Module 3: Introduction to the Normal Distribution

Rebecca C. Steorts

Announcements

- Thank you to everyone that completed the survey.
- I'll provide some actionable items below.

Feedback

- The github repo has been cleaned up extensively.
- Exams will be solely online (during class period) by feedback from class, and will be open note and open book.
- "Please move the lab time." These weren't set by me, and I don't have the authority to do this.
- "Please go faster/slower." "Great pace."
- "I'd like more OH hours." The TA's and I are pretty flexible if you have a question or need to meet. Can anyone not make any OH due to a class conflict? If so, see me after class.
- The class is quite divided regarding being online and being in person. We'll do out best to work this out in a safe way for everyone.
- "Lab is not helping me."

Resources

Don’t let the resources be overwhelming.

Instead, figure out what works the best for you and use them. What works for one student may be drastically different than what works for another student.

My goal is to set you up for success and provide you with many different ways to succeed so you’re not struggling.

If you find yourself struggling, please let me know so I can help you.

Agenda

- The normal distribution
- The variance versus precision
- The re-parameterized normal distribution
- Common properties
- The normal-uniform model
- The normal-normal model

Normal distribution

The normal distribution $\mathcal{N}(\mu, \sigma^2)$

- 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 $x \in \mathbb{R}$.

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.

Re-parameterized Normal

In this parametrization, the p.d.f. is

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

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

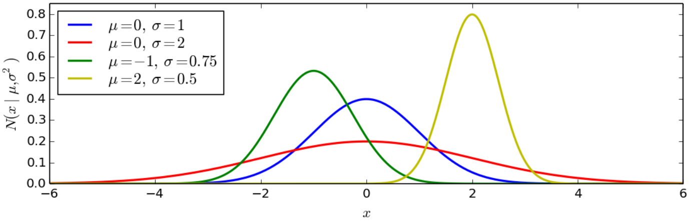

Figure 1: Normal distribution with various choices of μ and σ.

Normality?

- The central limit theorem (CLT) states that the sum of a large number of independent random variables tends to be approximately normally distributed.
- Real world data often appears approximately normal.

Normality?

- Human heights and other body measurements,
- Cumulative hydrologic measures such as annual rainfall or monthly river discharge,
- Errors in astronomical or physical observations,
- 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

Example: stock market indices, due to the effect of compound interest.

Properties of the Normal distribution

- 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

$$aX + bY \sim \mathcal{N}(a\mu + bm,\ a^2\sigma^2 + b^2 s^2). \tag{1}$$

- 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 $\mathcal{N}(m, s^2)$.

Normal-Uniform

$$X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2).$$

Assume the prior on θ is constant over the real line. We can write this as $p(\theta) \propto 1$, where $-\infty < \theta < \infty$.

Derive the posterior distribution.[1]

[1] Please observe that the posterior is not conjugate to the prior in this situation, so it's very important to make sure that the posterior is a proper distribution.

Solution

$$
\begin{aligned}
p(\theta \mid x_{1:n}) &\propto \mathcal{N}(x_{1:n} \mid \theta, \sigma^2) \times 1 && (2)\\
&\propto \big(\frac{\ell}{2\pi}\big)^{n/2} \exp\big(-\tfrac{1}{2}\ell \sum_i (x_i - \theta)^2\big)\\
&\propto \exp\big(-\tfrac{1}{2}\ell \sum_i (x_i - \bar{x} + \bar{x} - \theta)^2\big)\\
&\propto \exp\big(-\tfrac{1}{2}\ell \sum_i (x_i - \bar{x})^2\big) \exp\big(-\tfrac{1}{2}\ell \sum_i (\bar{x} - \theta)^2\big)\\
&\propto \exp\big(-\tfrac{1}{2}\ell \sum_i (\bar{x} - \theta)^2\big) && (3)\\
&= \exp\big(-\tfrac{n\ell}{2}(\theta - \bar{x})^2\big) = \mathcal{N}(\theta \mid \bar{x}, (n\ell)^{-1}). && (4)
\end{aligned}
$$

This implies that

$$\theta \mid x_{1:n} \sim N(\bar{x}, (n\ell)^{-1}) = N(\bar{x}, \sigma^2/n)$$

Commonly asked questions on the Normal-Uniform derivation

1. Why do we add and subtract $\bar{x}$?

This trick makes the cross product term 1.

2. Why is the cross product term 1? The cross product term is:

$$\begin{aligned}
&\exp\big(-\tfrac{2}{2}\ell\big)\sum_i (x_i-\bar{x})(\bar{x}-\theta)\\
&=\exp\big(-\ell)(\bar{x}-\theta)\sum_i (x_i-\bar{x})\\
&=e^0=1.
\end{aligned}$$

Normal-Normal

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

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.

Posterior derivation

We begin with the **likelihood** of the normal distribution. We work with proportionality with respect to θ as we will combine the **likelihood** and **prior** to find the posterior, where the parameter of interest is θ.

For any x and ℓ,

$$\begin{aligned} \mathcal{N}(x \mid \theta, \ell^{-1}) &= \sqrt{\frac{\ell}{2\pi}} \exp\big(-\tfrac{1}{2}\ell(x-\theta)^2\big) \\ &\underset{\theta}{\propto} \exp\big(-\tfrac{1}{2}\ell(x^2 - 2x\theta + \theta^2)\big) \\ &\underset{\theta}{\propto} \exp\big(\ell x\theta - \tfrac{1}{2}\ell\theta^2\big). \end{aligned} \tag{5}$$

Note: we drop the **constant term** and we will do this often when working with the normal distribution.

Posterior derivation (continued)

We now consider the **prior** distribution on θ.

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

$$\begin{aligned} \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) &= \sqrt{\frac{\lambda_o}{2\pi}} \exp\big(-\tfrac{1}{2}\lambda_o(\theta-\mu_o)^2\big) \\ &= \sqrt{\frac{\lambda_o}{2\pi}} \exp\big(-\tfrac{1}{2}\lambda_o(\mu_o-\theta)^2\big) \\ &= \mathcal{N}(\mu_0 \mid \theta, \lambda_0^{-1}) \underset{\theta}{\propto} \exp\big(\lambda_0\mu_0\theta - \tfrac{1}{2}\lambda_0\theta^2\big), \end{aligned} \tag{6}$$

where $x = \mu_0$ and $\ell = \lambda_0$.

Posterior derivation (continued)

Let

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

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

where step (a) uses Equations 5 and 6, and step (b) uses Equation 5 with $x = M$ and $\ell = L$.

Posterior derivation (continued)

Recall

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

It turns out that the posterior is

$$\boldsymbol{\theta}|x_{1:n} \sim \mathcal{N}(M, L^{-1}) \tag{7}$$

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

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

Heights of Adult Humans

- Heights tend to be normally distributed because there are many independent genetic and environmental factors which contribute additively to overall height
- This leads to a normal distribution due to the central limit theorem.

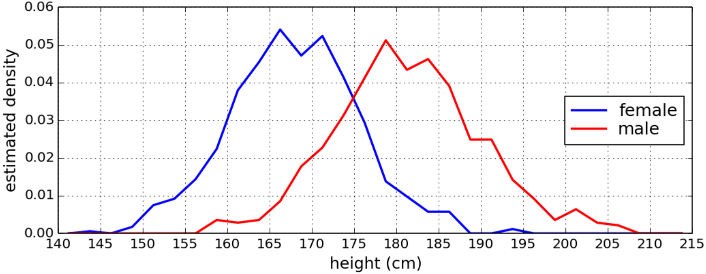

Figure 2: Estimated densities of the heights of Dutch women and Dutch men based on a sample of 695 women and 562 men.

Heights of Adult Humans

- Consider combined distribution of heights (pooling females and males together). Would this be normal?
- It is thought that such data is bimodal (having two maxima). Is it really bimodal? (See, Schilling et al. (2002))

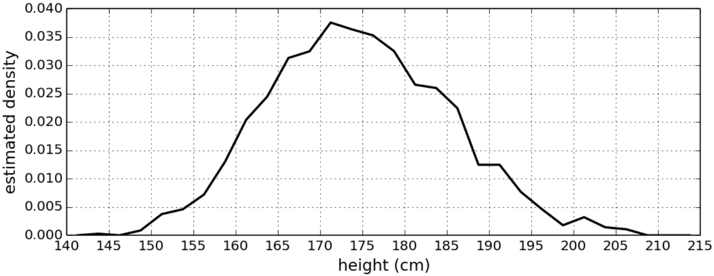

Figure 3: Estimated density for Dutch women and men together, assuming there is an equal proportion of women and men in the population.

Heights of Adult Humans, Combined

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?

Our Assumptions

- Assume female heights and male heights are each normally distributed.
- Assume both female heights and male heights have different means but the same standard deviation.
- Assume 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.)

Model

Let's put independent normal priors on θ_f and θ_m:

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

- 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 ht)	165 cm ($\approx$ 5 ft, 5 in)
$\mu_{0,m}$	(mean of prior on male mean ht)	178 cm ($\approx$ 5 ft, 10 in)
σ_0	(std. dev. of priors on mean ht)	15 cm ($\approx$ 6 in)

Bimodal Fact

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

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

Results

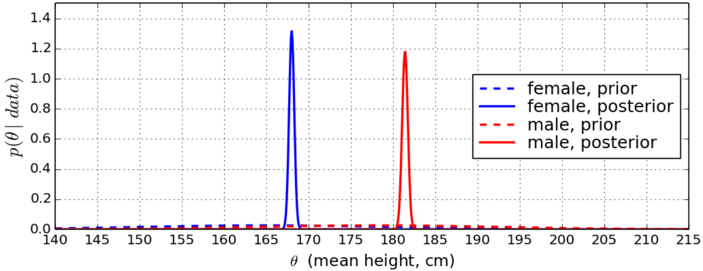

Figure 4: Priors and posteriors for the mean heights of Dutch women and men.

Results (continued)

We can compute the posteriors for θ_f and θ_m using Equation 7 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.

Results (continued)

By Equation 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.

Takeaways

- We reviewed the univariate normal distribution and properties of it (such as symmetry about the mean).
- We discussed types of data that empirically have a normal distribution to motivate its usage.
- We dervied the Normal-Uniform model, where the uniform distribution is the first improper prior we have seen in this course.
- We derived the Normal-Normal model.
- We considered an application of human heights from Schilling et al. (2002), where we showed that the posterior probability of bimodality was zero (regarding a plot of combined heights).

Detailed Takeaways for Exam

- Working with the univariate normal distribution
- Knowing the precision and variance relationship
- Knowing the shape of the normal distribution
- Knowing the CLT
- Knowing what type of data is approximately normal and what data is not
- Knowing properties of the normal that we will often work with such as symmetry about the mean
- Being able to derive the posterior of the normal-uniform model
- Understanding that a non-informative prior is highly informative and why.
- Being able to derive the normal-normal posterior distribution
- Knowing how to write out the posterior mean and variance intuitively

Detailed Takeaways for Exam (Continued)

- You should be able to work with a two-component mixture model (as in the case study)
- Suppose I provide you with a case study such as the one on human heights. You should be able to explain why it makes sense to model the data as normal and what paramaters would be reasonable given the description.
- Given as case study, you should be able to incorporate a pilot (or prior study) into your prior distribution and back up how you incorporate it.
- You should be able to update the posterior accordingly.
- You be able to state benefits of your model specification and weaknesses.
- You should be able to state a better model that would be more realistic once we have covered the normal-normal-gamma model (module 4).
- You should be able to discuss sensitity of the posterior analysis for any hyper-parameters.

Exercise

Recall that Recall

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

What happens to the posterior mean and the precision as $n \to \infty$.

Exercise

Let's consider the posterior mean.

$$\begin{aligned} M &= \frac{\lambda_0\mu_0 + \lambda \sum_{i=1}^n x_i}{\lambda_0 + n\lambda} \\ &= \frac{\lambda_0\mu_0 + \lambda n\bar{x}}{\lambda_0 + n\lambda} \\ &= \frac{\lambda_0\mu_0/n + \lambda\bar{x}}{\lambda_0/n + \lambda} \end{aligned}$$

As $n \to \infty$,

$$M \to \lambda\bar{x}/\lambda = \bar{x}.$$

Exercise

Let's consider the posterior variance.

$$L^{-1} = \frac{1}{\lambda_0 + n\lambda}$$

As $n \to \infty$,

$$L^{-1} = \frac{1}{\lambda_0 + n\lambda} = \frac{1/n}{\lambda_0/n + \lambda} \approx \frac{1}{n\lambda} \to 0$$

Interpretation

What can we learn from this example?

If our sample size is large enough (for any application or real world example), then

1. the posterior mean will tend to the the sample mean.
2. the posterior variance will tend to 0.

Module 3 Class Notes

Module 3 Class Notes can be found here:

https://github.com/resteorts/modern-bayes/tree/master/lectures
ModernBayes20/lecture-3/03-class-notes