The normal-normal model

Giorgio Corani - (IDSIA, SUPSI)

Bayesian Data Analysis and Probabilistic Programming

Credits

- Chap. 5 of *Bayes Rules! An Introduction to Applied Bayesian Modeling*
  - https://www.bayesrulesbook.com/chapter-5.html

The Normal model

- Let Y be a continuous random variable which can take values in $(-\infty,\infty)$
- The variability of Y might be well represented by a Normal model
$$Y \sim N(\mu, \sigma^2)$$

The Normal model

- The Normal pdf is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

- With:

$$\begin{aligned} E(Y) &= Mode(Y) = \mu \\ Var(Y) &= \sigma^2 \\ SD(Y) &= \sigma \end{aligned}$$

Standard deviation σ

- σ provides a sense of scale for Y.
- Roughly 95% of Y values are within 2 standard deviations of μ:

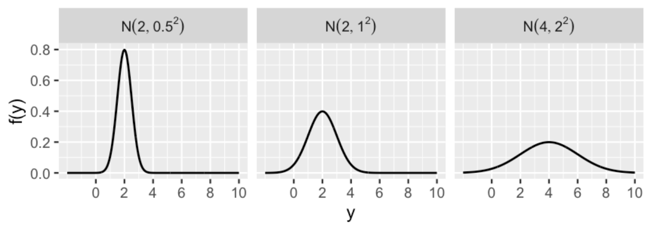
$$\mu \pm 2\sigma$$

- Roughly 99% of Y values are within 3 standard deviations of μ:

$$\mu \pm 3\sigma$$

The normal model

- The Normal model is bell-shaped and symmetric around μ.
- As σ gets larger, the pdf becomes more spread out.
- Though a Normal variable is defined in $(-\infty, \infty)$, the plausibility of values that are more than 3 standard deviations σ from the mean μ is negligible.



Example

- The volume of the hippocampus (a part of the brain) is researched in studies about the effect of concussions.
- In the general population, both halves of the hippocampus have a volume between 3.0 and 3.5 cm^3.
- Thus, the hippocampal volume is thought to vary, within the population, between 6 and 7 cm^3.
- The average volume μ is thought to be between 6.4 and 6.6 cm^3.

Normal prior

- Assuming symmetry, we formalize our prior information about μ as:

$$\mu \sim N(\mu', \sigma_\mu)$$

which in this example is :

$$\mu \sim N(6.5, 0.05)$$

- μ' is our prior guess on the value of μ.
- σ_μ represents our uncertainty on the guess μ'.
- According to this prior, μ lies with 95% probability in (6.4, 6.6).
- We allow the volume to range over $(-\infty, \infty)$, but values beyond $\mu \pm 3\sigma$ are given negligible probability.

Normal likelihood

- We now define a model for the distribution of the observations.
- We make a *second* assumption of normality.
- The hyppocampal volumes observed in n subjects ($y_1, y_2, ..., y_n$) are normally distributed $N(\mu, \sigma)$.

Normal likelihood

- μ is the mean volume in the population.
- σ expresses the spread of the measures within the population.
- We expect y to vary in (6-7); we interpret this interval as $\mu \pm 2\sigma$, hence it has length of 4σ.
- We thus set σ=0.25.

Independence

- We morever assume the observations $y_1, ..., y_n$ to be *independent* samples from $N(\mu, \sigma)$.
- This is realistic: the measure y_i tells us nothing about the measure y_{i+1} (assuming they refer to different subjects)

Likelihood

Assuming independence, the joint pdf of the n measures $(y_1, y_2, ..., y_n)$ is the product of the unique Normal pdfs $f(y_i \mid \mu)$:

$$f(\vec{y}|\mu) = \prod_{i=1}^{n} f(y_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right].$$

- $\vec{y}$ is the vector containing the measures $y_1, ...y_n$.

The Normal-Normal model

$$\mu \sim N(\mu', \sigma_\mu)$$
$$\vec{y} \sim N(\mu, \sigma)$$

- We treat μ', σ_μ and σ as fixed numbers.
- The likelihood assumes independence of the observations $y_1, y_2, \ldots, y_n$
- The only parameter of the model is μ.
- Later we will treat also σ as a parameter.

The Normal-Normal model

In the hyppocampus example, the normal-normal model is:

$$\mu \sim N(6.5, 0.05)$$
$$\vec{y} \sim N(\mu, 0.25)$$

- The only parameter of this model is μ.

Your turn: normal likelihood functions

- For a Normal random sample $y_i \sim N(\mu, \sigma)$ with σ=10 we observe:

$$y_1, y_2, y_3 = (-4.3, 0.7, -19.4)$$

- Specify and plot the corresponding likelihood function of μ.
  - Hint: define a range for the values of μ and make it a grid of e.g. 100 points
  - Compute the likelihood for each point of the grid.

Conjugacy of the normal-normal model

- Denote the sample mean as $\bar{y} = \frac{1}{n}\sum_i y_i$.
- The posterior density of μ is normal with updated parameters:

$$\mu|\vec{y} \;\sim\; N\Bigg(\underbrace{\mu' \frac{\sigma^2}{n\sigma_\mu^2 + \sigma^2} + \bar{y}\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \sigma^2}}_{\text{posterior mean}}, \underbrace{\frac{\sigma_\mu^2 \sigma^2}{n\sigma_\mu^2 + \sigma^2}}_{\text{posterior variance}}\Bigg).$$

Posterior mean

$$\mu|\vec{y} \;\sim\; N\Bigg(\mu' \underbrace{\frac{\sigma^2}{n\sigma_\mu^2+\sigma^2}}_{w} + \bar{y} \underbrace{\frac{n\sigma_\mu^2}{n\sigma_\mu^2+\sigma^2}}_{1-w},\; \frac{\sigma_\mu^2\sigma^2}{n\sigma_\mu^2+\sigma^2}\Bigg).$$

- The posterior mean is a weighted average of the prior mean μ' and the sample mean $\bar{y}$.
- As n increases, the posterior mean converges to $\bar{y}$.
- As n increases, the posterior variance decreases.
- The normal-normal is a *conjugate* model, since the posterior density is normal like the prior.

Your turn

With reference to the hyppocampus model:

- Which is the posterior mean, if we did 5 measures with $\bar{y} = 6.7$?
- Which is the posterior mean, if we did 35 measures with $\bar{y} = 6.7$?

Your turn

- Let μ be the average 3 p.m. temperature in Lugano.
- Your friend's prior understanding is that μ is around 15 degrees Celsius, though might be anywhere between 5 and 25 degrees.
- To learn about μ, he will analyze 1000 days of temperature data.
- Letting y_i denote the 3 p.m. temperature on day i, they'll assume that daily temperatures vary Normally around μ with a standard deviation of 5 degrees.
- Formalize a normal-normal model.

Your turn

- Solve exercises 5.9 and 5.10 from:

https://www.bayesrulesbook.com/chapter-5.html#exercises-4

- Compare the analytical posterior and the numerical posterior obtained via gridding (you create a grid of values of μ, multiply prior and likelihood, and normalize).

Treating σ as a parameter

- A more sophisticated approach is to treat σ as a parameter.
- We assigning a prior to it; it should cover a wide range of plausible values for σ, leaving out however values that make no sense.
- In this case there is no closed-form expression of the posterior.

Half-normal distribution

- σ is strictly positive; a suitable prior is the *half-normal* distribution.
- The half-normal is a Gaussian restricted to positive values.
- Sample s from a half-normal are obtained by:
  - sampling from a normal distribution
  - applying the absolute value to the sampled values
  - $s \sim |N(0, \xi)|$, where ξ is the standard deviation of the underlying normal. It is referred to as the *scale* of the half-normal.

The half-normal distribution

- It is asymmetric and right-skewed.
- It has long tails which are much larger than the median.

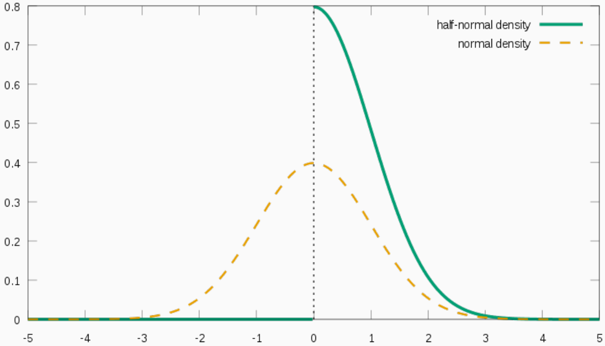
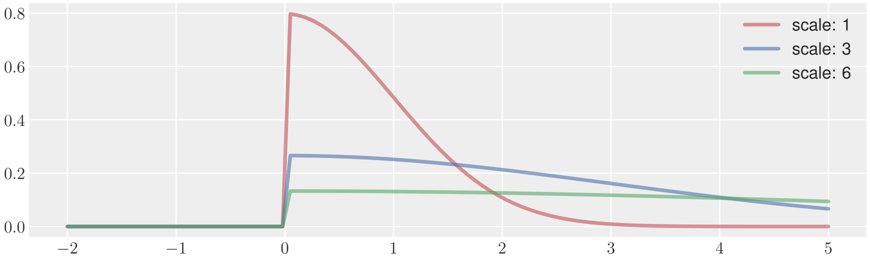


Figure 1: from wikipedia

Effect of the scale parameter ξ

- The half-normal pdf is characterized by a scale parameter (the standard deviation of the underlying normal).



Tuning the half-normal distribution

- You can tune the scale of the HN by considering a plausible value of σ, and choose the scale so that it is close to the median of the HN.
- E.g., assume a plausible value of σ is 7.5.
- With 95% probability the measures are lie in an interval of ± 15 around the mean.
- But we are uncertain about this statement, as the interval could be well of ± 30 .

Tuning the half-normal distribution

- We try different scales, until the median is about 7.5.
- Notice the long tails of the distribution, which allows to model to correct if our prior median guess (7.5) is underestimated.

```
pd.DataFrame(halfnorm.rvs(size=1000, scale=11)).describe()
```

```
##                  0
## count  1000.000000
## mean      8.536005
## std       6.543671
## min       0.027968
## 25%       3.454644
## 50%       6.905888
## 75%      12.643051
## max      36.857232
```

Probabilistic model with σ as parameter

$\mu \sim N(\mu', \sigma_\mu)$	prior beliefs about μ
$\sigma \sim$ Half-Normal(ξ)	prior beliefs about σ
$y \sim N(\mu, \sigma)$	the observation are normally distributedσ

- We cannot treat this model analytically, as the prior are no longer conjugates.
- We will implement it later via probabilistic programming.

Conceptual exercise

- Try to define a probabilistic model of the distribution of height of adult males in Switzerland

Prior for μ

- The mean height of the population could be 175.
- Keeping our prior broad, we state the mean height of the population to lie with 99% probability between 160 and 190 cm (the 99% interval roughly corresponds to $\mu \pm 3\sigma$).
  - $\mu \sim N(175, 5)$

Prior for σ

- We shall now assign a prior to σ. Within the population, we assume the height to lie with 99% probability between 100 and 250 (broad but realistic range).
- Hence the corresponding value of the standard deviation is (250-100)/6 = 25.

Tuning the half-normal

- A half-normal distribution with scale 35 has roughly median 25:

```
pd.DataFrame(halfnorm.rvs(size=1000, scale=35)).describe()
```

```
##                  0
## count  1000.000000
## mean     29.129983
## std      21.866708
## min       0.035822
## 25%      12.558447
## 50%      24.777304
## 75%      40.932584
## max     136.063258
```

Likelihood (distribution of the data)

- Under the assumption of normality and independence, the likelihood is:

$$y \sim \mathcal{N}(\mu, \sigma)$$

- No further specification is required.

The resulting model

$$
\begin{aligned}
\mu &\sim N(175, 5) \\
\sigma &\sim \text{half-normal}(35) \\
\vec{y} &\sim N(\mu, \sigma)
\end{aligned}
$$

Numerical exercise

- Compute the posterior distribution of the height model by using a bi-dimensional gridding.
- Assume to have collected two observations: 168 cm. and 178 cm.
- Plot the posterior joint and the posterior marginals.
- Solution: see code on *icorsi*.

Solutions

Solution of the likelihood exercise

- Compute the likelihood as a function of μ for a normal sample with σ=10, given the observations:

$$y_1, y_2, y_3 = (-4.3, 0.7, -19.4)$$

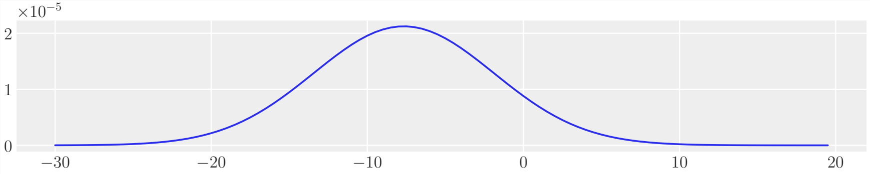
Solution of the likelihood exercise

```
#based on the observations, plausible values of mu range between -30 and 20.
mu = np.arange(-30, 20, 0.5)
sigma = 10

#a likelihood value for each value of mu
lik = norm.pdf(-4.3, loc=mu, scale=sigma)

#under independence, the likelihood of each observation multiplies
lik = lik * norm.pdf(0.7, loc=mu, scale=sigma)
lik = lik * norm.pdf(-19.4, loc=mu, scale=sigma)

plt.figure(figsize=(10, 2))
plt.plot(mu, lik)
```



- The function has its maximum in correspondence of $\bar{y}$.
- Small numerical values (10 e-5)

Solution

- The values in the previous slide are numerically small. With more data, and more likelihood multiplication, it will become numerically untractable.
- For this reason it is numerically better to work with the log of the likelihood (log-likelihood) and exponentiate back the results.
- You will see how do this in the labs.

Posterior mean for the hyppocampus model

$$\mu|\vec{y} \sim N\left(\mu' \underbrace{\frac{\sigma^2}{n\sigma_\mu^2 + \sigma^2}}_{w} + \bar{y} \underbrace{\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \sigma^2}}_{1-w}, \frac{\sigma_\mu^2 \sigma^2}{n\sigma_\mu^2 + \sigma^2}\right)$$

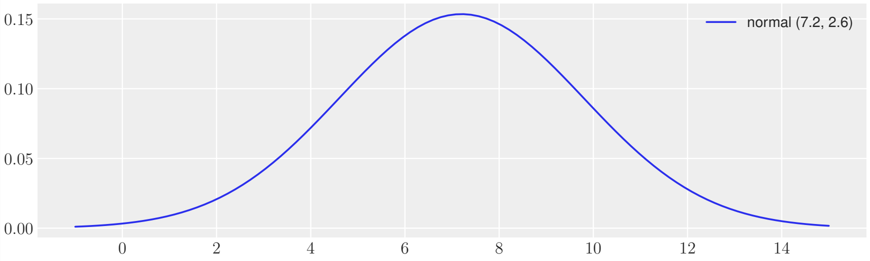
$$\mu' = 6.5 \;\; \sigma_\mu = 0.05; \sigma = 0.25$$

- For n=5, we get w=0.833 and posterior mean 6.53.
- For n=35, we get w=0.416 and posterior mean 6.61.
- When we have more data, the posterior mean is closer to the sample mean.

Solution of exercises 5.9 and 5.10

- https://www.bayesrulesbook.com/chapter-5.html#exercises-4

```
plt.figure(figsize=(10, 3))
x = np.linspace(-1, 15, 100)
mu = 7.2
sigma = 2.6
y = stats.norm.pdf(x, loc = mu, scale = sigma)
plt.plot(x, y, label='normal (%s, %s)' % (mu, sigma))
plt.legend(fontsize=12)
```



Questions b,c,d,e

- $P(X) >= 7.6, P(X) >= 4, P(X) < 0, P(X) > 8$

```
mu = 7.2
sigma = 2.6
p1 = 1 - stats.norm.cdf(7.6, loc = mu, scale = sigma)
p2 = 1 - stats.norm.cdf(4, loc = mu, scale = sigma)
p3 = stats.norm.cdf(0, loc = mu, scale = sigma)
p4 = 1 - stats.norm.cdf(8, loc = mu, scale = sigma)
p1
```

```
## 0.43886552075085816
```

```
p2
```

```
## 0.8907954066903792
```

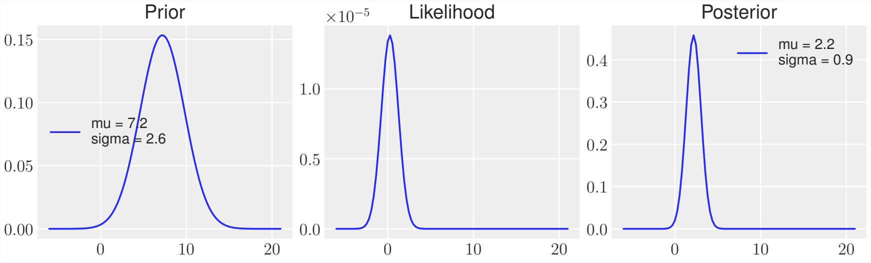
```
p3
```

```
## 0.00280944107441954
```

```
p4
```

```
## 0.3791582367631452
```

Prior, likelihood and posterior



Code of the previous figure

```
plt.figure(figsize=(10, 3))
mu = np.linspace(-6, 21, 100)

prior_mean = 7.2
prior_sigma = 2.6
sigma_lik = 2

#prior
prior = stats.norm.pdf(mu, prior_mean, prior_sigma)
plt.subplot(1, 3, 1)
plt.plot(mu, prior, label='mu = %s\n sigma = %s'
  % (prior_mean, prior_sigma))

plt.title('Prior')
plt.legend(fontsize=12)
```

Code of the previous figure (cont'd)

```
#likelihood
#code below could be  vectorized
y   = np.array([-0.7, 1.2, 4.5, -4])
lik = norm.pdf (y[0], loc = mu, scale = sigma_lik)
lik = lik * norm.pdf (y[1], loc = mu, scale = sigma_lik)
lik = lik * norm.pdf (y[2], loc = mu, scale = sigma_lik)
lik = lik * norm.pdf (y[3],  loc = mu, scale = sigma_lik)


plt.subplot(1, 3, 2)
plt.plot(mu, lik)
plt.title('Likelihood')
```

Code of the previous figure (cont'd)

```
#posterior
y_bar     = np.mean(y)
n         = len(y)
w_prior   = sigma_lik**2 / (n*prior_sigma + sigma_lik**2)

post_mean = prior_mean * w_prior + y_bar * (1 - w_prior)

post_var  = (prior_sigma**2 * sigma_lik**2)/
            (n * prior_sigma**2 + sigma_lik**2)
post_s    = np.sqrt(post_var)
posterior = stats.norm.pdf(mu, post_mean, post_var)

 plt.subplot(1, 3, 3)
plt.plot(mu, posterior, label='mu = %s\n sigma = %s' % (post_mean, post_s))
plt.legend(fontsize=12)
plt.title('Posterior')
```