Bayesian Modeling

The normal model

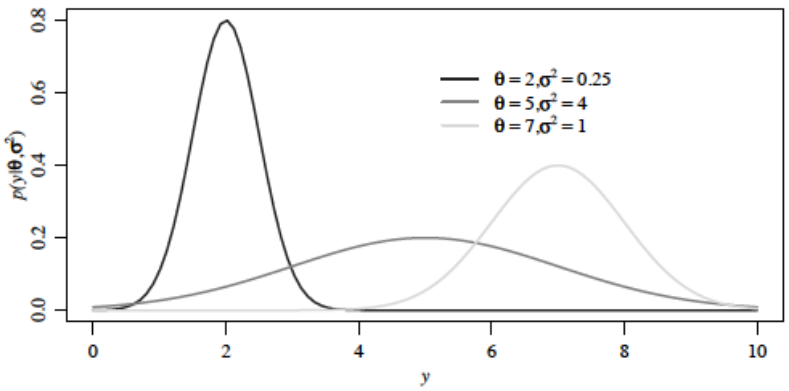
Yixin Wang

5 The normal model

5.1 The normal / Gaussian distribution

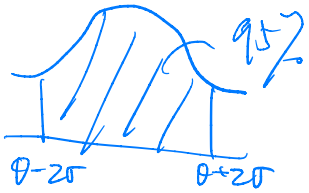
A random variable Y is said to be normally distributed with mean θ and variance $\sigma^2 > 0$ if the density of Y takes the form

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



location-scale family of distributions

location ≠ mean esp asymmetric dist.



$Z_{97.5\%} = 1.96$

$Z_{2.5\%} = -1.96$

Several important properties

- the distribution is symmetric about θ; the location, median and mean are all equal to θ
- σ^2 represents the spread of the mass: about 95% of the population lies within $(\theta \pm 2\sigma)$
- if $X \sim \text{Normal}(\mu, \tau^2)$ and $Y \sim \text{Normal}(\theta, \sigma^2)$, and X and Y are independent, then

$$aX + bY \sim \text{Normal}(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2).$$

Normal distribution is one of the most useful and widely utilized model in statistical sciences. Its important stems primarily from the central limit theorem, which says that under very general conditions, the empirical average of a collection of random variables is approximately distributed according to the Gaussian (normal) distribution.

$$E(aX+bY) = aE(X) + bE(Y) = a\mu + b\theta$$

$$\text{Var}(aX+bY) \overset{X \perp Y}{=} \text{Var}(aX) + \text{Var}(bY)$$

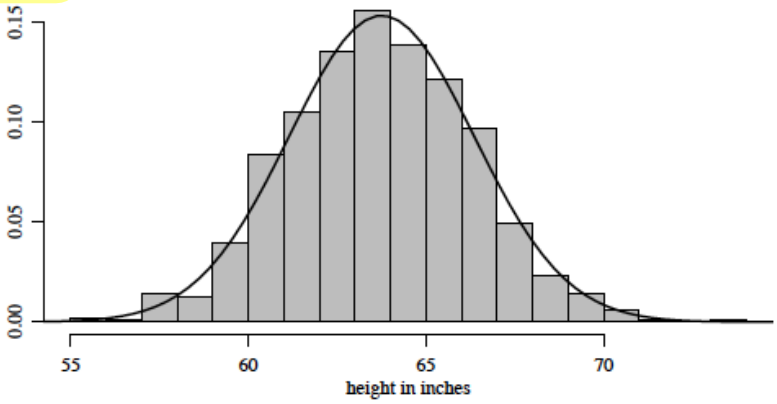
$$= a^2\text{Var}(X) + b^2\text{Var}(Y)$$

$$= a^2\tau^2 + b^2\sigma^2$$

Example 5.1. The following figure shows a normal density function overlay over the histogram of heights of $n = 1375$ women over age 18 collected in a study of 1100 English families from 1893 to 1898. One explanation for the variability in heights among these women is that the women were heterogenous in terms of a number of factors controlling human growth, such as genetics, diet, disease, stress and so on. Variability in such factors results in variability in height. Thus, letting y_i be the height in inches of woman i, a simple additive model for height might be

$$y_i = a + b \times \text{gene}_i + c \times \text{diet}_i + d \times \text{disease}_i + \ldots$$

where gene_i might denote the presence of a particular height-promoting gene, diet_i might measure some aspect of woman i's diet, and so on. Now, there may be a very large number of genes, dietary factors, and so on that contributes to a woman's height. If the effects of these factors are additive, then the height of a random woman may be modeled as a linear combination of a large number of random variables. The central limit theorem says that such a linear combination is approximately distributed according to a normal distribution.



CLT $Y_1, Y_2, \ldots$ iid some dist. with finite variance

Then $\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{n\to\infty}$ Normal Dist. (μ, σ^2)

$\mu = E[Y]$

$\sigma^2 = \frac{1}{n}\text{Var}(Y)$

---

generalization

$\sum_{i=1}^{n} a_i Y_i \to$ normal dist.

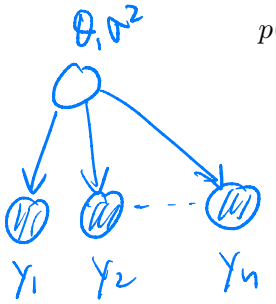
with suitable conditions on $\{a_i\}$.

Bernstein von Mises theorem: posterior → normal $n \to \infty$ under suitable conditions

n iid sample $Y_1, \dots, Y_n \in \mathbb{R}$

5.2 Inference of the mean with variance fixed

Given a sampling model $Y_1, \dots, Y_n | \theta, \sigma^2 \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2)$. The joint sampling pdf is



$$
\begin{aligned}
p(y_1, \dots, y_n | \theta, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \theta, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \theta)^2} \\
&= (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \Big\} \\
&= (2\pi\sigma^2)^{-n/2} \exp -\frac{1}{2}\Big\{ \frac{1}{\sigma^2} \sum y_i^2 - \frac{2\theta}{\sigma^2} \sum y_i + \frac{n\theta^2}{\sigma^2} \Big\}. \qquad (*)
\end{aligned}
$$

quadratic function of θ.

This expression shows that $\{\sum y_i^2, \sum y_i\}$ form a (two-dimensional) sufficient statistic for the normal model's parameters θ and σ^2. Equivalently, let $\bar{y} := \sum y_i / n$ and $s^2 := \sum (y_i - \bar{y})^2 / (n-1)$, then $(\bar{y}, s^2)$ is a sufficient statistic.

bijective transformation between $(\bar{y}, s^2)$ and $(\sum y_i^2, \sum y_i)$

Suppose that σ is fixed and known; the quantity of interest is θ. It is easy to see that the maximum likelihood estimate for θ is $\hat{\theta} = \bar{y}$.

MLE of θ

$\theta_{MLE} = \text{argmax } L(\theta) = \text{argmax } \log p(\text{Data} | \theta)$

$= \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$

↑ completing the square in $(*)$

Let us proceed to specifying a conjugate prior for θ. Given a (conditional) prior distribution $p(\theta|\sigma^2)$, the posterior pdf takes the form

$$\begin{aligned} p(\theta|y_1,\ldots,y_n) &\propto p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2) \\ &\propto p(\theta|\sigma^2)e^{-\frac{1}{2\sigma^2}\sum(\theta-y_i)^2}. \end{aligned}$$

↳ quadratic function of θ

The simplest possible form for a conjugate prior for θ is of the form $e^{c_1(\theta-c_2)^2}$. This suggests a normal distribution prior:

$$\text{Prior:}\theta \sim \text{Normal}(\mu_0, \tau_0^2).$$

μ_0, τ_0^2 hyperparameter (for the prior)

⇒ Normal dist is the conjugate prior with respect to the normal likelihood of the mean (location) parameter.

Continuing on the Bayesian update:

$$\begin{aligned} p(\theta|y_1,\ldots,y_n,\sigma^2) &\propto \exp -\frac{1}{2\tau_0^2}(\theta-\mu_0)^2 \times \exp -\frac{1}{2\sigma^2}\sum(\theta-y_i)^2 \\ &\propto \exp -\frac{1}{2}(a\theta^2 - 2b\theta + c), \end{aligned}$$

where it is easy to verify that

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2},$$

and c is independent of θ. Since the exponent of the posterior pdf is a quadratic form, with negative coefficient of the leading (second order) term, this must be the pdf of a normal distribution. Let us derive the corresponding mean and variance of the posterior.

$$\begin{aligned} p(\theta|\sigma^2,y_1,\ldots,y_n) &\propto \exp -\frac{1}{2}(a\theta^2 - 2b\theta) \\ &\propto \exp -\frac{1}{2}a(\theta - b/a)^2 \\ &= \text{Normal}(b/a, 1/a). \end{aligned}$$

$$\text{Normal}(\mu_n, \tilde{\tau}_n) = \text{Normal}\left(\frac{\frac{\mu_0}{\tau_0^2}+\frac{\sum y_i}{\sigma^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\right)$$

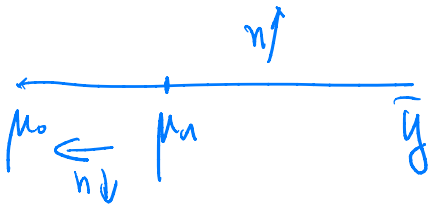
precision parametrization makes posterior update cleaner. instead of variance, also consider precision $\tilde{\tau} = \frac{1}{\tau}$

prior: $\tilde{\tau}_0 = \frac{1}{\tau_0^2}$ ← prior for θ

$\tilde{\sigma}^2 = \frac{1}{\sigma^2}$

posterior: $\tilde{\tau}_n = \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} = \tilde{\tau}_0 + n\tilde{\sigma}^2$

$$\mu_n = \frac{\tilde{\tau}_0\mu_0 + n\tilde{\sigma}^2 \cdot \bar{y}}{\tilde{\tau}_0 + n\tilde{\sigma}^2}$$ ← sample mean



Combining information Thus we have obtained that the posterior distribution of θ is indeed normal with mean μ_n and variance τ_n:

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}. \tag{5}$$

Not only is the posterior pdf remains a Gaussian, its corresponding parameters are obtained by combining information from the prior and the data in an intuitive way.

- Posterior variance: Inverse variance is often referred to as the *precision*.
  Let $\tilde{\sigma}^2 = 1/\sigma^2$ denote the sampling precision, $\tilde{\tau}_0^2 = 1/\tau_0^2$ the prior precision and $\tilde{\tau}_n^2 = 1/\tau_n^2$. Then

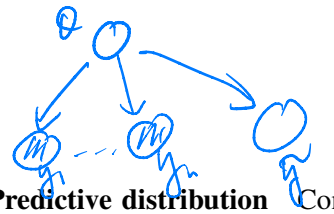
  $$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2,$$

  so the precision (for the parameter of interest) adds up with more data.

- Posterior mean:

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

  The posterior mean is a convex combination (i.e., weighted average) of the prior mean and the sample mean. The weights are corresponding precisions from either the prior or the data. The prior precision provides a shrinkage effect pulling the estimate toward the prior mean. As sample size n increases, the information from the data takes over.



$Y_1, \dots, Y_n, \tilde{Y} \mid \theta \overset{iid}{\sim} N(\cdot \mid \theta, \sigma^2)$ ← fixed

$\theta \sim N(\cdot \mid \mu_0, \tau_0^2)$

Predictive distribution Consider predicting a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \dots, Y_n = y_n)$. That is to find $p(\tilde{y}|y_1, \dots, y_n)$.

In general, to find the predictive distribution we need to perform an integration over the unknown θ. For the normal model, the situation is very easy (without having to perform this integration), due to the fact that a linear combination of normal random variables is another normal random variable.

In particular, for our model

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

Since $\theta|y_1, \dots, y_n \sim \text{Normal}(\mu_n, \tau_n^2)$ and ϵ is also normal and (conditionally) independent of θ. So,

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

$$p(\tilde{y} \mid y_1, \dots, y_n) = \int p(\tilde{y} \mid \theta) \cdot p(\theta \mid y_1, \dots, y_n)\, d\theta$$

$$= \int \text{Normal}(\tilde{y} \mid \theta, \sigma^2) \cdot \text{Normal}(\theta \mid \mu_n, \tau_n^2)\, d\theta$$

$$= \text{Normal}(\tilde{y} \mid \mu_n, \tau_n^2 + \sigma^2)$$

prior predictive is also normal.

Example 5.2. (Midge wing length) We are given a data set on the wing length in millimeters of nine members of a species of midge (small, two-winged flies). From these nine measurements we wish to make inference about the population mean θ.

From previous studies, the wing lengths are typically around 1.9mm, so we set $\mu_0 = 1.9$. We also know that the wing length are positive-valued, but since we are using a normal prior, we need to set for τ_0 so that most of the mass is concentrated on the positive values. Conservatively, we set $\mu_0 - 2\tau_0 > 0$, so $\tau_0 < 1.9/2 = 0.95$.

The observations are: $\{1.64, 1.70, 1.72, 1.74, 1.82, 1.82.1.82, 1.90, 2.08\}$, giving $\bar{y} = 1.804$. Using the above formulas for posterior computation,

$$\begin{aligned} \mu_n &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1.11 \times 1.9 + \frac{9}{\sigma^2} \times 1.804}{1.11 + \frac{9}{\sigma^2}}, \\ \tau_n &= \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{1.11 + \frac{9}{\sigma^2}}. \end{aligned}$$

If we set $\sigma^2 := s^2 = 0.017$, then posterior distribution $\theta|y_1, \ldots, y_n, \sigma^2 = 0.017 \sim \text{Normal}(1.805, 0.002)$. A 95% quantile-based ~~confidence~~ credible interval for θ according to this posterior distribution is (1.72,1.89). Of course, this result is based on a point estimate of $\sigma^2 := s^2$ which is in fact only a rough estimate based on only nine observations. Next section we will study techniques for properly handling unknown variance.

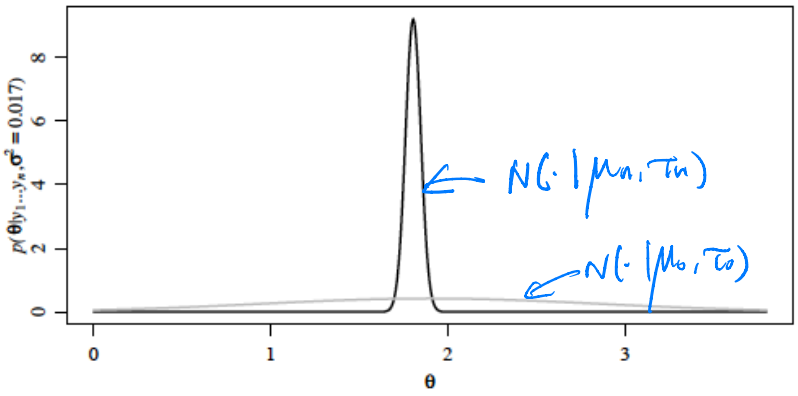


Figure 5.1: Prior and posterior ditributions for the population mean wing lengh.

5.3 Joint inference for the mean and variance

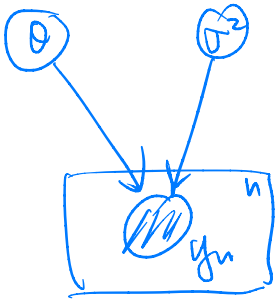
We need to specify a prior distribution on θ and σ^2. By Bayes' rule

$$\begin{aligned} p(\theta, \sigma^2|y_1, \ldots, y_n) &\propto p(\theta, \sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2) \\ &\propto p(\theta, \sigma^2)\frac{1}{\sigma^n}e^{-\frac{1}{2\sigma^2}\sum(\theta - y_i)^2}. \end{aligned} \tag{6}$$

It is not immediately obvious how to come up with a conjugate prior jointly for θ and σ^2. In the previous section, σ^2 is assumed to be fixed — from there it is simple to find that a normal prior for θ yields a normal posterior, *conditionally* on σ. This suggests that we may wish to set $\theta|\sigma^2 \sim \text{Normal}(\mu_0, \tau_0^2)$, for some suitable choice of μ_0, τ_0 which may be dependent on σ^2. This suggests a prior according to which θ and σ^2 may be coupled (i.e., dependent). The question is how. Moreover, this still does not tell us how to place a suitable prior on σ^2, since we still need to specify the joint prior distribution

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

Normal model



$\tilde{\sigma}^2 = \frac{1}{\sigma^2}$

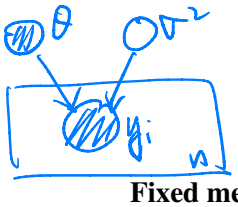
For σ^2 fixed / known

conjugate prior

$\theta|\sigma^2 \sim N(\mu_0, \tilde{\tau}_0^2)$

Bayes $\theta|y_1, \ldots, y_n, \sigma^2 \sim N(\mu_n, \tilde{\tau}_n^2)$

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



Fixed mean, varying variance To get a sense of what the form for a conjugate prior of σ^2 may be, let us take a step back, by assuming that θ is fixed.

Simplifying from (6)

$$\begin{aligned} p(\sigma^2|\theta, y_1, \ldots, y_n) &\propto p(\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2) \\ &\propto p(\sigma^2)\frac{1}{\sigma^n}e^{-\frac{1}{2\sigma^2}\sum(\theta-y_i)^2}. \end{aligned} \tag{7}$$

It is more convenient to look at the posterior pdf in terms of the precision $\tilde{\sigma}^2 = 1/\sigma^2$, we see that the simplest form for a conjugate prior for $\tilde{\sigma}$ will be one of the form $\tilde{\sigma}^{c_1}e^{-c_2\tilde{\sigma}^2}$. This gives us a Gamma prior for the precision parameter. In particular, we set

$$\tilde{\sigma}^2 \sim \text{Gamma}(a, b)$$

This is equivalent to saying that $\sigma^2 \sim \text{InvGamma}(a, b)$, and can be taken as a definition of the Inverse Gamma distribution.

Still assume θ is fixed. but σ^2 is random & latent.

what is the posterior predictive distribution of $\tilde{y}$?

$\sigma^2 \sim$ Inv Gamma (a, b)

$$p(\tilde{y} \mid y_1 \cdots y_n; a, b, \theta) = \int p(\tilde{y}, \sigma^2 \mid y_1 \cdots y_n; a, b, \theta)\, d\sigma^2$$

fixed parameters

$$= \int p(\tilde{y} \mid \sigma^2, y_1 \cdots y_n; a, b, \theta)\, p(\sigma^2 \mid y_1 \cdots y_n; a, b, \theta)\, d\sigma^2$$

$$= \int p(\tilde{y} \mid \sigma^2; a, b, \theta)\, p(\sigma^2 \mid y_1 \cdots y_n; a, b, \theta)\, d\sigma^2$$

Normal — Inv Gamma

Recall the Gamma pdf: $p(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}$, for $y > 0$. Let $z = 1/y$, so that $y = 1/z$ and $dy/dz = -1/z^2$. By the change of variable formula,

$$p(z|a,b) = p(y(z)|a,b)|dy/dz| = \frac{b^a}{\Gamma(a)} y(z)^{a-1} e^{-by(z)} (1/z^2) = \frac{b^a}{\Gamma(a)} z^{-a-1} e^{-b/z},$$

which gives the pdf for InvGamma(a,b).

Now, combining the inverse-gamma prior for σ^2 with the normal likelihood, we find that

$$\begin{aligned} p(\sigma^2|a,b,\theta,y_1,\ldots,y_n) &\propto p(\sigma^2) \times \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum(\theta-y_i)^2} \\ &\propto (\sigma^2)^{-a-1} e^{-b/\sigma^2} \times \sigma^{-n} e^{-\frac{1}{2\sigma^2}\sum(\theta-y_i)^2} \\ &\propto (\sigma^2)^{-(a+n/2)-1} e^{-(b+\frac{1}{2}\sum(\theta-y_i)^2)/\sigma^2} \\ &= \text{InvGamma}(a+\frac{n}{2}, b+\frac{1}{2}\sum(\theta-y_i)^2) \\ &=: \text{InvGamma}(a_n, b_n). \end{aligned} \tag{8}$$

We proceed to finding the predictive distribution. Note that this can be viewed as a mixture distribution of Gaussians, with the location fixed, and the precision parameter varying according to the Gamma distribution Gamma(a_n, b_n). We also note that the representation in the precision parameter is more convenient because it allows us to directly utilize the relevant identity that arise from Gamma pdf's normalizing constant. Thus, in what's followed we may switch back and forth between the two representations, in terms of $\tilde{\sigma}^2$ and σ^2.

$$
\begin{aligned}
p(\tilde{y}|a, b, y_1, \ldots, y_n) &= \int p(\tilde{y}|\theta, \tilde{\sigma}^2) \times p(\tilde{\sigma}^2|a, b, \theta, y_1, \ldots, y_n) \mathrm{d}\tilde{\sigma}^2. \\
&= \int \left[\left(\frac{\tilde{\sigma}^2}{2\pi} \right)^{1/2} \exp -\frac{\tilde{\sigma}^2}{2}(\tilde{y} - \theta)^2 \right] \times \left[\frac{b_n^{a_n}}{\Gamma(a_n)} (\tilde{\sigma}^2)^{a_n - 1} e^{-b_n \tilde{\sigma}^2} \right] \mathrm{d}\tilde{\sigma}^2 \\
&= \frac{b_n^{a_n}}{\Gamma(a_n)} \frac{1}{(2\pi)^{1/2}} \frac{\Gamma(a_n + 1/2)}{(b_n + (\tilde{y} - \theta)^2/2)^{a_n + 1/2}} \\
&= \frac{\Gamma(a_n + 1/2)}{\Gamma(a_n)} \frac{1}{(2\pi b_n)^{1/2} (1 + (\tilde{y} - \theta)^2/(2b_n))^{a_n + 1/2}}. \qquad (9)
\end{aligned}
$$

We arrive at the well-known Student's t distribution, which has three parameters, location parameter θ, scale parameter b_n/a_n and $2a_n$ degrees of freedom. The variance of the predictive distribution is, provided $2a_n > 2$,

$$(b_n/a_n)\frac{2a_n}{2a_n - 2} = b_n/(a_n - 1).$$

It is interesting to note that the predictive distribution of the data becomes heavier tailed than the normal sampling model (inverse squared tail vs inverse exponential tail), thanks to the uncertainty about the variance/precision parameter that is integrated out.

$p(\theta, \sigma^2 | \text{Data}) \propto p(\theta, \sigma^2) * p(y_1, \dots, y_n | \theta, \sigma^2)$
joint prior

Both mean and variance parameter varying Now we are ready to handle the case both θ and σ^2 vary. As we have seen in the previous pages, it may be more convenient in our derivation to work with the precision parameter $\tilde{\sigma}^2$ instead.

It is tempting to place independent prior distributions on θ and $\tilde{\sigma}^2$: say a normal prior on θ and independently, a Gamma prior on $\tilde{\sigma}^2$. The reader can verify without difficulty that this won't give us a conjugate prior because the posterior for either θ or $\tilde{\sigma}^2$ will not be normal or Gamma, respectively. (What would be the form of the posteriors then?)

The issue is that conditionally given the observations $y_1, \dots, y_n$, parametes θ and $\tilde{\sigma}^2$ are dependent even if they are independent a priori. So we need to construct a prior distribution according to which θ and $\tilde{\sigma}^2$ are dependent to begin with. Here is how: use the decomposition

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

$p(\theta, \tilde{\sigma}^2 | \text{Data}) = p(\tilde{\sigma}^2 | \text{Data}) * p(\theta | \tilde{\sigma}^2, \text{Data})$

and set the prior as

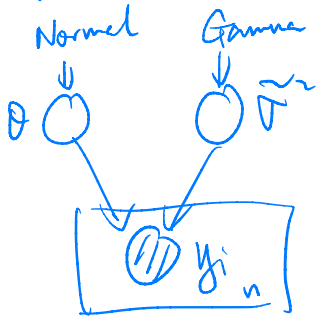
$$\begin{aligned} \tilde{\sigma}^2 &\sim \text{Gamma}(a, b) \\ \theta | \tilde{\sigma}^2 &\sim \text{Normal}(\mu_0, \kappa_0 \tilde{\sigma}^2). \end{aligned}$$

precision.

$\text{Variance} = \frac{1}{\kappa_0 \tilde{\sigma}^2} = \frac{1}{\kappa_0} \sigma^2$

The key is in the second line, which allows the coupling of θ and $\tilde{\sigma}^2$ via the conditional prior's variance.

why not independent prior?



suppose independent prior for θ & σ^2

$p(\theta, \sigma^2) \sim p(\theta) * p(\sigma^2)$ $\quad \theta \perp \sigma^2$

$\sim \text{Normal}(\theta | \dots) * \text{Inv Gamma}(\sigma^2 | \dots)$

$p(\theta | \tilde{\sigma}^2, \text{Data}) = \text{Normal}$ conjugate

$\tilde{\sigma}^2$ Random

$\Rightarrow p(\theta | \text{Data}) \neq \text{Normal}$

$p(\theta, \sigma^2 | \text{Data})$ if conjugacy, then we need $\theta \perp \sigma^2$ in posterior $p(\theta, \sigma^2 | \text{Data})$

$\neq p(\theta | \text{Data}) \cdot p(\sigma^2 | \text{Data})$ — so not conjugate

even if $\theta \perp \sigma^2$ a priori

$\theta \not\perp \sigma^2 | y_1, \dots, y_n$

$\tilde{\sigma}^2 \sim \text{Gamma}(a, b)$

For ease of interpretation later, we set $a = \nu_0/2, b = \nu_0\sigma_0^2/2$ (which gives the prior expectation for $\tilde{\sigma}^2$ to equal $a/b = 1/\sigma_0^2 =: \tilde{\sigma}_0^2$).

The sampling/likelihood model is the same as before:

$$Y_1, \ldots, Y_n | \theta, \tilde{\sigma}^2 \overset{iid}{\sim} \text{Normal}(\theta, \tilde{\sigma}^2).$$

Now we verify that the specified prior is indeed conjugate. Decompose the posterior distribution similarly:

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

From the previous section, we already have

$$\boxed{\theta | y_1, \ldots, y_n, \tilde{\sigma}^2 \sim \text{Normal}(\mu_n, \tilde{\tau}_n^2)}$$

Recall $\theta | \tilde{\sigma}^2 \sim \text{Normal}(\mu_0, \kappa_0 \tilde{\sigma}^2)$

where

$$\begin{aligned} \tilde{\tau}_n^2 &= \kappa_0 \tilde{\sigma}^2 + n\tilde{\sigma}^2 =: \kappa_n \tilde{\sigma}^2 \\ \mu_n &= \frac{\kappa_0 \tilde{\sigma}^2 \mu_0 + (n\tilde{\sigma}^2)\bar{y}}{\kappa_0 \tilde{\sigma}^2 + n\tilde{\sigma}^2} = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n}. \end{aligned}$$

Same calculation as fixed variance & unknown mean

(1)

In short, the conditional posterior of θ, namely $p(\theta | \tilde{\sigma}^2, y_1, \ldots, y_n)$, has the same form as that of the conditional prior $p(\theta | \tilde{\sigma}^2)$.

(2) also the posterior of $\tilde{\sigma}^2$, $p(\tilde{\sigma}^2 | y_1, \ldots, y_n)$ also has the same form as that of prior $p(\tilde{\sigma}^2)$

→ next page

⇒ joint posterior of $\theta, \tilde{\sigma}^2$ has the same form as that of prior

Recall marginal prior for $\tilde{\sigma}^2$ is

$\tilde{\sigma}^2 \sim \text{Gamma}(a, b) \equiv \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$

Next, we check the marginal posterior of $\tilde{\sigma}^2$. For this computation, we need to integrate out θ (unlike the previous detour where θ is fixed, and $\tilde{\sigma}^2$ varies).

previous section $p(\tilde{\sigma}^2|\theta, y_1, \ldots, y_n)$

marginal posterior

$$\begin{aligned}
p(\tilde{\sigma}^2|y_1,\ldots,y_n) &\propto p(\tilde{\sigma}^2)p(y_1,\ldots,y_n|\tilde{\sigma}^2) \\
&\propto p(\tilde{\sigma}^2)\int p(y_1,\ldots,y_n|\theta,\tilde{\sigma}^2)p(\theta|\tilde{\sigma}^2)\mathrm{d}\theta \\
&\propto (\tilde{\sigma}^2)^{a-1}e^{-b\tilde{\sigma}^2}\int (\tilde{\sigma}^2)^{n/2}e^{-\frac{1}{2}\tilde{\sigma}^2\sum(\theta-y_i)^2}(\kappa_0\tilde{\sigma}^2)^{1/2}e^{-\frac{1}{2}\kappa_0\tilde{\sigma}^2(\theta-\mu_0)^2}\mathrm{d}\theta \\
&\propto (\tilde{\sigma}^2)^{a+n/2-1}e^{-b\tilde{\sigma}^2}(\kappa_0\tilde{\sigma}^2)^{1/2}\int e^{-\frac{1}{2}\tilde{\sigma}^2[\sum(\theta-y_i)^2+\kappa_0(\theta-\mu_0)^2]}\mathrm{d}\theta.
\end{aligned}$$

We quickly see that in the integrand the form of a Gaussian pdf, so the integral can be simplified by utilizing the formula for normalizing the Gaussian pdf: $\int e^{-\frac{1}{2}\tilde{\sigma}^2(y-\mu)^2}\mathrm{d}y = \sqrt{2\pi/\tilde{\sigma}^2}$. Accordingly, the integral is precisely

$$\begin{aligned}
&\sqrt{2\pi/[(\kappa_0+n)\tilde{\sigma}^2]}\exp\Big\{-\frac{1}{2}\tilde{\sigma}^2\Big\{-\frac{(\mu_0\kappa_0+n\bar{y})^2}{\kappa_0+n}+\sum y_i^2+\kappa_0\mu_0^2\Big\}\Big\} \\
= \;&\sqrt{2\pi/[(\kappa_0+n)\tilde{\sigma}^2]}\exp\Big\{-\frac{1}{2}\tilde{\sigma}^2\Big\{\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}+\sum y_i^2-n\bar{y}^2\Big\}\Big\}.
\end{aligned}$$

Plugging back for the posterior of $\tilde{\sigma}^2$, keeping only relevant terms

$$\begin{aligned}
p(\tilde{\sigma}^2|y_1,\ldots,y_n) &\propto (\tilde{\sigma}^2)^{a+n/2-1}e^{-b\tilde{\sigma}^2}\exp\Big\{-\frac{1}{2}\tilde{\sigma}^2\Big\{\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}+(n-1)s^2\Big\}\Big\} \\
&= \text{Gamma}\Big(a+\frac{1}{2}n, b+\frac{1}{2}\Big\{\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}+(n-1)s^2\Big\}\Big) \\
&= \text{Gamma}\Big(\nu_0/2+n/2, (1/2)\Big\{\nu_0\sigma_0^2+\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}+(n-1)s^2\Big\}\Big) \\
&=: \boxed{\text{Gamma}(\nu_n/2, \nu_n\sigma_n^2/2)},
\end{aligned}$$

prior; data

where the posterior distribution's parameters take the form

$$\begin{aligned}
\nu_n &= \nu_0+n \\
\sigma_n^2 &= \frac{1}{\nu_n}\Big\{\nu_0\sigma_0^2+\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}+(n-1)s^2\Big\}.
\end{aligned}$$

$\nu_n\sigma_n^2$

How to make sense of the contribution of the prior information and the data in these expressions? The posterior mean of $\tilde{\sigma}^2$ is $1/\sigma_n^2$, while the posterior variance is of the order $1/\nu_n\sigma_n^4$. In the above formula for $\nu_n\sigma_n^2$, it is clear that $\nu_0\sigma_0^2$ represents the information from the prior for σ^2. The term $(n-1)s^2$ represents the variability of the observed data from the sample mean.

Finally, the middle term $\frac{\kappa_0 n(\mu_0-\bar{y})^2}{\kappa_0+n}$ represents the contribution to the variance parameter σ^2 due to the coupling between the location parameter θ and precision $\tilde{\sigma}^2$ according to the conditional prior

$$\theta|\tilde{\sigma}^2 \sim \text{Normal}(\mu_0, \kappa_0\tilde{\sigma}^2).$$

According to this prior, θ is drawn from a mixture of normal distributions centering on μ_0 with varying precision proportional to κ_0. This seems to be relatively strong opinion for a prior specification, which entails the "biased" contribution of the middle term that increases with both κ_0 and the variability of sample mean about μ_0 toward the estimate of the variance σ^2.

One may harshly criticize the prior due to the implication discussed as being too strong. We don't necessarily defend this at all cost: after all we have arrived at this prior construction mainly from a mathematical/ computational viewpoint, i.e., to obtain a conjugate prior. So, there's a bias — the incurred bias is a cost one has to pay for the mathematical/ computational convenience. Whether it is worth it depends on the modeler and the data at hand. Note that when the sample size is large, the bias incurred by our prior construction will be washed away by the last term $(n-1)s^2$, which is purely driven by the data set.

sampling model: $y_1, \dots, y_n \mid \theta, \sigma \sim N(\cdot \mid \theta, \sigma^2) \equiv N(\cdot \mid \theta, \tilde{\sigma}^2)$ ← precision

conjugate prior jointly from $(\theta, \tilde{\sigma}^2)$:

$$\begin{cases} \tilde{\sigma}^2 \sim \text{Gamma}\left(a = \nu_0/2,\ b = \nu_0\sigma_0^2/2\right) \\ \theta \mid \tilde{\sigma}^2 \sim \text{Normal}\left(\mu_0,\ \kappa_0\tilde{\sigma}^2\right) \end{cases}$$

↑ dependent on $\tilde{\sigma}^2$

Apply Bayes rule

$$p(\theta, \tilde{\sigma}^2 \mid \text{Data}) = p(\tilde{\sigma}^2 \mid \text{Data}) * p(\theta \mid \tilde{\sigma}^2, \text{Data})$$

where $\tilde{\sigma}^2 \mid \text{Data} \sim \text{Gamma}\left(\nu_n/2,\ \nu_n\sigma_n^2/2\right)$

$$\theta \mid \tilde{\sigma}^2, \text{Data} \sim \text{Normal}\left(\mu_n,\ \kappa_n\tilde{\sigma}^2\right)$$

$$\begin{cases} \kappa_n = \kappa_0 + n \\ \mu_n = \dfrac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \end{cases}$$

$$\begin{cases} \nu_n = \gamma_0 + n \\ \nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \dfrac{\kappa_0 n(\mu_0 - \bar{y})^2}{\kappa_0 + n} + (n-1)s^2 \end{cases}$$

↑ marginal prior ↑ conditional prior ↑ data

Example 5.3. (Midge wing length — continued). Our sampling model for midge wing lengths is $Y|\theta, \tilde{\sigma} \sim \text{Normal}(\theta, \tilde{\sigma}^2)$ and we will place a joint prior on $\theta, \tilde{\sigma}$ via

$$\begin{aligned} \tilde{\sigma}^2 &\sim \text{Gamma}(a = \nu_0/2, b = \nu_0\sigma_0^2/2) \\ \theta|\tilde{\sigma}^2 &\sim \text{Normal}(\mu_0, \kappa_0\tilde{\sigma}^2). \end{aligned}$$

Previous studies suggest that the true mean and standard deviation should not be too far from 1.9 mm and 0.1 mm, respectively. So we may set $\mu_0 = 1.9$ and $\sigma_0^2 = 0.01$.

The Gamma prior implies the prior mean for the precision is $a/b = 1/\sigma_0^2 = \tilde{\sigma}_0^2 = 100$, and prior variance for the precision is $a/b^2 = 2/(\nu_0\sigma_0^4)$. We set $\nu_0 = 1$ to allow for a reasonably large variance.

As for κ_0: we also set $\kappa_0 = 1$. Since $\tilde{\sigma}$ is a priori distributed over a large range of values, this implies that we assume θ to be only weakly coupled to $\tilde{\sigma}^2$.

From the sample, $\bar{y} = 1.804$ and $s^2 = 0.0169$. Applying to the posterior computation derived earlier:

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n} = \frac{1.9 + 9 \times 1.804}{1 + 9} = 1.814 \\ \sigma_n^2 &= \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + \frac{\kappa_0 n(\mu_0 - \bar{y})^2}{\kappa_0 + n} + (n-1)s^2\right\} = \frac{0.010 + 0.008 + 0.135}{10} = 0.015. \end{aligned}$$

Compared to the point estimate presented earlier, the posterior mean for θ is comparable, but the uncertainty captured by σ_n^2 is considerably larger. But we can say much more.

In particular, the joint posterior distribution is given by

$$\begin{aligned} \theta|y_1, \ldots, y_n, \sigma^2 &\sim \text{Normal}(\mu_n = 1.814, \tilde{\tau}_n^2 = \kappa_n\tilde{\sigma}^2 = 10\tilde{\sigma}^2) \\ \tilde{\sigma}^2|y_1, \ldots, y_n &\sim \text{Gamma}(\nu_n/2 = 10/2, \nu_n\sigma_n^2/2 = 10 \times 0.015/2). \end{aligned}$$

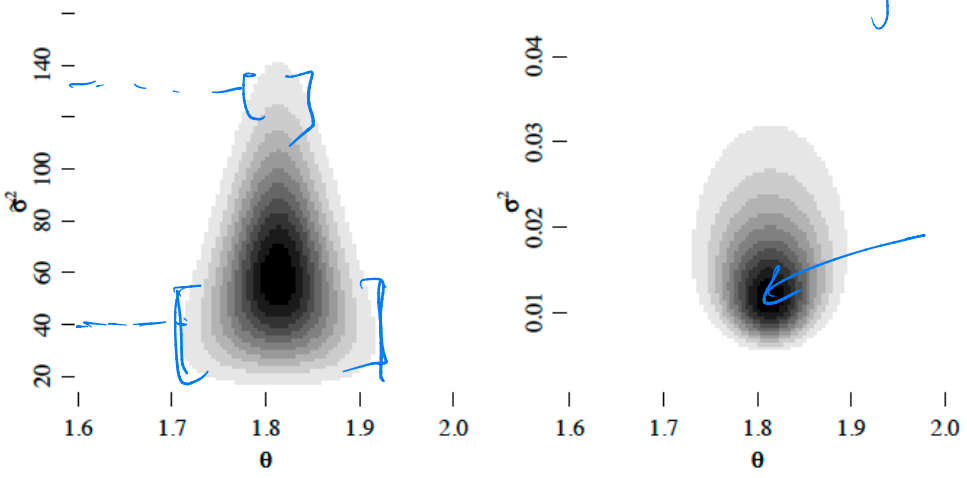


Figure 5.2: Joint posterior distributions for $(\theta, \tilde{\sigma}^2)$ and (θ, σ^2).

These plots were obtained by computing the joint pdf at pairs of values of $(\theta, \tilde{\sigma}^2)$ and (θ, σ^2) on a grid. Note also that samples from this posterior can be easily obtained via Monte Carlo sampling.

These plots tell us about where most of the mass of the posterior for (θ, σ^2) is, and to some extent the relationship between the two parameters. When $\tilde{\sigma}^2$ is small (σ^2 is large) there are more uncertainties about θ. Moreover, the contours are more peaked as function of θ for low values of σ^2 than high values.

Hyperparameters and improper priors Hyperparamerers are parameters specified for the prior distributions. In our previous example, two of them are κ_0 and ν_0. They may be regarded as the prior sample size, because according to the Bayesian update

$$\kappa_n = \kappa_0 + n$$
$$\nu_n = \nu_0 + n.$$

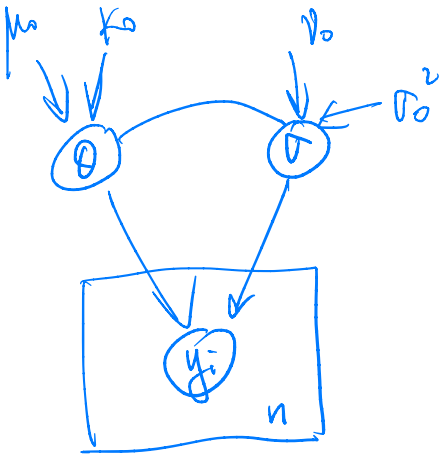
cf. Beta-Bernoulli model

When κ_0 and ν_0 are relatively small compared to n, the effects of these hyperparameters are negligible. Of interest is when n is itself quite small. Is this still possible to have a prior specification whose impact relative to the impact from the data is minimal?

The smaller ν_0 is, the "flatter" the marginal prior distribution for $\tilde{\sigma}$; the smaller κ_0 and ν_0 are, the flatter the marginal prior distribution for θ. (Recall our earlier computation in Eq. (9) that a mixture of fixed-mean normal distributions with a Gamma mixing on the precision is a Student's t distribution). In other words, the priors can be viewed as "less discriminative"; and hence "more objective".

Weak priors

let $\kappa_0 \to 0$, $\nu_0 \to 0$



Let us perform the formal computation, by letting $\kappa_0, \nu_0 \to 0$

correspondence to frequentist estimates

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n} \to \bar{y} \\ \sigma_n^2 &= \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + \frac{\kappa_0 n(\mu_0 - \bar{y})^2}{\kappa_0 + n} + (n-1)s^2\right\} \to \frac{n-1}{n}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2. \end{aligned}$$

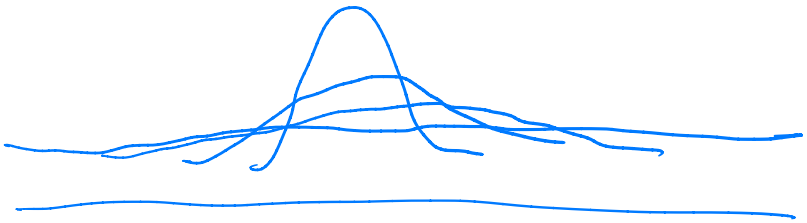
This leads to the following "posterior distribution", which is free of hyperparameters:

$$\begin{aligned} \tilde{\sigma}^2|y_1, \ldots, y_n &\sim \text{Gamma}(n/2, (n/2)\frac{1}{n}\sum(y_i - \bar{y})^2) \\ \theta|\tilde{\sigma}^2, y_1, \ldots, y_n &\sim \text{Normal}(\bar{y}, \frac{1}{n}\sigma^2). \end{aligned} \tag{10}$$

There does not exist a valid prior distribution for the above "posterior distribution", which appears only as the limit of a sequence of posterior distributions that arise from the sequence of prior distributions according to which $\kappa_0, \nu_0 \to 0$. If one still wish to employ such posterior distribution, one need to utilize a notion of improper prior distribution. In other words, we may still speak of a posterior distribution, even though the corresponding prior is not a proper probability distribution.

prior: $N(\mu_0, \infty)$ ← invalid

$N(\mu_0, \sigma_0^2)$ $\sigma_0^2 \to \infty$



$\iint \frac{1}{\sigma^2} d\theta d\sigma^2 = +\infty$

Consider function $\tilde{p}(\theta, \sigma^2) = 1/\sigma^2$. This is not a proper distribution because it is not integrable over (θ, σ^2). Thus we will treat this as an *improper* prior distribution and apply the Bayes' rule to obtain:

$$p(\theta, \sigma^2|y) \propto p(y|\theta, \sigma^2) \times \tilde{p}(\theta, \sigma^2).$$

$p(\theta, \sigma^2|y) = \frac{p(y|\theta,\sigma^2) \cdot \tilde{p}(\theta,\sigma^2)}{\iint p(y|\theta,\sigma^2) \tilde{p}(\theta,\sigma^2) d\theta d\sigma^2}$

Then we have a valid distribution over (θ, σ^2): in fact, it can be easily verified that the induced marginal θ is the same as that of (10), while the marginal for $\tilde{\sigma}^2$ is Gamma$((n-1)/2, (1/2)\sum(y_i - \bar{y})^2$. In addition, integrating over $\tilde{\sigma}^2$, following a computation similar to Eq. (9) we find that

$$\left.\frac{\theta - \bar{y}}{s/\sqrt{n}}\right| y_1, \ldots, y_n \sim t_{n-1}. \tag{11}$$

Bayesian statement

everything can still follow through even if the prior is improper (aka not integrable)

Remark 5.1. Some remarks.

(i) The use of improper priors is not considered to be truly Bayesian, but it can be justified (informally) by the limiting argument presented above, and formally via a decision-theoretic framework. It is one area where one can find the meeting points between Bayesian and frequentist approaches.

(ii) It is interesting to compare with the sampling distribution of the t statistic, conditional on θ but unconditional on the data:

$$\left.\frac{\bar{Y}-\theta}{s/\sqrt{n}}\right|\theta \sim t_{n-1}. \qquad \text{frequentist statement} \tag{12}$$

Eq. (12) is a statement about the data: it says that *before* we sample the data, our uncertainty about the scaled deviation of the sample mean $\bar{Y}$ from the population mean θ has a t_{n-1} distribution. Eq. (11) says that *after* we sample the data, our uncertainty is still represented with a t_{n-1} distribution, except that it is our uncertainty about θ given the information provided by the data $\bar{y}$.

5.4 Normal model for non-normal data

People apply normal models to non-normal data all the time. In this section, we have seen examples of modeling heights for a human population and modeling flies' wing length. In both cases, the data are positive valued, whereas normal distributions are supported on the entire real line. However, the quantity of interest is the population mean, which can be treated as approximately normally distributed according to the central limit theorem.

As another example, consider the number of children for a group of women over age 40, and consider estimating the mean number of children for this population, based on the samples $Y_1, \ldots, Y_n$. In the previous section, we considered a Poisson sampling model, which is motivated by the fact that Y_i are integer-valued. Obviously it makes no sense to assume $Y_i|\theta, \sigma^2 \sim \text{Normal}(\theta, \sigma^2)$. However, it is still reasonable to assume that the population mean θ is normally distributed (a priori).

By the CLT, we know that

$$p(\bar{y}|\theta, \sigma^2) \approx \text{Normal}(\bar{y}|\theta, \sqrt{\sigma^2/n}),$$

where σ^2 denotes the population variance, with the approximation becoming increasingly accurate as n gets larger.

If σ^2 is known, then we may consider placing a normal prior on θ and obtain the posterior for θ via

$$p(\theta|\bar{y}, \sigma^2) \propto p(\theta) \times p(\bar{y}|\theta, \sigma^2).$$

If σ^2 is unknown, we may consider to bring in the point estimate s^2 and *conditioning* on it:

$$p(\theta, \sigma^2|\bar{y}, s^2) \propto p(\theta, \sigma^2) \times p(\bar{y}, s^2|\theta, \sigma^2).$$

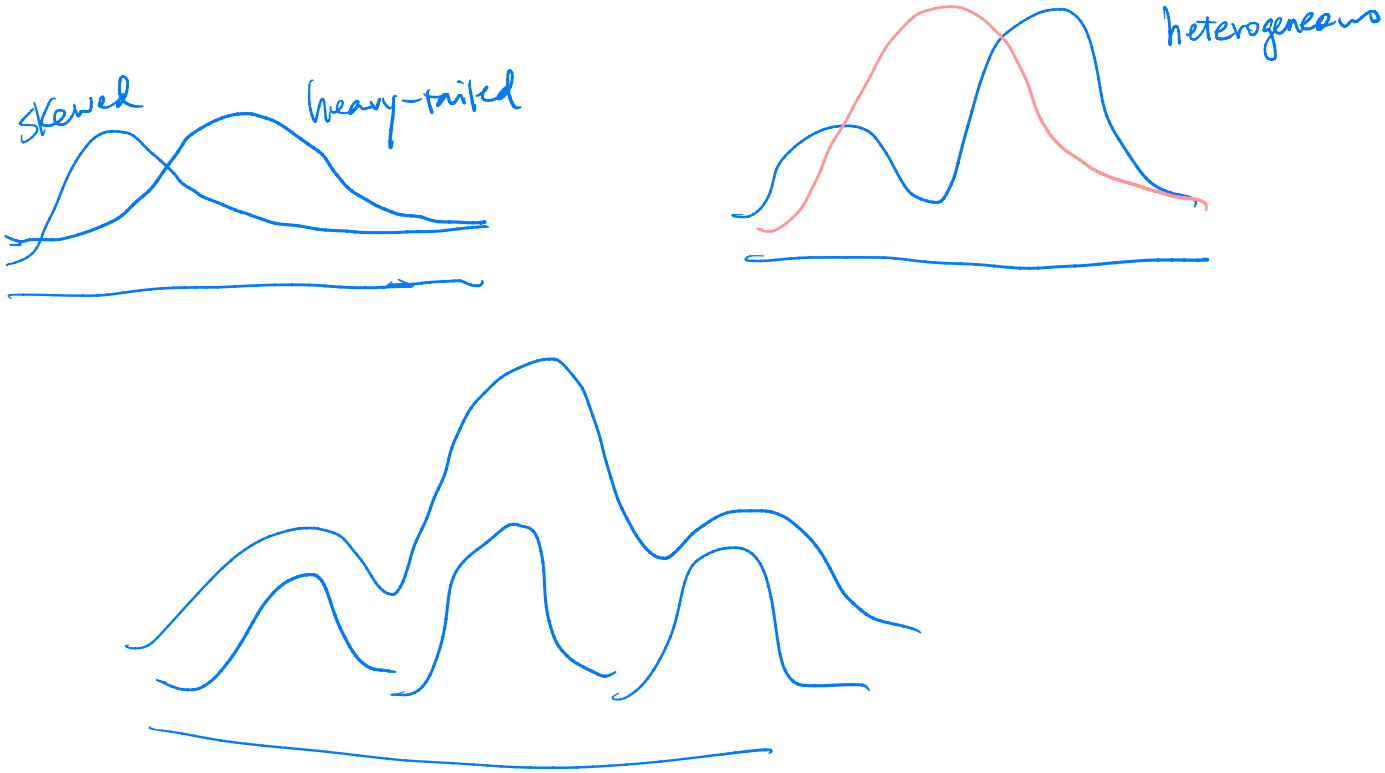
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The likelihood term $p(\bar{y}, s^2|\theta, \sigma^2)$ may be approximated by applying a normal sampling model $p(\bar{y}|\theta, \sigma^2)$ for $\bar{y}$ and Gamma sampling model $p(s^2|\bar{y}, \theta, \sigma)$ for s^2, conditionally on $\bar{y}$.

Hence, we have seen that when the sample size is reasonably large, the above approximation treatment is quite reasonable and can lead to good practical results.

When are normal models not appropriate?

- when the quantity of interest is not about the population mean and/or variance but requires tail behavior of the population, while the population's distribution is clearly not normal (e.g., heavy-tailed or skewed distributions). For instance, we may be interested in the group of people with large number of children.
- when the population is highly heterogeneous and we are interested in learning about such heterogeneity. For instance, the population's distribution may be multi-modal, and so it makes more sense to represent it as a mixture of sub-populations each of which have their own parameters of interest. One is not interested in the population mean as much as the parameters of each sub-population.
- even when the normal model is not appropriate, normal distributions frequently serve as an useful building block: recall that heavier-tailed distributions such as t-distribution can be viewed as a mixture of normals with variance parameters varying, while multi-modal distributions can be approximated by a mixture of normal distributions with mean parameters or both type of parameters varying.



The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.