Chapter 4: Univariate Normal Model

Contents

1 The normal distribution

The normal distribution $\mathcal{N}(\mu,\sigma^2)$ (sometimes called the Gaussian distribution) with mean $\mu\in\mathbb{R}$ and variance $\sigma^2>0$ (standard deviation $\sigma=\sqrt{\sigma^2}$) has p.d.f.

$$\mathcal{N}(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big)$$

for $x\in\mathbb{R}$. In Bayesian calculations, it is often more convenient to write the p.d.f. in terms of the precision, or inverse variance, $\lambda=1/\sigma^2$ rather than the variance. In this parametrization, the p.d.f. is

$$\mathcal{N}(x\mid\mu,\lambda^{-1})=\sqrt{\frac{\lambda}{2\pi}}\exp\big(-\tfrac{1}{2}\lambda(x-\mu)^2\big)$$

since $\sigma^2=1/\lambda=\lambda^{-1}$.

The normal distribution has certain special properties that give it a unique position in probability and statistics. Foremost among these is the central limit theorem (CLT), which states that the sum of a large number of independent random variables tends to be approximately normally distributed. The CLT explains why real-world data so often appears approximately normal, and from a modeling perspective, it helps us to understand when a normal model would be appropriate.

Jeffrey W. Miller (2015). *Lecture Notes on Bayesian Statistics.* Duke University, Durham, NC.

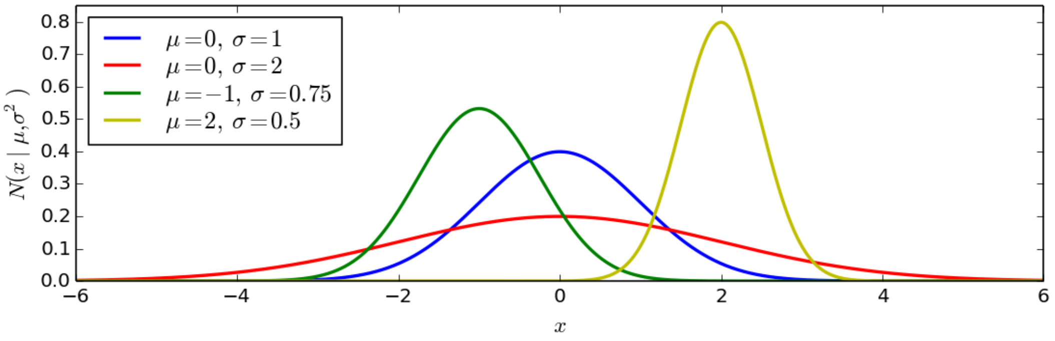

Figure 1: Normal distribution with various choices of μ and σ.

Many real-world quantities tend to be normally distributed—for instance, human heights and other body measurements, cumulative hydrologic measures such as annual rainfall or monthly river discharge, errors in astronomical or physical observations, and diffusion of a substance in a liquid or gas. Some things are products of many independent variables (rather than sums), and in such cases the logarithm will be approximately normal since it is a sum of many independent variables—this is often the case for economic quantities such as stock market indices, due to the effect of compound interest.

Basic properties of $\mathcal{N}(\mu, \sigma^2)$

- Mean, median, and mode are all the same (μ)
- Symmetric about the mean
- 95% probability within $\pm 1.96\sigma$ of the mean (roughly, $\pm 2\sigma$)
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(m, s^2)$ independently, then

$$aX + bY \sim \mathcal{N}(a\mu + bm,\ a^2\sigma^2 + b^2 s^2). \tag{1.1}$$

- Careful: `rnorm`, `dnorm`, `pnorm`, and `qnorm` in `R` take the mean and standard deviation σ as arguments (not mean and variance σ^2). For example, `rnorm(n,m,s)` generates n normal random variables from $\mathcal{N}(m, s^2)$.

The normal distribution is also special due to its analytic tractability—inference for complex models constructed by combining normal distributions can often be done analytically. This makes it especially convenient to work with from a computational standpoint.

2 Conjugate prior for the mean

Suppose we are using an i.i.d. normal model with mean θ and precision λ:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \lambda^{-1}).$$

Assume the precision $\lambda = 1/\sigma^2$ is known and fixed, and θ is given a $\mathcal{N}(\mu_0, \lambda_0^{-1})$ prior:

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu_0, \lambda_0^{-1})$$

i.e., $p(\theta) = \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1})$. This is sometimes referred to as a ***Normal–Normal model***. It turns out that the posterior is

$$\boldsymbol{\theta}|x_{1:n} \sim \mathcal{N}(M, L^{-1}) \tag{2.1}$$

i.e., $p(\theta|x_{1:n}) = \mathcal{N}(\theta \mid M, L^{-1})$, where $L = \lambda_0 + n\lambda$ and

$$M = \frac{\lambda_0\mu_0 + \lambda \sum_{i=1}^n x_i}{\lambda_0 + n\lambda}.$$

Thus, the normal distribution is, itself, a conjugate prior for the mean of a normal distribution with known precision.

2.1 Derivation of the posterior

There are various ways of deriving Equation 2.1, with “completing the square” being perhaps the most common. Here, we take a slightly more streamlined approach. First, note that for any x and ℓ,

$$\begin{aligned}
\mathcal{N}(x \mid \theta, \ell^{-1}) &= \sqrt{\frac{\ell}{2\pi}} \exp\big(-\tfrac{1}{2}\ell(x-\theta)^2\big) \\
&\underset{\theta}{\propto} \exp\big(-\tfrac{1}{2}\ell(x^2 - 2x\theta + \theta^2)\big) \\
&\underset{\theta}{\propto} \exp\big(\ell x\theta - \tfrac{1}{2}\ell\theta^2\big).
\end{aligned} \tag{2.2}$$

Due to the symmetry of the normal p.d.f.,

$$\mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) = \mathcal{N}(\mu_0 \mid \theta, \lambda_0^{-1}) \underset{\theta}{\propto} \exp\big(\lambda_0\mu_0\theta - \tfrac{1}{2}\lambda_0\theta^2\big) \tag{2.3}$$

by Equation 2.2 with $x = \mu_0$ and $\ell = \lambda_0$. Therefore, defining L and M as above,

$$\begin{aligned}
p(\theta|x_{1:n}) &\propto p(\theta)p(x_{1:n}|\theta) \\
&= \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) \prod_{i=1}^n \mathcal{N}(x_i \mid \theta, \lambda^{-1}) \\
&\overset{\text{(a)}}{\propto} \exp\big(\lambda_0\mu_0\theta - \tfrac{1}{2}\lambda_0\theta^2\big) \exp\big(\lambda(\textstyle\sum x_i)\theta - \tfrac{1}{2}n\lambda\theta^2\big) \\
&= \exp\Big((\lambda_0\mu_0 + \lambda \textstyle\sum x_i)\theta - \tfrac{1}{2}(\lambda_0 + n\lambda)\theta^2\Big) \\
&= \exp(LM\theta - \tfrac{1}{2}L\theta^2) \\
&\overset{\text{(b)}}{\propto} \mathcal{N}(M \mid \theta, L^{-1}) = \mathcal{N}(\theta \mid M, L^{-1}),
\end{aligned}$$

where step (a) uses Equations 2.2 and 2.3, and step (b) uses Equation 2.2 with $x = M$ and $\ell = L$.

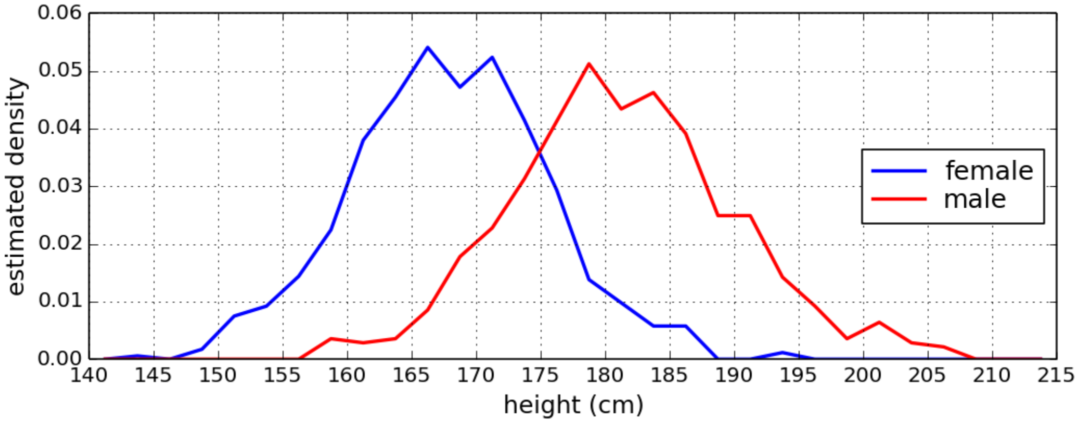

Figure 2: Heights of Dutch women and Dutch men.

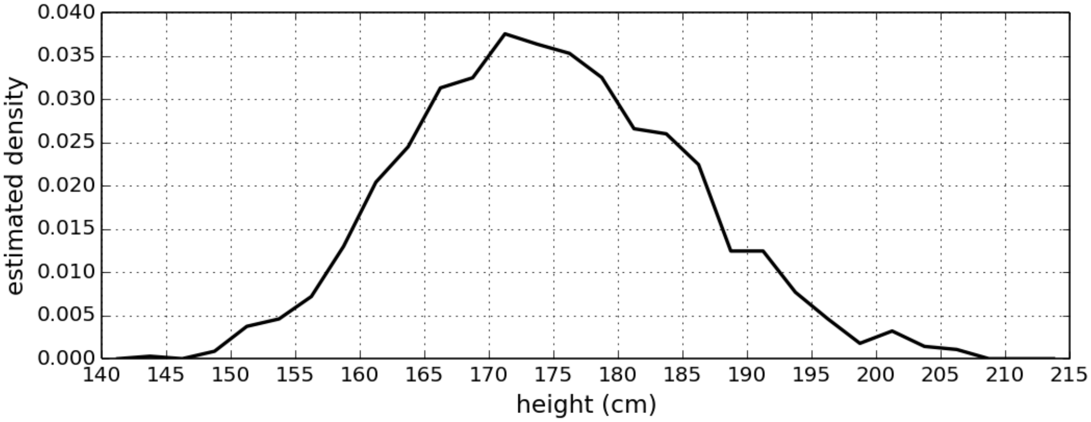

Figure 3: Heights of Dutch women and men, combined.

2.2 Example: Is human height bimodal?

The distribution of heights of adult humans—when separated according to sex (female or male)—is a classic example of a normal distribution. It seems that the reason why height tends to be normally distributed is because there are many independent genetic and environmental factors which contribute additively to overall height, and this leads to a normal distribution due to the central limit theorem. However, the combined distribution of heights (pooling females and males together) is not normal, and is often said to be bimodal—that is, having two modes (i.e., two maxima). But is it really bimodal?[1]

Figure 2 shows estimated densities of the heights of Dutch women and Dutch men, based on a sample of 695 women and 562 men (Krul et al., 2011)[2], and Figure 3 shows the estimated

[1] This example is inspired by Schilling et al. (2002).

[2] Data from the `selfreport` dataset in the `MICE` package for `R`.

density for women and men together, assuming there is an equal proportion of women and men in the population. At a glance, while the heights of women and men separately do appear to be roughly normally distributed, the combined distribution does not look bimodal. How could we test whether it is bimodal in a more precise way?

Let's assume female heights and male heights are each normally distributed. To keep things relatively simple, let's assume they have the same standard deviation, and also that there is an equal proportion of women and men in the population. Then, it is known that the combined distribution is bimodal if and only if the difference between the means is greater than twice the standard deviation (Helguerro, 1904).

Model

In mathematical notation: Assume the female heights are

$$X_1, \ldots, X_k \overset{\text{iid}}{\sim} \mathcal{N}(\theta_f, \sigma^2),$$

where $k = 695$, the male heights are

$$Y_1, \ldots, Y_\ell \overset{\text{iid}}{\sim} \mathcal{N}(\theta_m, \sigma^2),$$

where $\ell = 562$, and the p.d.f. of the combined distribution of heights is

$$\tfrac{1}{2}\mathcal{N}(x \mid \theta_f, \sigma^2) + \tfrac{1}{2}\mathcal{N}(x \mid \theta_m, \sigma^2).$$

(This is an example of what is called a two-component ***mixture*** distribution.) Let's put independent normal priors on θ_f and θ_m:

$$p(\theta_f, \theta_m) = p(\theta_f)p(\theta_m) = \mathcal{N}(\theta_f \mid \mu_{0,f}, \sigma_0^2)\mathcal{N}(\theta_m \mid \mu_{0,m}, \sigma_0^2).$$

In Section 3, we will see how to put a prior on σ^2 (or equivalently, on $\lambda = 1/\sigma^2$), but for now, let's assume σ^2 is known. For the purposes of this example, let's use $\sigma = 8$ centimeters (about 3 inches). Based on common knowledge of typical human heights, let's choose the prior parameters (a.k.a. hyperparameters) as follows:

$\mu_{0,f}$	(mean of prior on female mean height)	165 centimeters ($\approx$ 5 feet, 5 inches)
$\mu_{0,m}$	(mean of prior on male mean height)	178 centimeters ($\approx$ 5 feet, 10 inches)
σ_0	(std. dev. of priors on mean height)	15 centimeters ($\approx$ 6 inches)

Another way to choose these parameters would be to estimate them from the distribution of the mean heights in various countries around the world—and the Dutch are known for being especially tall, so that could also be taken into account. Note that σ_0 represents our uncertainty about the mean heights, not about the heights of individuals.

It is known (Helguerro, 1904) that the combined distribution is bimodal if and only if

$$|\theta_f - \theta_m| > 2\sigma.$$

So, to address our question of interest ("Is human height bimodal?"), we would like to compute the posterior probability that this is the case, i.e., we want to know

$$\mathbb{P}(\text{bimodal} \mid \text{data}) = \mathbb{P}\big(|\boldsymbol{\theta}_f - \boldsymbol{\theta}_m| > 2\sigma \mid x_{1:k}, y_{1:\ell}\big).$$

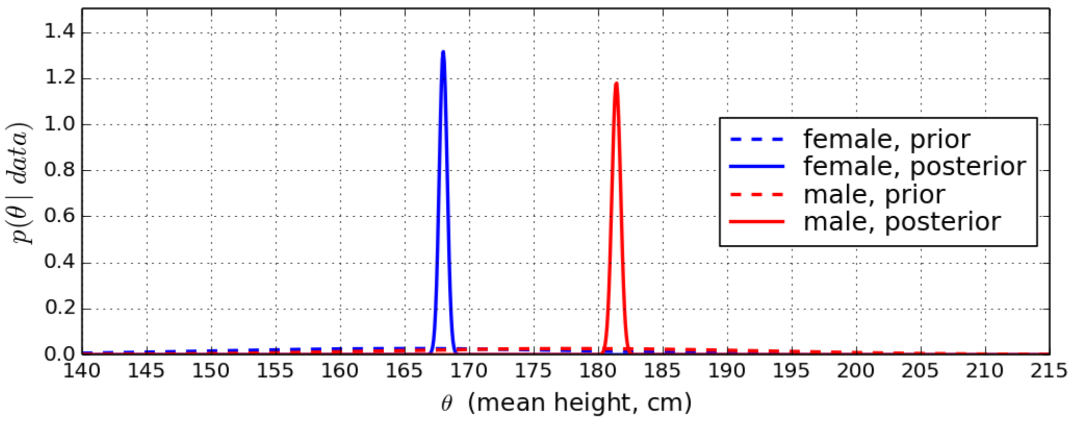

Figure 4: Priors and posteriors for the mean heights of Dutch women and men.

Results

We can compute the posteriors for θ_f and θ_m using Equation 2.1 for each of them, independently. Figure 4 shows the priors and posteriors.

- Sample means: $\bar{x} = 168.0$ cm (5 feet 6.1 inches) for females, and $\bar{y} = 181.4$ cm (5 feet 11.4 inches) for males.
- Posterior means: $M_f = 168.0$ cm for females, and $M_m = 181.4$ cm for males. (Essentially identical to the sample means, due to the relatively large sample size and relatively weak prior.)
- Posterior standard deviations: $1/\sqrt{L_f} = 0.30$ cm and $1/\sqrt{L_m} = 0.34$ cm.

By Equation 1.1 (a linear combination of independent normals is normal),

$$\boldsymbol{\theta}_m - \boldsymbol{\theta}_f \mid x_{1:k}, y_{1:\ell} \sim \mathcal{N}(M_m - M_f,\, L_m^{-1} + L_f^{-1}) = \mathcal{N}(13.4, 0.45^2)$$

so we can compute $\mathbb{P}(\text{bimodal} \mid \text{data})$ using the normal c.d.f. Φ:

$$\begin{aligned}\mathbb{P}(\text{bimodal} \mid \text{data}) &= \mathbb{P}\big(|\boldsymbol{\theta}_m - \boldsymbol{\theta}_f| > 2\sigma \mid x_{1:k}, y_{1:\ell}\big)\\ &= \Phi(-2\sigma \mid 13.4, 0.45^2) + \big(1 - \Phi(2\sigma \mid 13.4, 0.45^2)\big)\\ &= 6.1 \times 10^{-9}.\end{aligned}$$

Intuitive interpretation: The posteriors are about 13 or 14 centimeters apart, which is under the $2\sigma = 16$ threshold for bimodality, and they are sufficiently concentrated that the posterior probability of bimodality is essentially zero.

3 Conjugate prior for the mean and precision

Now, suppose that both the mean μ and the precision $\lambda = 1/\sigma^2$ are unknown, with $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \lambda^{-1})$ as before. The NormalGamma(m, c, a, b) distribution, with $m \in \mathbb{R}$

and $c, a, b > 0$, is a joint distribution on (μ, λ) obtained by letting

$$\begin{aligned} \boldsymbol{\lambda} &\sim \text{Gamma}(a, b) \\ \boldsymbol{\mu}|\lambda &\sim \mathcal{N}(m, (c\lambda)^{-1}). \end{aligned}$$

In other words, the joint p.d.f. is

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu \mid m, (c\lambda)^{-1})\,\text{Gamma}(\lambda \mid a, b)$$

which we will denote by $\text{NormalGamma}(\mu, \lambda \mid m, c, a, b)$ following our usual convention. It turns out that this provides a conjugate prior for (μ, λ). Indeed, the posterior is

$$\boldsymbol{\mu}, \boldsymbol{\lambda}|x_{1:n} \sim \text{NormalGamma}(M, C, A, B) \tag{3.1}$$

i.e., $p(\mu, \lambda|x_{1:n}) = \text{NormalGamma}(\mu, \lambda \mid M, C, A, B)$, where

$$\begin{aligned} M &= \frac{cm + \sum_{i=1}^n x_i}{c + n} \\ C &= c + n \\ A &= a + n/2 \\ B &= b + \tfrac{1}{2}\big(cm^2 - CM^2 + \textstyle\sum_{i=1}^n x_i^2\big). \end{aligned}$$

For interpretation, B can also be written (by rearranging terms) as

$$B = b + \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{2}\frac{cn}{c + n}(\bar{x} - m)^2. \tag{3.2}$$

Interpretation of posterior parameters

- M: Posterior mean for μ. It is a weighted average (convex combination) of the prior mean and the sample mean:

$$M = \frac{c}{c + n}m + \frac{n}{c + n}\bar{x}.$$

- C: "Sample size" for estimating μ. (The standard deviation of $\mu|\lambda$ is $\lambda^{-1/2}/\sqrt{C}$.)
- A: Shape for posterior on λ. Grows linearly with sample size.
- B: Rate (1/scale) for posterior on λ. Equation 3.2 decomposes B into the prior variation, observed variation (sample variance), and variation between the prior mean and sample mean:

$$B = (\text{prior variation}) + \tfrac{1}{2}n(\text{observed variation}) + \tfrac{1}{2}\tfrac{cn}{c+n}(\text{variation bw means}).$$

3.1 Derivation of the posterior

First, consider the NormalGamma density. Dropping constants of proportionality, multiplying out $(\mu - m)^2 = \mu^2 - 2\mu m + m^2$, and collecting terms, we have

$$\begin{aligned}
\mathrm{NormalGamma}(\mu, \lambda \mid m, c, a, b) &= \mathcal{N}(\mu \mid m, (c\lambda)^{-1}) \,\mathrm{Gamma}(\lambda \mid a, b) \\
&= \sqrt{\frac{c\lambda}{2\pi}} \exp\Big(-\tfrac{1}{2}c\lambda(\mu - m)^2\Big) \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda) \\
&\underset{\mu,\lambda}{\propto} \lambda^{a-1/2} \exp\Big(-\tfrac{1}{2}\lambda(c\mu^2 - 2cm\mu + cm^2 + 2b)\Big). \qquad (3.3)
\end{aligned}$$

Similarly, for any x,

$$\begin{aligned}
\mathcal{N}(x \mid \mu, \lambda^{-1}) &= \sqrt{\frac{\lambda}{2\pi}} \exp\Big(-\tfrac{1}{2}\lambda(x - \mu)^2\Big) \\
&\underset{\mu,\lambda}{\propto} \lambda^{1/2} \exp\Big(-\tfrac{1}{2}\lambda(\mu^2 - 2x\mu + x^2)\Big). \qquad (3.4)
\end{aligned}$$

Using Equations 3.3 and 3.4, we get

$$\begin{aligned}
p(\mu, \lambda | x_{1:n}) &\underset{\mu,\lambda}{\propto} p(\mu, \lambda) p(x_{1:n} | \mu, \lambda) \\
&= \mathrm{NormalGamma}(\mu, \lambda \mid m, c, a, b) \prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \lambda) \\
&\underset{\mu,\lambda}{\propto} \lambda^{a-1/2} \exp\Big(-\tfrac{1}{2}\lambda(c\mu^2 - 2cm\mu + cm^2 + 2b)\Big) \\
&\qquad \times \lambda^{n/2} \exp\Big(-\tfrac{1}{2}\lambda(n\mu^2 - 2(\textstyle\sum x_i)\mu + \sum x_i^2)\Big) \\
&= \lambda^{a+n/2-1/2} \exp\Big(-\tfrac{1}{2}\lambda\big((c+n)\mu^2 - 2(cm + \textstyle\sum x_i)\mu + cm^2 + 2b + \sum x_i^2\big)\Big) \\
&\overset{\text{(a)}}{=} \lambda^{A-1/2} \exp\Big(-\tfrac{1}{2}\lambda\big(C\mu^2 - 2CM\mu + CM^2 + 2B\big)\Big) \\
&\overset{\text{(b)}}{\propto} \mathrm{NormalGamma}(\mu, \lambda \mid M, C, A, B)
\end{aligned}$$

where step (b) is by Equation 3.3, and step (a) holds if $A = a + n/2$, $C = c + n$, $CM = (cm + \sum x_i)$, and

$$CM^2 + 2B = cm^2 + 2b + \sum x_i^2.$$

This choice of A and C match the claimed form of the posterior, and solving for M and B, we get $M = (cm + \sum x_i)/(c + n)$ and

$$B = b + \tfrac{1}{2}(cm^2 - CM^2 + \textstyle\sum x_i^2),$$

as claimed.

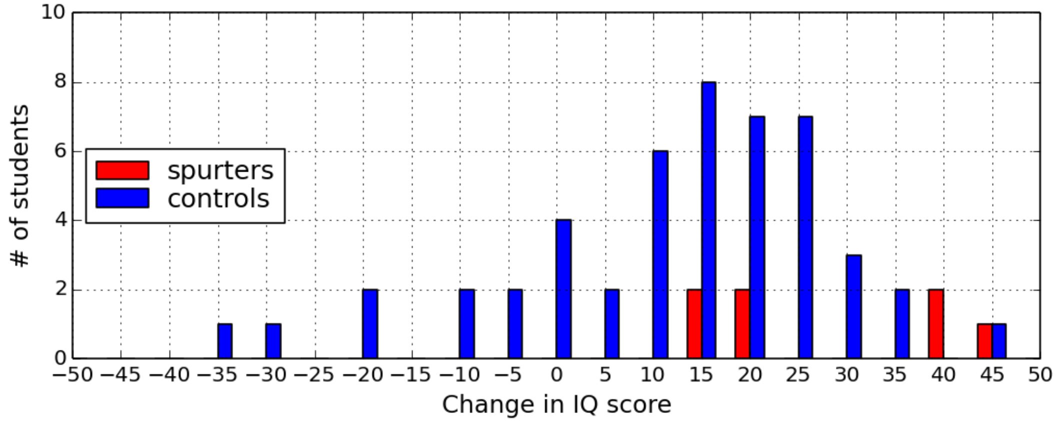

Figure 5: Histograms of change in IQ score for the two groups.

3.2 The Pygmalion effect

Do a teacher's expectations influence student achievement? In a famous study, Rosenthal and Jacobson (1968) performed an experiment in a California elementary school to try to answer this question. At the beginning of the year, all students were given an IQ test. For each class, the researchers randomly selected around 20% of the students, and told the teacher that these students were "spurters" that could be expected to perform particularly well that year. (This was not based on the test—the spurters were randomly chosen.) At the end of the year, all students were given another IQ test. The change in IQ score for the first-grade students was:[3]

spurters (S)
$x = (18, 40, 15, 17, 20, 44, 38)$

controls (C)
$y = (–4, 0, –19, 24, 19, 10, 5, 10, 29, 13, –9, –8, 20, –1, 12, 21, –7, 14, 13, 20, 11, 16, 15, 27, 23, 36, –33, 34, 13, 11, –19, 21, 6, 25, 30, 22, –28, 15, 26, –1, –2, 43, 23, 22, 25, 16, 10, 29)$

Summary statistics:

- spurters: $n_S = 7$, $\bar{x} = 27.4$, $\hat{\sigma}_x = 11.7$
- controls: $n_C = 48$, $\bar{y} = 12.0$, $\hat{\sigma}_y = 16.1$

See histograms in Figure 5. The average increase in IQ score is larger for the spurters. How strongly does this data support the hypothesis that the teachers' expectations caused the spurters to perform better than their classmates?

[3] The original data is not available. This data is from the `ex1321` dataset of the `R` package `Sleuth3`, which was constructed to match the summary statistics and conclusions of the original study.

Model

IQ tests are purposefully calibrated to make the scores normally distributed, so it makes sense to use a normal model here:

$$\text{spurters: } X_1, \ldots, X_{n_S} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_S, \lambda_S^{-1})$$
$$\text{controls: } Y_1, \ldots, Y_{n_C} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_C, \lambda_C^{-1}).$$

We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$? We don't know the standard deviations $\sigma_S = \lambda_S^{-1/2}$ and $\sigma_C = \lambda_C^{-1/2}$, and the sample seems too small to estimate them very well. The frequentist approach to this problem is rather complicated when $\sigma_S \neq \sigma_C$ (involving approximate t-distributions based on the Welch–Satterthwaite degrees of freedom).

On the other hand, it is easy using a Bayesian approach: we just need to compute the posterior probability that $\mu_S > \mu_C$:

$$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid x_{1:n_S}, y_{1:n_C}).$$

Let's use independent NormalGamma priors:

$$\text{spurters: } (\boldsymbol{\mu}_S, \boldsymbol{\lambda}_S) \sim \text{NormalGamma}(m, c, a, b)$$
$$\text{controls: } (\boldsymbol{\mu}_C, \boldsymbol{\lambda}_C) \sim \text{NormalGamma}(m, c, a, b)$$

with the following hyperparameter settings, based on subjective prior knowledge:

- $m = 0$ (Don't know whether students will improve or not, on average.)
- $c = 1$ (Unsure about how big the mean change will be—prior certainty in our choice of m assessed to be equivalent to one datapoint.)
- $a = 1/2$ (Unsure about how big the standard deviation of the changes will be.)
- $b = 10^2 a$ (Standard deviation of the changes expected to be around $10 = \sqrt{b/a} = \mathbb{E}(\lambda)^{-1/2}$.)

Aside: How to check whether a prior conforms to our beliefs?

1. Draw some samples from the prior and look at them—this is probably the best general strategy. See Figure 6. It's also a good idea to look at sample hypothetical datasets $X_{1:n}$ drawn using these sampled parameter values.
2. Plot the c.d.f. and check various quantiles (first quartile, median, third quartile), if univariate.
3. Plot the p.d.f., but beware—it can be misleading.
4. Look at various moments (e.g., mean, standard deviation), but beware—they can be misleading.

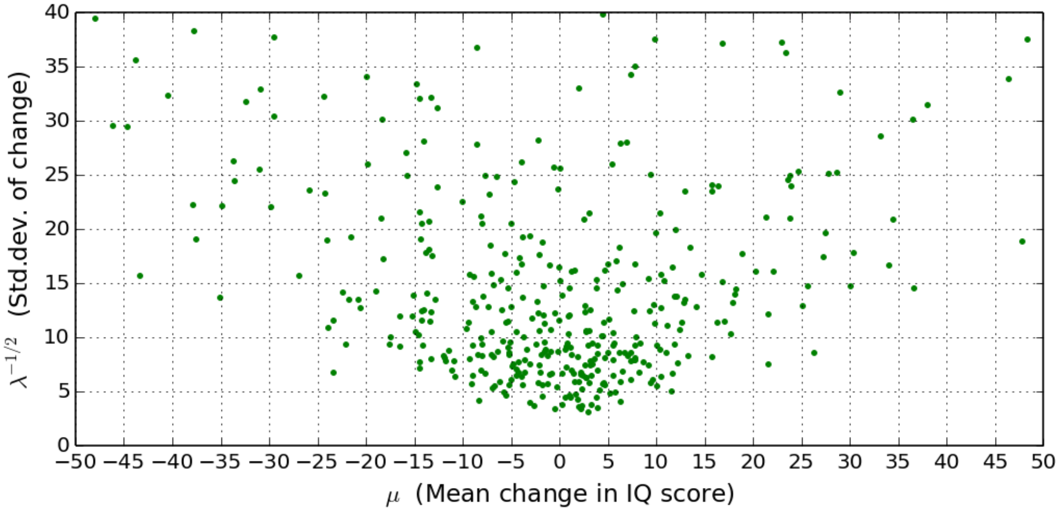

Figure 6: Samples of (μ, σ) from the prior.

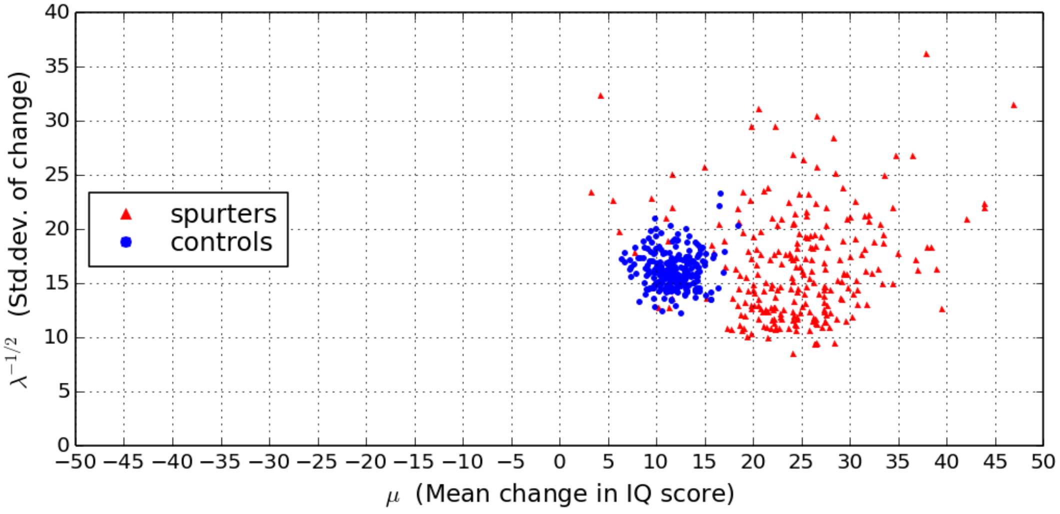

Figure 7: Samples of (μ, σ) from the posteriors for the two groups.

Results

Using Equation 3.1, the posterior parameters are

for spurters:

$$\begin{aligned}
M &= \frac{1 \cdot 0 + 7 \cdot 27.43}{1+7} = 24.0 \\
C &= 1 + 7 = 8 \\
A &= 1/2 + 7/2 = 4 \\
B &= 100/2 + \tfrac{1}{2} \cdot 7 \cdot 11.66^2 + \tfrac{1}{2}\frac{1 \cdot 7}{1+7}(27.43 - 0)^2 = 855.0
\end{aligned}$$

for controls:

$$\begin{aligned}
M &= \frac{1 \cdot 0 + 48 \cdot 12.04}{1+48} = 11.8 \\
C &= 1 + 48 = 49 \\
A &= 1/2 + 48/2 = 24.5 \\
B &= 100/2 + \tfrac{1}{2} \cdot 48 \cdot 16.10^2 + \tfrac{1}{2}\frac{1 \cdot 48}{1+48}(12.04 - 0)^2 = 6344.0
\end{aligned}$$

and so the posteriors are

$$\begin{aligned}
\boldsymbol{\mu}_S, \boldsymbol{\lambda}_S \mid x_{1:n_S} &\sim \text{NormalGamma}(24.0, 8, 4, 855.0) \\
\boldsymbol{\mu}_C, \boldsymbol{\lambda}_C \mid y_{1:n_C} &\sim \text{NormalGamma}(11.8, 49, 24.5, 6344.0).
\end{aligned}$$

Figure 7 shows a scatterplot of samples from the posteriors. Now, we can answer our original question: "What is the posterior probability that $\mu_S > \mu_C$?" The easiest way to do this is to take a bunch of samples from each of the posteriors, and see what fraction of times we have $\mu_S > \mu_C$. This is an example of a Monte Carlo approximation (much more to come on this in the future). To do this, we draw $N = 10^6$ samples from each posterior:

$$\begin{aligned}
(\mu_S^{(1)}, \lambda_S^{(1)}), \ldots, (\mu_S^{(N)}, \lambda_S^{(N)}) &\sim \text{NormalGamma}(24.0, 8, 4, 855.0) \\
(\mu_C^{(1)}, \lambda_C^{(1)}), \ldots, (\mu_C^{(N)}, \lambda_C^{(N)}) &\sim \text{NormalGamma}(11.8, 49, 24.5, 6344.0)
\end{aligned}$$

and obtain the approximation

$$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid x_{1:n_S}, y_{1:n_C}) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\mu_S^{(i)} > \mu_C^{(i)}\right) = 0.97.$$

This is consistent with a visual inspection of the scatterplots of posteriors in Figure 7.

Interpretation: The posterior probability that the spurter group had a higher mean change in IQ score is about 0.97. Thus, this data seems to support the hypothesis that the teachers' expectations did in fact play a role. (Note: The results of this study have been contested, since it has been difficult to replicate.)

Carl Friedrich Gauss

James Clerk Maxwell

Adolphe Quetelet

3.3 Inverse Gamma

If X is Gamma distributed then the distribution of $1/X$ is called the Inverse Gamma distribution. More precisely, if $X \sim \text{Gamma}(a, b)$ and $Y = 1/X$ then $Y \sim \text{InvGamma}(a, b)$, and the p.d.f. of Y is

$$\text{InvGamma}(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-b/y).$$

So, putting a $\text{Gamma}(a, b)$ prior on the precision λ is equivalent to putting an $\text{InvGamma}(a, b)$ prior on the variance $\sigma^2 = 1/\lambda$. The Inverse Gamma can be used to define a NormalInvGamma distribution for use as a prior on (μ, σ^2), which is sometimes more convenient than (but equivalent to) using a NormalGamma prior on (μ, λ).

4 History

In 1809, Carl Friedrich Gauss (1777–1855) proposed the normal distribution as a model for the errors made in astronomical measurements, as a formal way of justifying the use of the sample mean, by showing it to be the most likely estimate—that is, the maximum likelihood estimate—of the true value (and more generally, to justify the method of least squares in linear regression). With astonishing speed, following Gauss' proposal, Laplace proved the central limit theorem in 1810. Laplace also calculated the normalization constant of the normal distribution, which is not a trivial task. James Clerk Maxwell (1831–1879) showed that the normal distribution arose naturally in physics, particularly in thermodynamics. Adolphe Quetelet (1796–1874) pioneered the use of the normal distribution in the social sciences.

5 Exercises

Normal–Normal model

1. Derive the posterior predictive density $p(x_{n+1}|x_{1:n})$ for the Normal–Normal model from Section 2. (HINT: There is an easy way to do this and a hard way. The easy way uses Equation 1.1, writing $X_{n+1} = \boldsymbol{\theta} + Z$ given $x_{1:n}$, where $Z \sim \mathcal{N}(0, \lambda^{-1})$.)

2. Get a cup of hot tea or coffee and take a little break.

NormalGamma–Normal model

Two competitors for the snowiest city in the world are Aomori City in Japan, and Valdez in the state of Alaska. Here are annual snowfall records, in inches/year, for the two cities:

Aomori, 1954–2014
188.6, 244.9, 255.9, 329.1, 244.5, 167.7, 298.4, 274.0, 241.3, 288.2, 208.3, 311.4,
273.2, 395.3, 353.5, 365.7, 420.5, 303.1, 183.9, 229.9, 359.1, 355.5, 294.5, 423.6,
339.8, 210.2, 318.5, 320.1, 366.5, 305.9, 434.3, 382.3, 497.2, 319.3, 398.0, 183.9,
201.6, 240.6, 209.4, 174.4, 279.5, 278.7, 301.6, 196.9, 224.0, 406.7, 300.4, 404.3,
284.3, 312.6, 203.9, 410.6, 233.1, 131.9, 167.7, 174.8, 205.1, 251.6, 299.6, 274.4,
248.0

Valdez, 1976–2013
351.0, 379.3, 196.1, 312.3, 301.4, 240.6, 257.6, 304.5, 296.0, 338.8, 299.9, 384.7,
353.5, 312.8, 550.7, 327.1, 515.8, 343.4, 341.6, 396.9, 267.3, 230.6, 277.4, 341.0,
377.0, 391.3, 337.0, 250.4, 353.7, 307.7, 237.5, 275.2, 271.4, 266.5, 318.7, 215.5,
438.3, 404.6

Assume that for each city independently, the data is i.i.d. normal.

3. Do you think an i.i.d. normal model is appropriate here? Why or why not?

4. Is the mean annual snowfall for Valdez higher than that of Aomori? To address this question, perform an analysis like the one for the Pygmalion effect in Section 3.2. In particular, your analysis should involve computing the posterior probability that the mean annual snowfall for Valdez higher than that of Aomori. Choose prior parameters (hyperparameters) according to your personal subjective prior.

5. (Continuation of Exercise 4) Try different values for the hyperparameters, to see what effect they have on the results. Report your results for three different settings of the hyperparameters.

Supplementary material

- Hoff (2009), Chapter 5.
- mathematicalmonk videos, Machine Learning (ML) 7.9 and 7.10
  https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

References

- Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? The American Statistician, 56(3), 223-229.

- Krul, A. J., Daanen, H. A., & Choi, H. (2011). Self-reported and measured weight, height and body mass index (BMI) in Italy, the Netherlands and North America. The European Journal of Public Health, 21(4), 414-419.

- Helguerro, F. (1904), Sui Massimi Delle Curve Dimorfiche. Biometrika, 3, 85-98.

- Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. The Urban Review, 3(1), 16-20.