data (unimodal in the prior vs bimodal in data).
  - ⋆⋆ There is no frequentist counterpart for the CI for θ in the above model.

- **Example** Let $x \mid \theta$ be the shifted exponential with density

$$f(x \mid \theta) = \exp\{-(x-\theta)\}1(\theta \leq x).$$

Let θ be half-Cauchy,

$$\pi(\theta) = \frac{2}{\pi(1+\theta^2)}, \quad \theta > 0.$$

Find the posterior and show that $(1-\alpha)100\%$ HPD credible set is of the form $[\beta, x]$ for some $\beta \in (0, x)$.

- **Example** Let $\eta = e^{\theta}$ and find the posterior $\pi^{\star}(\eta \mid x)$. Show that $\pi^{\star}(\eta \mid x)$ is decreasing in η and that the credible set for η is of the form $[1, \gamma]$, for some $\gamma < e^{x}$.

$\star\star$ Transform the interval of η back to the space of θ and observe $[\log 1, \log \gamma] = [0, \beta'] \neq [\beta, x]$.

- One undesirable property of credible sets is the lack of invariance with respect to monotone transformations.
- For a solution, read JB pages 144-145.
- A HPD credible sets can be found for multivariate cases. See JB p143

† Predictive Inference

- Predict a random variable $y \sim g(y \mid \theta)$ based on observations of $x \sim f(x \mid \theta)$.

  ⋆⋆ no need to be $g = f$

  ⋆⋆ easily can be extended to the case of $y \sim g(y \mid \theta, x)$.

- Find the predictive density of y given x, when the prior for θ is π,

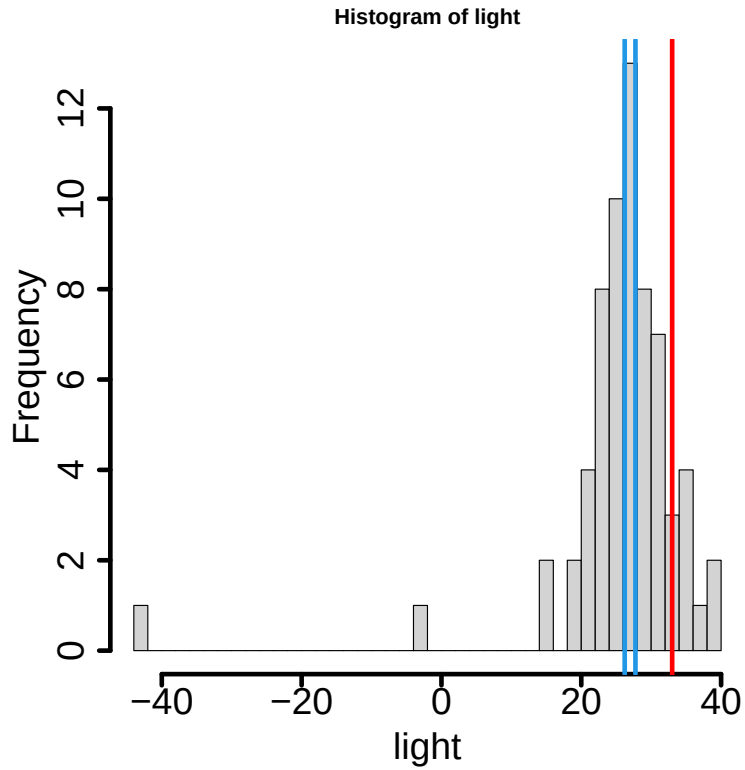
$$p(y \mid x) = \int_{\Theta} g(y \mid \theta)\pi(\theta \mid x)d\theta.$$

† Predictive Inference (contd)

- Point estimation: use the loss function and find the Bayes actions minimizing $\mathrm{E}(L(y, a) \mid x)$.
- Posterior predictive interval for y.

† Example: Estimating the speed of light (BDA p 66)

- Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters. He made 66 measurements. Consider the problem of estimating the speed of light.



† Example: Estimating the speed of light (contd)

- Consider the normal model and assume that all 66 measurements are independent draws from $\mathsf{N}(\theta, \sigma^2)$.

  $\Leftrightarrow$ Assume $x_i \mid \theta, \sigma^2 \overset{iid}{\sim} \mathsf{N}(\theta, \sigma^2)$, $i = 1, \ldots, n$ with $n = 66$

  $\Rightarrow$ inferential goal: posterior inference for θ (so σ^2 is a nuisance parameter)

- Build a prior model for unknown random model parameters θ and σ^2.

  $\Leftrightarrow$ Consider a semi-conjugate prior distribution and let

$$\theta \sim \mathsf{N}(\mu, \tau^2) \text{ and } \sigma^2 \sim \mathsf{IG}(a_0, b_0),$$

  where μ, τ^2, a_0 and b_0 are fixed.

† Example: Estimating the speed of light (contd)

- Use prior information and specify the values of the fixed hyper-parameter values.

```
> ## \theta ~ N(\mu, \tau2)
> hyper$mu <- 33
> hyper$tau2 <- 100
>
> ## \sig2 ~ IG(a0, b0)
> hyper$a0 <- 0.1
> hyper$b0 <- 0.1
```

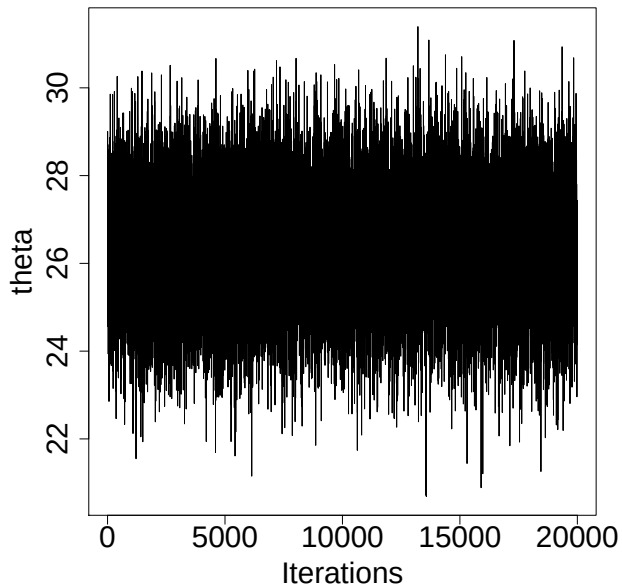
- Find the joint posterior distribution of all random parameters θ and σ^2.

- Find the posterior computation strategy.

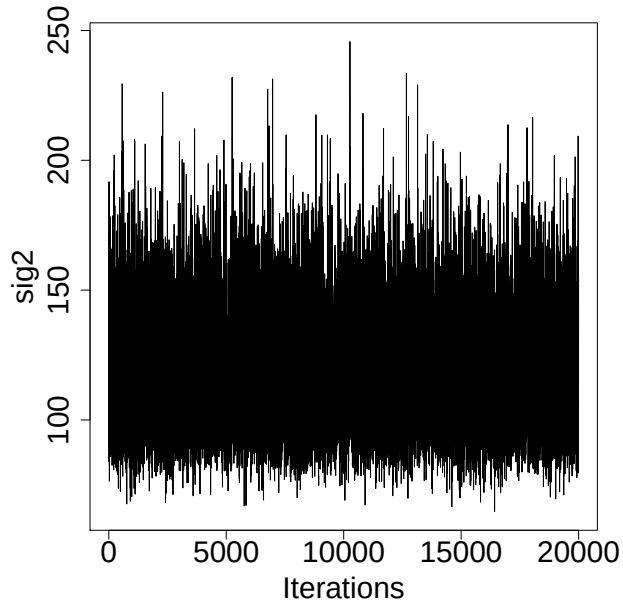
  ⋆⋆ Use the Gibbs sampler and derive the full conditional distributions.

† Example: Estimating the speed of light (contd)

- Check mixing and convergence of the Markov chain.



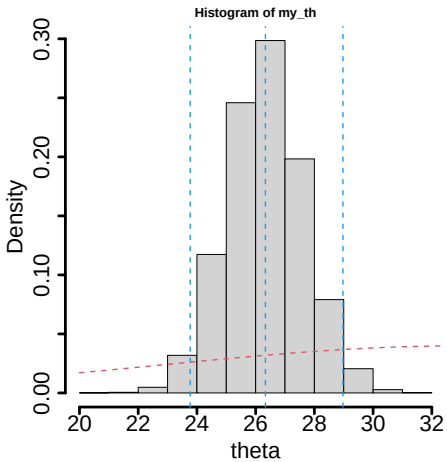
(a) θ



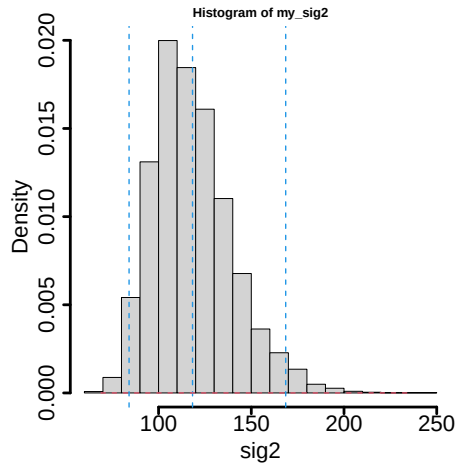
(b) σ^2

† Example: Estimating the speed of light (contd)

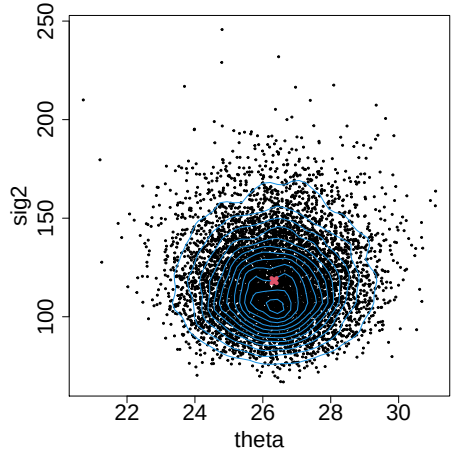
- Posterior summary of θ and σ^2



(a) θ



(b) σ^2



(c) Joint

† Example: Estimating the speed of light (contd)

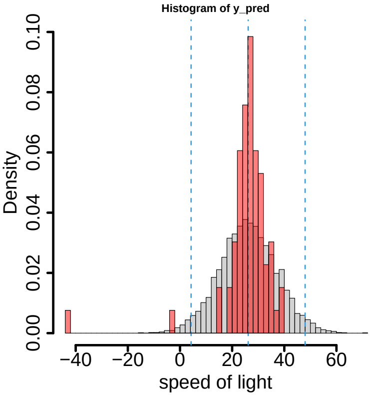
- Posterior summary of θ and σ^2 (think about the implied loss function!)

```
> ### summaries of the margianl posterior of theta
> post_m_th <- mean(my_th)
> post_sd_th <- sd(my_th)
> ci_th <- quantile(my_th, prob=c(0.025, 0.975))
> post_m_th
[1] 26.30754
> post_sd_th
[1] 1.355212
> ci_th
    2.5%    97.5%
23.66675 29.01357
>
> ### summaries of the margianl posterior of sig2
> post_m_sig2 <- mean(my_sig2)
> post_sd_sig2 <- sd(my_sig2)
> ci_sig2 <- quantile(my_sig2, prob=c(0.025, 0.975))
> post_m_sig2
[1] 119.0088
> post_sd_sig2
[1] 21.49393
> ci_sig2
     2.5%     97.5%
 84.55515 167.76078
>
```

† Example: Estimating the speed of light (contd)

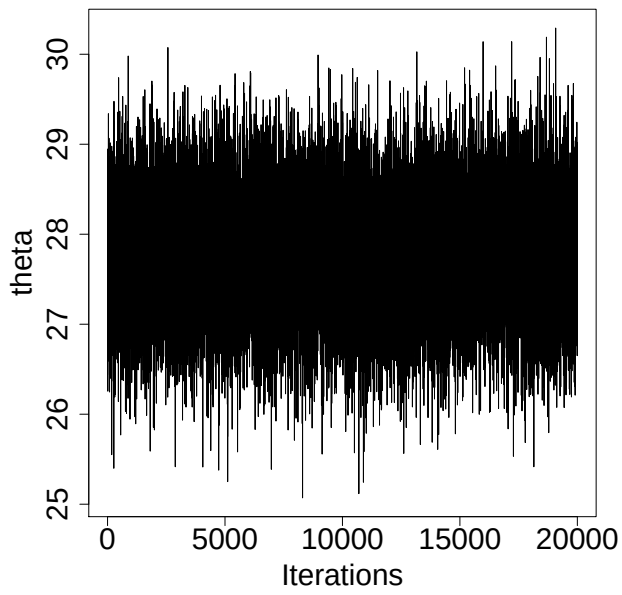
- Summary of the posterior predictive distribution of unobserved *y*

```
> ###%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
> #%%%## preditive distribution
> y_pred <- rnorm(length(my_th), my_th, sqrt(my_sig2))
> mean(y_pred)
[1] 26.24271
> sd(y_pred)
[1] 11.01951
> quantile(y_pred, prob=c(0.025, 0.975))
     2.5%      97.5%
 4.343315 47.816016
>
```

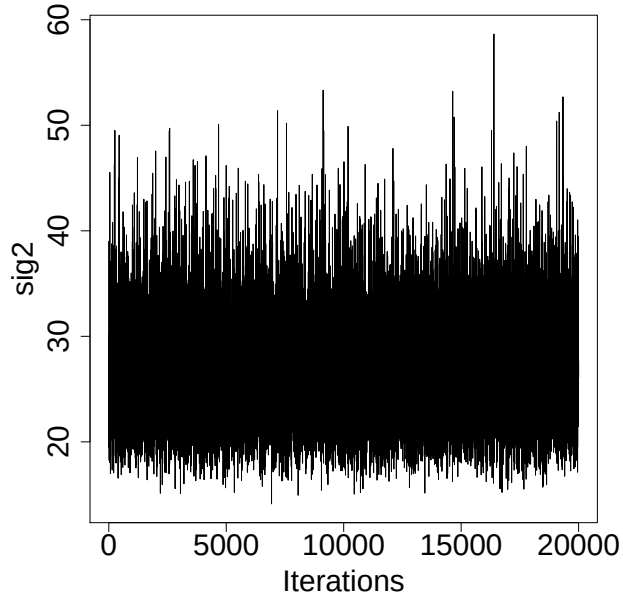


† Example: Estimating the speed of light (contd - redo)

- Remove the two outlying measurements and reanlyze the data with the same model.
- Check mixing and convergence of the Markov chain.



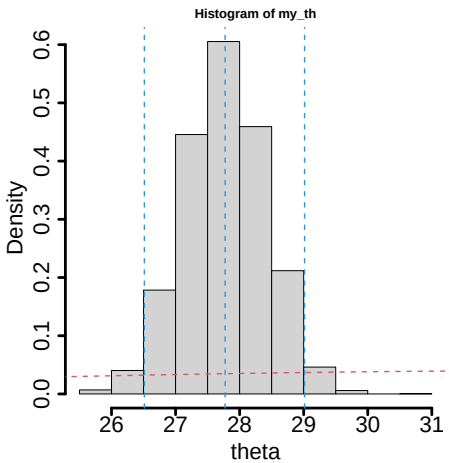
(a) θ



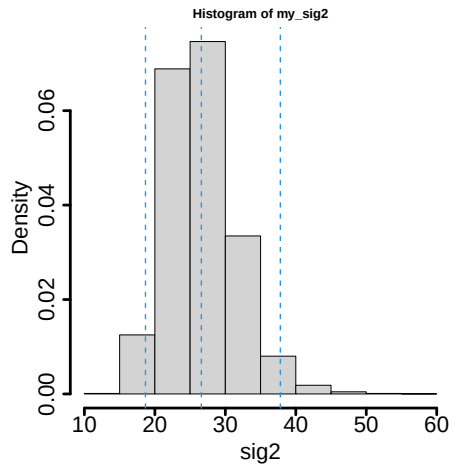
(b) σ^2

† Example: Estimating the speed of light (contd - redo)

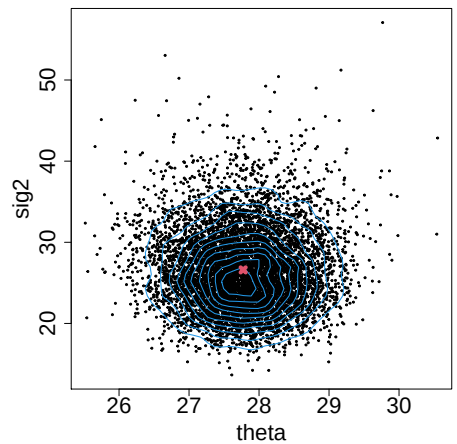
- Posterior summary of θ and σ^2



(a) θ



(b) σ^2



(c) Joint

† Example: Estimating the speed of light (contd - redo)

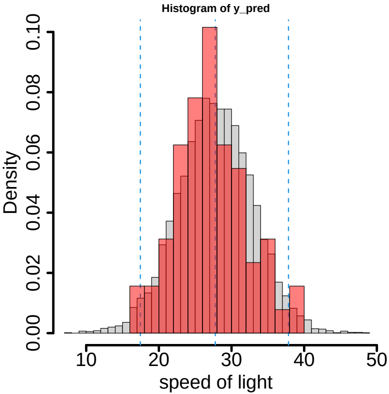
- Posterior summary of θ and σ^2

```
> ### summaries of the margianl posterior of theta
> post_m_th <- mean(my_th)
> post_sd_th <- sd(my_th)
> ci_th <- quantile(my_th, prob=c(0.025, 0.975))
> post_m_th
[1] 27.77308
> post_sd_th
[1] 0.6430802
> ci_th
    2.5%    97.5%
26.51241 29.01371
>
> ### summaries of the margianl posterior of sig2
> post_m_sig2 <- mean(my_sig2)
> post_sd_sig2 <- sd(my_sig2)
> ci_sig2 <- quantile(my_sig2, prob=c(0.025, 0.975))
> post_m_sig2
[1] 26.597
> post_sd_sig2
[1] 4.862329
> ci_sig2
    2.5%    97.5%
18.66827 37.80316
```

† Example: Estimating the speed of light (contd - redo)

- Summary of the posterior predictive distribution of unobserved *y*

```
>
> ###%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
> #%%%## preditive distribution
> y_pred <- rnorm(length(my_th), my_th, sqrt(my_sig2))
> mean(y_pred)
[1] 27.76714
> sd(y_pred)
[1] 5.173034
> quantile(y_pred, prob=c(0.025, 0.975))
    2.5%    97.5%
17.45357 37.83828
>
```



† **Hypothesis Testing** (CR Chapter 5 and JB Sec 4.3.3)

- Consider a statistical model $f(x \mid \theta)$ with $\theta \in \Theta$.
- Specify
  - ⋆⋆ Null hypothesis $H_0 : \theta \in \Theta_0$
  - ⋆⋆ Alternative hypothesis $H_1 : \theta \in \Theta_1$

  where $\Theta_0, \Theta_1 \subset \Theta$, $\theta \in \Theta_0 \cup \Theta_1 = \Theta$.
- Consider a test procedure $\psi \in \mathcal{D} = \{0, 1\}$, where
  - ⋆⋆ $\psi = 1$: conclude $H_0 : \theta \in \Theta_0$
  - ⋆⋆ $\psi = 0$: conclude $H_1 : \theta \in \Theta_1$

- Consider the loss function proposed by Neyman-Pearson (a.k.a. the 0-1 loss function)

$$L(\theta, \psi) = \begin{cases} 1 & \text{if } \psi \neq \mathbb{I}_{\Theta_0}(\theta), \\ 0 & \text{otherwise.} \end{cases}$$

- Base the decision upon the posterior probability that the hypothesis is true.

- Recall: the *Bayesian decision* is

$$\psi^{\pi}(x) = \begin{cases} 1 & \text{if } P^{\pi}(\theta \in \Theta_0 \mid x) > P^{\pi}(\theta \in \Theta_0^c \mid x), \\ 0 & \text{otherwise.} \end{cases}$$

$\star\star$ $P^{\pi}(\theta \in \Theta_0 \mid x) = P^{\pi}(H_0 \text{ is true} \mid x)$

$\Rightarrow$ we choose the hypothesis with the largest posterior probability.

- Consider the weighted 0-1 loss function

$$L(\theta, \psi) = \begin{cases} 0 & \text{if } \psi = \mathbb{I}_{\Theta_0}(\theta), \\ a_0 & \text{if } \theta \in \Theta_0 \text{ and } \psi = 0, \\ a_1 & \text{if } \theta \notin \Theta_0 \text{ and } \psi = 1. \end{cases}$$

⋆⋆ For a wrong answer under H_0, we lose by a_0.

⋆⋆ Suppose $a_0 > a_1$ (i.e., larger a_0/a_1) $\Rightarrow$ We lose more when we reject the true H_0 than when we reject the true H_1 (the more important a wrong answer under H_0 is relative to H_1).

⋆⋆ A large value of a_0/a_1 guards against falsely rejecting H_0.

- **Prop 5.2.2** The Bayesian estimator associated with a prior distribution π is

$$\psi^{\pi}(x) = \begin{cases} 1 & \text{if } P^{\pi}(\theta \in \Theta_0 \mid x) > \frac{a_1}{a_0+a_1}, \\ 0 & \text{otherwise.} \end{cases}$$

$\star\star$ Reject H_0 if the posterior probability of H_0 is too small.

$\star\star$ *How small?* smaller than $a_1/(a_0 + a_1)$

$\star\star$ The acceptance level $a_1/(a_0+a_1)$ is determined by the choice of the loss function.

$\star\star$ Note that ψ^{π} only depends on a_0/a_1 (rather than their actual values).

- **Example 5.2.4** Consider $x \sim \mathrm{N}(\theta, \sigma^2)$ and $\theta \sim \mathrm{N}(\mu, \tau^2)$. Test $H_0 : \theta < 0$ under the $a_0 - a_1$ loss.

† Bayes Factor – another way to do a testing.

- **Def 5.2.5** The Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses , i.e.,

$$B_{01}^{\pi} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}$$

$$\Rightarrow \underbrace{\frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)}}_{\text{posterior odds}} = \underbrace{B_{01}^{\pi}}_{\text{Bayes Factor}} \underbrace{\frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}}_{\text{prior odds}}$$

⋆⋆ The Bayes factor evaluates the modification of the odds of Θ_0 against Θ_1 due to data (naturally compared to 1)

⋆⋆ The Bayes factor can be interpreted as how much the data favors $H_0 : \theta \in \Theta_0$ over $H_1 : \theta \in \Theta_1$.

• **JB Example 1 - p147** Recall the IQ test problem. The child taking the IQ test is to be classified as having below average IQ (less than 100) or above average IQ (greater than 100). Formally, test $H_0 : \theta \geq 100$ versus $H_1 : \theta < 100$. Recall $\theta \sim \mathrm{N}(100, 225)$ and $\theta \mid x \sim \mathrm{N}(110.39, 63.23)$. We find

$$P(\theta \geq 100) = 0.5, \qquad P(\theta < 100) = 0.5$$
$$P(\theta \geq 100 \mid x) = 0.894, \qquad P(\theta < 100 \mid x) = 0.106.$$

- ⋆⋆ In prior, H_0 and H_1 are viewed as equally plausible (the prior odds is 1).
- ⋆⋆ The Bayes factor $B_{01}^{\pi}(x) = (0.894/0.106)/(0.5/0.5) = 8.44$. That is, the odds of Θ_0 against Θ_1 increased by 8.44 times after observing data.
- ⋆⋆ In other words, the data is in favor of Θ_0.

† Bayes Factor for Simple Hypotheses

- Consider testing a simple null hypothesis against a simple alternative hypothesis; $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ where $\Theta = \{\theta_0, \theta_1\}$.
- $\rho_0 = \pi(\theta \in \Theta_0)$ and $\pi(\theta \in \Theta_1) = 1 - \rho_0 = \rho_1$.
- Find

$$B_{01}^{\pi}(x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)}.$$

⋆⋆ The Bayes factor does not depend on prior.

⋆⋆ A testing procedure solely based on the Bayes factor becomes the classical *likelihood ratio*.

† Bayes Factor for a General Case

- $\rho_0 = \pi(\theta \in \Theta_0)$ and $\pi(\theta \in \Theta_1) = 1 - \rho_0 = \rho_1$. And let

$$\theta \sim \begin{cases} \pi_0(\theta) & \text{if } \theta \in \Theta_0, \\ \pi_1(\theta) & \text{if } \theta \in \Theta_1. \end{cases}$$

$\Rightarrow$ Observe that $\pi(\theta) = \rho_0 \pi_0(\theta) + (1 - \rho_0)\pi_1(\theta)$.

- We find

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x \mid \theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x \mid \theta)\pi_1(\theta)d\theta} = \frac{m_0(x)}{m_1(x)}.$$

$\star\star$ Note that we have the marginals instead of the likelihoods.

$\star\star$ Observe $B_{01}^{\pi}(x)$ depends on *both* prior and data.

$\star\star$ When $H_0 : \theta = \theta_0$, observe $B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{m_1(x)}$.

† How to connect the Bayes Factor to the decision theoretic testing procedure?

- Recall

$$\psi^{\pi}(x) = \begin{cases} 1 & \text{if } P^{\pi}(\theta \in \Theta_0 \mid x) > \frac{a_1}{a_0+a_1}, \\ 0 & \text{otherwise.} \end{cases}$$

- This is equivalent to accepting H_0 when

$$B_{01}^{\pi}(x) > \frac{a_1}{a_0} \Big/ \frac{\rho_0}{\rho_1} = \frac{a_1\rho_1}{a_0\rho_0},$$

where $\rho_0 = \pi(\theta \in \Theta_0)$ and $\rho_1 = \pi(\theta \in \Theta_1)$.

- Many Bayesians consider the Bayes factor on its own ground (*outside a true decision-theoretic setting*).

- Jeffreys developed a scale to judge the evidence in favor of or against H_0 brought by the data,

  * $\log_{10}(B_{01}^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is *poor*,

  * if it is between 0.5 and 1, it is *substantial*,

  * if it is between 1 and 2, it is *strong*, and

  * if it is above 2, it is *decisive*

  ⋆⋆ This gives guidelines for Bayesian testing of hypotheses depending on the value of log-Bayes factor.

- The exact bounds can be driven based on a loss function.

† Testing a point null hypothesis

- Consider $H_0 : \theta = \theta_0$ vs $H_0 : \theta \neq \theta_0$.
- Consider the following prior

$$\theta \begin{cases} = \theta_0 & \text{with probability } \rho_0, \\ \sim g_1(\theta) & \text{with probability } \rho_1 = 1 - \rho_0, \end{cases}$$

where probability distribution $g_1(\theta)$ gives probability zero to the event $\theta = \theta_0$.

$\Rightarrow$ We rewrite

$$\pi(\theta) = \rho_0 \delta_{\theta_0} + (1 - \rho_0) g_1(\theta),$$

where δ_{θ_0} is the Dirac mass at θ_0.

- Let's find $\pi(\theta = \theta_0 \mid x)$.

$\star\star$ *Step 1* The marginal distribution for X is

$$\begin{aligned} m(x) &= \int f(x \mid \theta)\pi(\theta)d\theta \\ &= \rho_0 f(x \mid \theta_0) + (1-\rho_0)\int f(x \mid \theta)g_1(\theta)d\theta \\ &= \rho_0 f(x \mid \theta_0) + (1-\rho_0)m_1(x). \end{aligned}$$

- Let's find $\pi(\theta = \theta_0 \mid x)$ (contd).

  $\star\star$ *Step 2* The posterior probability of $\theta = \theta_0$ is

$$\begin{aligned}
\pi(\theta = \theta_0 \mid x) &= \frac{\rho_0 f(x \mid \theta_0)}{m(x)} \\
&= \frac{\rho_0 f(x \mid \theta_0)}{\rho_0 f(x \mid \theta_0) + (1 - \rho_0) m_1(x)} \\
&= \left\{ 1 + \frac{1 - \rho_0}{\rho_0} \frac{m_1(x)}{f(x \mid \theta_0)} \right\}^{-1}.
\end{aligned}$$

• **Example 5.2.8** (Example 5.2.4 continued) Consider the test of $H_0 : \theta = 0$. It seems reasonable to choose π_1 as $\mathrm{N}(\mu, \tau^2)$ and $\mu = 0$, if no additional information is available. Find the posterior probability, $\pi(\theta = 0 \mid x)$.

- **Example 5.2.8** (contd)

⋆⋆ $\rho_0 = 1/2$ and $\tau = \sigma$

Table 5.2.2. *Posterior probabilities of* $\theta = 0$ *for different values of* $z = x/\sigma$ *and for* $\tau = \sigma$.

z	0	0.68	1.28	1.96
$\pi(\theta = 0\|z)$	0.586	0.557	0.484	0.351

⋆⋆ $\rho_0 = 1/2$ and $\tau^2 = 10\sigma^2$ (more diffuse prior)

Table 5.2.3. *Posterior probabilities of* $\theta = 0$ *for* $\tau^2 = 10\sigma^2$ *and* $z = x/\sigma$.

z	0	0.68	1.28	1.96
$\pi(\theta = 0\|x)$	0.768	0.729	0.612	0.366

- **(JB p151)** Let's change the example a bit. Now we have $x_i \mid \theta \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$, $i = 1, \ldots, n$ and let $\sigma = \tau$. Show the posterior probability on the null hypothesis is shown below to be given by

$$\pi(\theta = 0 \mid \bar{x}) = \frac{1}{1 + \frac{1}{\sqrt{n+1}} \exp\left\{ \frac{g^2}{2(1+1/n)} \right\}},$$

where $g = \frac{\sqrt{n|\bar{x}|}}{\sigma}$.

$\star\star$ Find the p-value.

- Values of the Posterior Probabilities $P(H_0 \mid x)$.

p-value	g	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 1000$
0.100	1.645	0.42	0.44	0.47	0.56	0.65	0.72	0.89
0.050	1.960	0.35	0.33	0.37	0.42	0.52	0.60	0.82
0.010	2.576	0.21	0.13	0.14	0.16	0.22	0.27	0.53
0.001	3.291	0.086	0.026	0.024	0.026	0.034	0.045	0.124

- Observe when $g = 1.96$ for $n = 50$

$\star\star$ The frequentist researcher could reject H_0 at $p = 0.05$

$\star\star$ The Bayesian hypothesis tester, the evidence against the null hypothesis is weaker (little or no evidence against H_0).

$\star\star$ (but keep in mind) p-values are not a posterior probability of a hypothesis! For more, read JB 4.3.3.

- Testing with a noninformative prior

Example 5.2.9 Consider $x \sim \mathrm{N}(\theta, 1)$ and test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. For the diffuse distribution, $\pi(\theta) = 1$, find $P(\Theta_0 \mid x)$.

• **Example 5.2.8** (contd, Section 5.2.5) Assume $x \sim \mathrm{N}(\theta, 1)$. Consider the test of $H_0 : \theta = 0$ to test against $H_1 : \theta \neq 0$. To express vague prior information, assume the improper prior $\pi(\theta) = c$ on $\{\theta \neq 0\}$.

- **Example 5.2.8** (contd)

  ⋆⋆ $\pi(\theta = \theta_0 \mid x)$ for the Jeffreys prior $\pi(\theta) = 1$

Table 5.2.5. *Posterior probabilities of* $\theta = 0$ *for the Jeffreys prior* $\pi(\theta) = 1$.

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.285	0.195	0.089	0.055	0.014

  ⋆⋆ $\pi(\theta = \theta_0 \mid x)$ for the Jeffreys prior $\pi(\theta) = 10$

Table 5.2.6. *Posterior probabilities of* $\theta = 0$ *for the Jeffreys prior* $\pi(\theta) = 10$.

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.0384	0.0236	0.0101	0.00581	0.00143

• **Example 5.2.8**(contd) Another illustration of the delicate issue of improper priors in testing setting.

$$\begin{aligned}\pi(\theta = \theta_0 \mid x) &= \left\{1 + \frac{\rho_1}{\rho_0}\frac{m_1(x)}{f(x \mid \theta_0)}\right\}^{-1} \\ &= \left\{1 + \frac{\rho_1}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right\}^{-1}.\end{aligned}$$

⋆⋆ For every x, as $\tau^2 \to \infty$, $\pi(\theta = \theta_0 \mid x) \to 1$.

⋆⋆ Compare to $\pi(\theta = \theta_0 \mid x)$ with the improper prior $\pi(\theta) = 1$ on $\{\theta \neq 0\}$,

$$\pi(\theta = 0 \mid x) = \frac{1}{1 + \sqrt{2\pi}\exp(x^2/2)}$$

⋆⋆ i.e., limiting arguments are not valid in the testing settings and prevent an alternative derivation of noninformative answers.

† Testing with noninformative priors. (contd)

- In many (**not all**) one-sided testing situations (& estimation situations), vague prior information tends to result in posterior probabilities that are similar to p-values.
- Improper priors should not be used at all in tests – DeGroot, 1973
- A testing problem cannot be treated in a coherent way if no prior information is available.
- Read CR 5.2.5 very interesting things regarding testing with improper priors.

- How about Bayes factor when the prior is improper?
  - ⋆⋆ Intrinsic Bayes factor and fractional Bayes factor.

    Both use some part of data to make the improper prior proper ("propertize" the improper prior) and proceed the posterior inference **as if** it were a regular proper prior for the remainder of the sample.

  - ⋆⋆ CR 5.2.6 Pseudo-Bayes Factor.