5.1-5.3

The normal model

Inference for the mean, conditional on the variance

Joint inference for the mean and variance

5.1 The normal model

A random variable Y is said to be normally distributed with mean θ and variance $\sigma^2 > 0$ if the density of Y is given by

$$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2}, \quad -\infty < y < \infty.$$

- The distribution is symmetric about θ, and the mode, median and mean are all equal to θ
- About 95% of the population lies within two standard deviations of the mean (more precisely, 1.96 standard deviations)
- If $X \sim \text{normal}(\mu, \tau^2)$, $Y \sim \text{normal}(\theta, \sigma^2)$ and X and Y are independent, then $aX + bY \sim \text{normal}(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$;

5.1 The normal model

The importance of the normal distribution stems primarily from the central limit theorem, which says that under very general conditions, the sum (or mean) of a set of random variables is approximately normally distributed. In practice, this means that the normal sampling model will be appropriate for data that result from the additive effects of a large number of factors.

5.1 The normal model

Central limit theorem (CLT)

- 每次從一個具有有限平均和變異數的隨機變數裡抽出n個數，並算出平平均
- 重複抽m次，而這m個平均數會呈現常態分佈
- 抽越多次(m越大)，越收斂於常態分佈

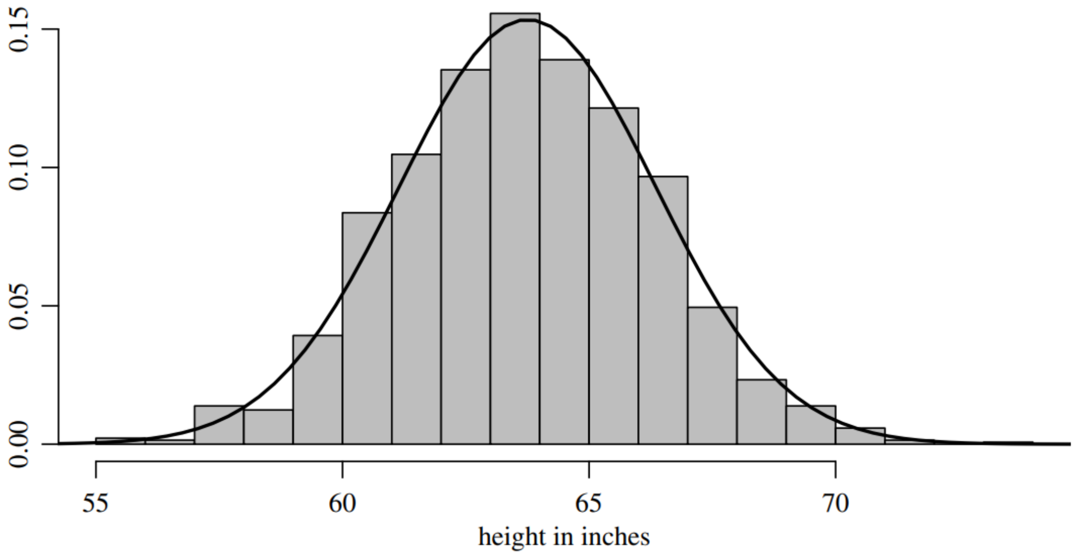


Fig. 5.2. Height data and a normal density with $\theta = 63.75$ and $\sigma = 2.62$.

5.1 The normal model

Additive effect:

When **two predictors do not interact**, we say that each predictor has an "**additive effect**" on the response. More formally, a regression model contains additive effects **if the response function can be written as a sum of functions of the predictor variables**:

$$\mu_y = f_1(x_1) + f_2(x_2) + \ldots + f_{p-1}(x_{p-1})$$

If the effects of these factors are approximately additive, then each height measurement yi is roughly equal to a linear combination of a large number of terms.

$$\begin{aligned}
y_1 &= a + b \times \text{gene}_1 + c \times \text{diet}_1 + d \times \text{disease}_1 + \cdots \\
y_2 &= a + b \times \text{gene}_2 + c \times \text{diet}_2 + d \times \text{disease}_2 + \cdots \\
&\vdots \\
y_n &= a + b \times \text{gene}_n + c \times \text{diet}_n + d \times \text{disease}_n + \cdots
\end{aligned}$$

5.2 Inference for the mean, conditional on the variance(variance 已知)

Suppose our model is $\{Y_1, \ldots, Y_n | \theta, \sigma^2\} \sim$ i.i.d. normal (θ, σ^2). Then the joint sampling density is given by

$$
\begin{aligned}
p(y_1, \ldots, y_n | \theta, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \theta, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \theta}{\sigma}\right)^2} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2} \sum \left(\frac{y_i - \theta}{\sigma}\right)^2\right\}.
\end{aligned}
$$

Expanding the quadratic term in the exponent, we see that $p(y_1, \ldots, y_n|\theta, \sigma^2)$ depends on $y_1, \ldots, y_n$ through

$$\sum_{i=1}^{n} \left(\frac{y_i - \theta}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum y_i^2 - 2\frac{\theta}{\sigma^2} \sum y_i + n\frac{\theta^2}{\sigma^2}.$$

From this you can show that $\{\sum y_i^2, \sum y_i\}$ make up a two-dimensional sufficient statistic. Knowing the values of these quantities is equivalent to knowing the values of $\bar{y} = \sum y_i/n$ and $s^2 = \sum(y_i - \bar{y})^2/(n-1)$, and so $\{\bar{y}, s^2\}$ are also a sufficient statistic.

Sufficient statistic:

「沒有任何其他可以以同一樣本中計算得出的統計量可以提供任何有關未知母數的額外訊息」

- making inference for θ when σ^2 is known
- use a conjugate prior distribution for θ

Posterior distribution

Prior distribution

$$\boxed{p(\theta|y_1,\ldots,y_n,\sigma^2)} \propto \boxed{p(\theta|\sigma^2)} \times e^{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2}$$

$$\propto p(\theta|\sigma^2) \times \boxed{e^{c_1(\theta-c_2)^2}}.$$

Must have

If $p(\theta|\sigma^2)$ is normal and $y_1, \ldots, y_n$ are i.i.d. normal(θ, σ^2),

then $p(\theta| y_1, \ldots, y_n, \sigma^2)$ is also a normal density.

Let's evaluate this claim: If $\theta \sim$ normal (μ_0, τ_0^2), then

$$\boxed{p(\theta|y_1,\ldots,y_n,\sigma^2)} = p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2)/p(y_1,\ldots,y_n|\sigma^2)$$

Posterior distribution

$$\propto p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2)$$

$$\propto \exp\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\}\exp\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\}.$$

Adding the terms in the exponents and ignoring the -1/2 for the moment, we have

$$\frac{1}{\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}(\sum y_i^2 - 2\theta\sum y_i + n\theta^2) = a\theta^2 - 2b\theta + c, \text{ where}$$

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}, \quad \text{and } c = c(\mu_0, \tau_0^2, \sigma^2, y_1, \ldots, y_n).$$

Now let's see if $p(\theta|\sigma^2, y_1, \ldots, y_n)$ takes the form of a normal density:

$$
\begin{aligned}
\underline{p(\theta|\sigma^2, y_1, \ldots, y_n)} &\propto \exp\{-\frac{1}{2}(a\theta^2 - 2b\theta)\} \\
&= \exp\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\} \\
&\propto \exp\{-\frac{1}{2}a(\theta - b/a)^2\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\theta - \boxed{b/a}}{\boxed{1/\sqrt{a}}}\right)^2\right\}
\end{aligned}
$$

the role of the mean

the role of standard deviation

- **This function has exactly the same shape as a normal density curve.**
- **Since probability distributions are determined by their shape, this is indeed a normal density**
- **refer to the mean and variance of this density as μ_n and τ_n^2, where**

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}.$$

Combining information

- The (conditional) posterior parameters τ_n^2 and μ_n combine the prior parameters τ_0^2 and μ_0 with terms from the data.

  - Posterior variance and precision: The formula for $1/\tau_n^2$ is

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \tag{5.1}$$

and so the prior *inverse* variance is combined with the inverse of the data variance. Inverse variance is often referred to as the *precision*. For the normal model let,

$\tilde{\sigma}^2 = 1/\sigma^2 =$ sampling precision, i.e. how close the y_i's are to θ;
$\tilde{\tau}_0^2 = 1/\tau_0^2 =$ prior precision;
$\tilde{\tau}_n^2 = 1/\tau_n^2 =$ posterior precision.

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

and so posterior information = prior information + data information.

Combining information

Posterior mean: Notice that

$$\mu_n = \boxed{\frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}}\mu_0 + \boxed{\frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}}\bar{y},$$

The weight on the prior mean is 1/ τ^2 **The weight on the sample mean is n/ σ^2**

The posterior mean is a weighted average of the prior mean and the sample mean.

Combining information

- If the prior mean were based on κ_0 prior observations from the same (or similar) population as $y_1, \ldots, y_n$, then we might want to set $\tau_0^2 = \sigma^2 / \kappa_0$, the variance of the mean of the prior observations. In this case, the formula for the posterior mean reduces to

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}.$$

Prediction

Consider predicting a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \dots, Y_n = y_n)$. To find the predictive distribution, let's use the following fact:

$$\{\tilde{Y}|\theta,\sigma^2\} \sim \text{ normal}(\theta,\sigma^2) \Leftrightarrow \boxed{\tilde{Y} = \theta + \tilde{\epsilon},} \ \{\tilde{\epsilon}|\theta,\sigma^2\} \sim \text{ normal}(0,\sigma^2)\,.$$

$\tilde{Y}$ is equal to θ plus some mean-zero normally distributed noise.

posterior mean of Y

$$\begin{aligned}\mathrm{E}[\tilde{Y}|y_1,\dots,y_n,\sigma^2] &= \mathrm{E}[\theta+\tilde{\epsilon}|y_1,\dots,y_n,\sigma^2]\\ &= \mathrm{E}[\theta|y_1,\dots,y_n,\sigma^2] + \mathrm{E}[\tilde{\epsilon}|y_1,\dots,y_n,\sigma^2]\\ &= \mu_n + 0 = \mu_n\end{aligned}$$

posterior variance of Y

$$\begin{aligned}\mathrm{Var}[\tilde{Y}|y_1,\dots,y_n,\sigma^2] &= \mathrm{Var}[\theta+\tilde{\epsilon}|y_1,\dots,y_n,\sigma^2]\\ &= \mathrm{Var}[\theta|y_1,\dots,y_n,\sigma^2] + \mathrm{Var}[\tilde{\epsilon}|y_1,\dots,y_n,\sigma^2]\\ &= \tau_n^2 + \sigma^2\,.\end{aligned}$$

Prediction

- the sum of independent normal random variables is also normal

$$\tilde{Y} = \theta + \tilde{\epsilon}.$$

$$\tilde{Y}|\sigma^2, y_1, \ldots, y_n \sim \text{ normal}(\mu_n, \tau_n^2 + \sigma^2)\,.$$

Posterior variance τ_n^2 of θ 反映了我們對 θ 的不確定性。

隨著樣本量 (n) 增加，我們對 θ 的估計會越來越準確，因此 τ_n^2會趨近於零。

隨著樣本量的增加，我們對 θ 的不確定性會減少，但樣本的內在變異性 (σ^2)是不可避免的，

因此對於新樣本 $\tilde{Y}$ 的不確定性永遠不會低於 σ^2。

5.3 Joint inference for the mean and variance

1. posterior inference

$$p(\theta, \sigma^2 | y_1, \ldots, y_n) = p(y_1, \ldots, y_n | \theta, \sigma^2) p(\theta, \sigma^2) / p(y_1, \ldots, y_n) .$$

2. developing a simple conjugate class of prior distributions

$$p(\theta, \sigma^2) = p(\theta | \sigma^2) p(\sigma^2)$$

3. If σ^2 were known, then a conjugate prior distribution for θ was normal (μ_0, τ_0^2), the particular case in which $\tau_0^2 = \sigma^2 / \kappa_0$:

the sample size from a set of prior observations.

$$p(\theta, \sigma^2) = p(\theta | \sigma^2) p(\sigma^2) = \text{dnorm}(\theta, \boxed{\mu_0}, \tau_0 = \sigma / \sqrt{\boxed{\kappa_0}}) \times p(\sigma^2) .$$

the mean from a set of prior observations.

- For σ^2, we need a family of prior distributions that has support on $(0,\infty)$.

➔Gamma family is, but it is not conjugate for the normal variance.

➔However, it does turn out to be a conjugate class of densities for $1/\sigma^2$ (the precision)

➔When using such a prior distribution, we say that σ^2 has an inverse-gamma distribution:

$$\begin{aligned} \text{precision} = 1/\sigma^2 &\sim \text{gamma}(a,b) \\ \text{variance} = \sigma^2 &\sim \text{inverse-gamma}(a,b) \end{aligned}$$

為了後續的解讀，將 a & b 換個形式

$$\begin{aligned}\text{precision} &= 1/\sigma^2 \sim \text{gamma}(a,b) \\ \text{variance} &= \sigma^2 \sim \text{inverse-gamma}(a,b)\end{aligned} \rightarrow \quad 1/\sigma^2 \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right).$$

- $\mathrm{E}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$;
- $\mathrm{mode}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$, so $\mathrm{mode}[\sigma^2] < \sigma_0^2 < \mathrm{E}[\sigma^2]$;
- $\mathrm{Var}[\sigma^2]$ is decreasing in ν_0.

(σ_0^2, ν_0) **are the <u>sample variance</u> of prior observations and the <u>sample size</u> of prior observations.**

Posterior inference

Suppose our prior distributions and sampling model are as follows:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\ \theta|\sigma^2 &\sim \text{normal}(\mu_0, \sigma^2/\kappa_0) \\ Y_1, \ldots, Y_n|\theta, \sigma^2 &\sim \text{ i.i.d. normal } (\theta, \sigma^2) . \end{aligned}$$

如同prior distribution 可以分解成 $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$

posterior distribution 也可以變成 $p(\theta, \sigma^2|y_1, \ldots, y_n) = \boxed{p(\theta|\sigma^2, y_1, \ldots, y_n)}p(\sigma^2|y_1, \ldots, y_n)$.

可以從數據中知道

Plugging in $\sigma^2/\ \kappa_0$ for τ_0^2, we have $\{\theta|y_1, \ldots, y_n, \sigma^2\} \sim \text{normal}(\mu_n, \sigma^2/\kappa_n)$, where

$$\kappa_n = \kappa_0 + n \text{ and } \mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n} .$$

If μ_0 is the mean of κ_0 prior observations, $E[\theta| y_1, \ldots, y_n, \sigma^2]$ is the sample mean of the current and prior observations.

$\text{Var}[\theta| y_1, \ldots, y_n, \sigma^2]$ is σ^2 divided by the total number of observations, both prior and current.

Posterior inference

- The posterior distribution of σ^2 can be obtained by performing an integration over the unknown value θ

$$
\begin{aligned}
p(\sigma^2|y_1,\ldots,y_n) &\propto p(\sigma^2)p(y_1,\ldots,y_n|\sigma^2) \\
&= p(\sigma^2)\int p(y_1,\ldots,y_n|\theta,\sigma^2)p(\theta|\sigma^2)\,d\theta\,.
\end{aligned}
$$

posterior sums of squares

$$
\begin{aligned}
\Rightarrow \quad \{1/\sigma^2|y_1,\ldots,y_n\} &\sim \text{gamma}(\nu_n/2, \boxed{\nu_n\sigma_n^2}/2), \text{ where} \\
\nu_n &= \boxed{\nu_0} + n \\
\sigma_n^2 &= \frac{1}{\nu_n}\left[\boxed{\nu_0\sigma_0^2} + \boxed{(n-1)s^2} + \frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2\right].
\end{aligned}
$$

Sum of square

prior sums of squares

$\boxed{\nu_0}$ is as a prior sample size, from which a prior sample variance of σ_0^2 has been obtained.

Posterior inference

posterior sums of squares

$$\{1/\sigma^2|y_1,\ldots,y_n\} \sim \text{gamma}(\nu_n/2, \boxed{\nu_n\sigma_n^2}/2), \text{ where}$$

$$\nu_n = \nu_0 + n$$

Sum of square

$$\sigma_n^2 = \frac{1}{\nu_n}\boxed{[\nu_0\sigma_0^2}+\boxed{(n-1)s^2}+\frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2].$$

prior sums of squares

等號兩邊同乘 ν_n

➔ $$\nu_n\sigma_n^2 = [\nu_0\sigma_0^2 + (n-1)s^2 + \boxed{\frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2}].$$?

posterior sum of squares = prior sum of squares + data sum of squares +.....

! : A large value of ($\bar{y} - \mu_0)^2$ increases the posterior probability of a large σ^2
If we want to think of μ_0 as the sample mean of κ_0 prior observations with variance σ^2 ,
then is an estimate of σ^2

Monte Carlo sampling

- Goal: population mean θ **determined by the marginal posterior distribution of θ given the data**
- Want to know: $\boxed{\mathrm{E}[\theta|y_1,\ldots,y_n] \quad \mathrm{sd}[\theta|y_1,\ldots,y_n] \quad \Pr(\theta_1 < \theta_2|\, y_{1,1},\ldots,\, y_{n_2,2})}$
- But we only know that the conditional distribution of θ given the data and σ^2 is normal, and that σ^2 given the data is inverse-gamma.
- 生成 θ 的 marginal samples ，那麼我們可以使用Monte Carlo method 來近似上面我們想知道的數值
- $p(\theta| y_1, \ldots, y_n)$ ➔marginal samples of θ ➔ can use Monte Carlo method to approximate
- samples of θ and σ^2 can be generated by their joint posterior distribution

Monte Carlo sampling

Consider simulating parameter values using the following Monte Carlo procedure:

$$
\begin{aligned}
\sigma^{2(1)} &\sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), & \theta^{(1)} &\sim \text{normal}(\mu_n, \sigma^{2(1)}/\kappa_n) \\
&\vdots & &\vdots \\
\sigma^{2(S)} &\sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), & \theta^{(S)} &\sim \text{normal}(\mu_n, \sigma^{2(S)}/\kappa_n)\,.
\end{aligned}
$$

are independent samples from the joint posterior distribution of $p(\theta, \sigma^2 \mid y_1, \ldots, y_n)$

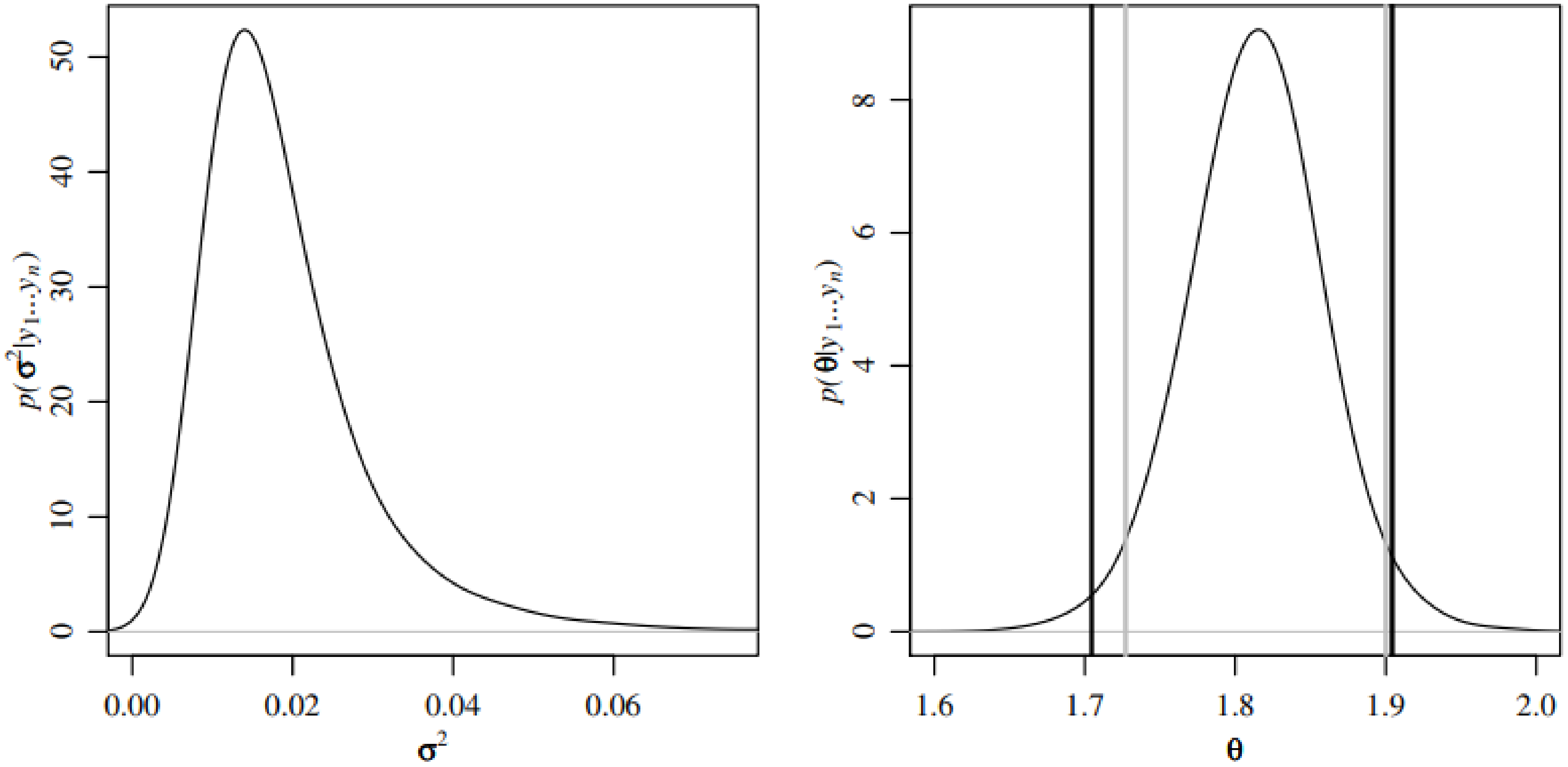
Each $\theta^{(s)}$ is sampled from its conditional distribution given the data and $\sigma^2 = \sigma^{2(S)}$

The simulated sequence $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ can be seen as independent samples from the marginal posterior distribution of $p(\theta \mid y_1, \ldots, y_n)$

while $\theta^{(1)}, \ldots, \theta^{(s)}$ are indeed each conditional samples, they are each conditional on different values of σ 2 . Taken together, they constitute marginal samples of θ

Monte Carlo sampling

- Black part: (1.70, 1.90), a frequentist 95% confidence interval obtained from the t-test.
- Grey part: (1.73, 1.90), a 95% quantile-based posterior interval for θ



Monte Carlo samples from and estimates of the joint and
marginal distributions of the population mean and variance.

Monte Carlo sampling

- Why they are so close?
- Because p(θ|y1, . . . , yn), the marginal posterior distribution of θ, can be obtained in a closed form.

the posterior distribution of $t(\theta) = \frac{(\theta - \mu_n)}{\sigma_n/\sqrt{\kappa_n}}$, given $\bar{y}$ and s^2, has a t-distribution with $\nu_0 + n$ degrees of freedom. If κ_0 and ν_0 are small, then the posterior distribution of $t(\theta)$ will be very close to the t_{n-1} distribution.

當 ν_0，κ_0 很小時，先驗分佈對後驗分佈的影響減少，後驗分佈更多地依賴於數據，

因此後驗分佈 t(θ)會近似於 t_{n-1}分佈。

Improper priors

ν_0, κ_0 as prior sample sizes, 越小越客觀，但越來越小會發生什麼事?

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_0 + n}$$
$$\sigma_n^2 = \frac{1}{\nu_0 + n}[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2],$$

and so as $\kappa_0, \nu_0 \to 0$,

$$\mu_n \to \bar{y}, \text{ and}$$
$$\sigma_n^2 \to \frac{n-1}{n} s^2 = \frac{1}{n} \sum (y_i - \bar{y})^2.$$

$$\{1/\sigma^2 | y_1, \ldots, y_n\} \sim \text{gamma}(\frac{n}{2}, \frac{n}{2} \frac{1}{n} \sum (y_i - \bar{y})^2)$$
$$\{\theta | \sigma^2, y_1, \ldots, y_n\} \sim \text{normal}(\bar{y}, \frac{\sigma^2}{n}).$$

Improper priors

Somewhat more formally, if we let $\tilde{p}(\theta, \sigma^2) = 1/\sigma^2$ (which is not a probability density) and set $p(\theta, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta, \sigma^2) \times \tilde{p}(\theta, \sigma^2)$, we get the same "conditional distribution" for θ but a gamma$(\frac{n-1}{2}, \frac{1}{2}\sum(y_i - \bar{y})^2)$ distribution for $1/\sigma^2$ (Gelman et al (2004), Chapter 3). You can integrate this latter joint distribution over σ^2 to show that

$$\frac{\theta - \bar{y}}{s/\sqrt{n}}|y_1, \ldots, y_n \sim t_{n-1}.$$

It is interesting to compare this result to the sampling distribution of the t-statistic, conditional on θ but unconditional on the data:

$$\frac{\bar{Y} - \theta}{s/\sqrt{n}}|\theta \sim t_{n-1}.$$

Improper priors

- 在你抽樣數據之前，對於樣本平均 (\bar{Y}) 與母體平均 (θ) 之間的偏差的不確定性，可以用 t_{n-1} 分佈來表示。這表示在抽樣之前，我們對這個偏差的了解是基於 t_{n-1}分佈。
- 在你抽樣數據之後，對於這個偏差的不確定性仍然可以用 t_{n-1}分佈來表示。不同的是，在抽樣之前，樣本平均 (\bar{Y}) 和母體平均 (θ) 都是未知的；而在抽樣之後，樣本平均 (\bar{Y}) 已知，這提供了關於 (θ) 的信息。
- 沒有適當的先驗概率分佈可以導致上述的 t_{n-1}後驗分佈，因此基於這個後驗分佈的推論並不是正式的貝葉斯推論。然而，有時候這樣的極限處理會導致合理的答案。
- Stein (1955) 的理論結果表明，從決策理論的角度來看，任何合理的估計量都是貝葉斯估計量或一系列貝葉斯估計量的極限，並且任何貝葉斯估計量都是合理的（技術術語是可允許的；