BINOMIAL, POISSON, AND NORMAL MODELS

BST228 Applied Bayesian Analysis

Speaker notes

No notes on this slide.

RECAP

- Binomial likelihood with beta prior.
- Poisson likelihood with gamma prior.
- Posterior predictive distribution.

Speaker notes

- Binomial likelihood for # events in finite population (North Carolina low birth weight; Warfarin complications).
- Beta prior is conjugate; we can derive posterior in closed form.
- Poisson likelihood # events with given rate (Prussian soldiers kicked by horses & hospital admissions).
- Gamma prior is conjugate.
- Why are these different?
  - Babies either have low birth weight or not; soldiers can be kicked a lot.
  - Poisson to binomial: 3 of 104 soldiers were kicked.
  - Binomial to Poisson: 17 babies with LBW born.
- Posterior predictive is distribution of future outcomes given observed outcomes.
  - Extra uncertainty compared with MLE is important, especially for small sample sizes.

OUTLINE

- Wrap up Poisson and binomial models.
- Why non-informative priors are often informative.
- Normal model as a two-parameter distribution.

Speaker notes

- Wrap up count outcomes by considering another examples with binomial or Poisson likelihood: asthma mortality rates. Sometimes choosing the *right* model is not straightforward.
- Sometimes uninformative priors are quite informative depending on the parameterization of the model.
- Normal model has two parameters: location and scale. It is the fundamental building block of most hierarchical models (random effects for between-subject variability, time series models, least-squares regression, Gaussian processes, …).

ASTHMA MORTALITY

In a city of $n = 200,000$, $y = 3$ people died of asthma in 2018.

Speaker notes

- What is an appropriate likelihood for this problem? Raise hands for binomial, Poisson, another likelihood.

ASTHMA MORTALITY

What is the probability to die of asthma in a given year?

➡️ Binomial likelihood.

What is the rate at which people die of asthma?

➡️ Poisson likelihood.

Speaker notes

- Data may not be enough to tell us about the appropriate model.
- The model also depends on the question we want to answer.
- Formulating a model is a science but also sometimes an art.
- Incorporating your and your collaborators' experience and domain knowledge is essential for building "good" models.

DERIVATION OF POSTERIOR FOR BINOMIAL LIKELIHOOD

We have the binomial likelihood and conjugate beta prior with hyperparameters a_0 and b_0 such that

$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$
$$p(\theta) = \frac{\theta^{a_0 - 1} (1-\theta)^{b_0 - 1}}{B(a_0, b_0)},$$

where $B(a_0, b_0)$ is a normalization constant. Neglecting constants in θ, the posterior is

$$p(\theta \mid y, n, a_0, b_0) \propto \theta^{a_0 + y - 1} (1-\theta)^{b_0 + n - y - 1}$$

which has the kernel of a beta distribution. The posterior is thus a beta distribution with updated parameters $a_n = a_0 + y$ and $b_n = b_0 + n - y$.

Speaker notes

No notes on this slide.

PAIRED EXERCISE

- Identify values for hyperparameters a_0 and b_0.
- Obtain posterior parameters for $n = 200{,}000$ and $y = 3$.
- Sample from the posterior and estimate posterior mean using R.

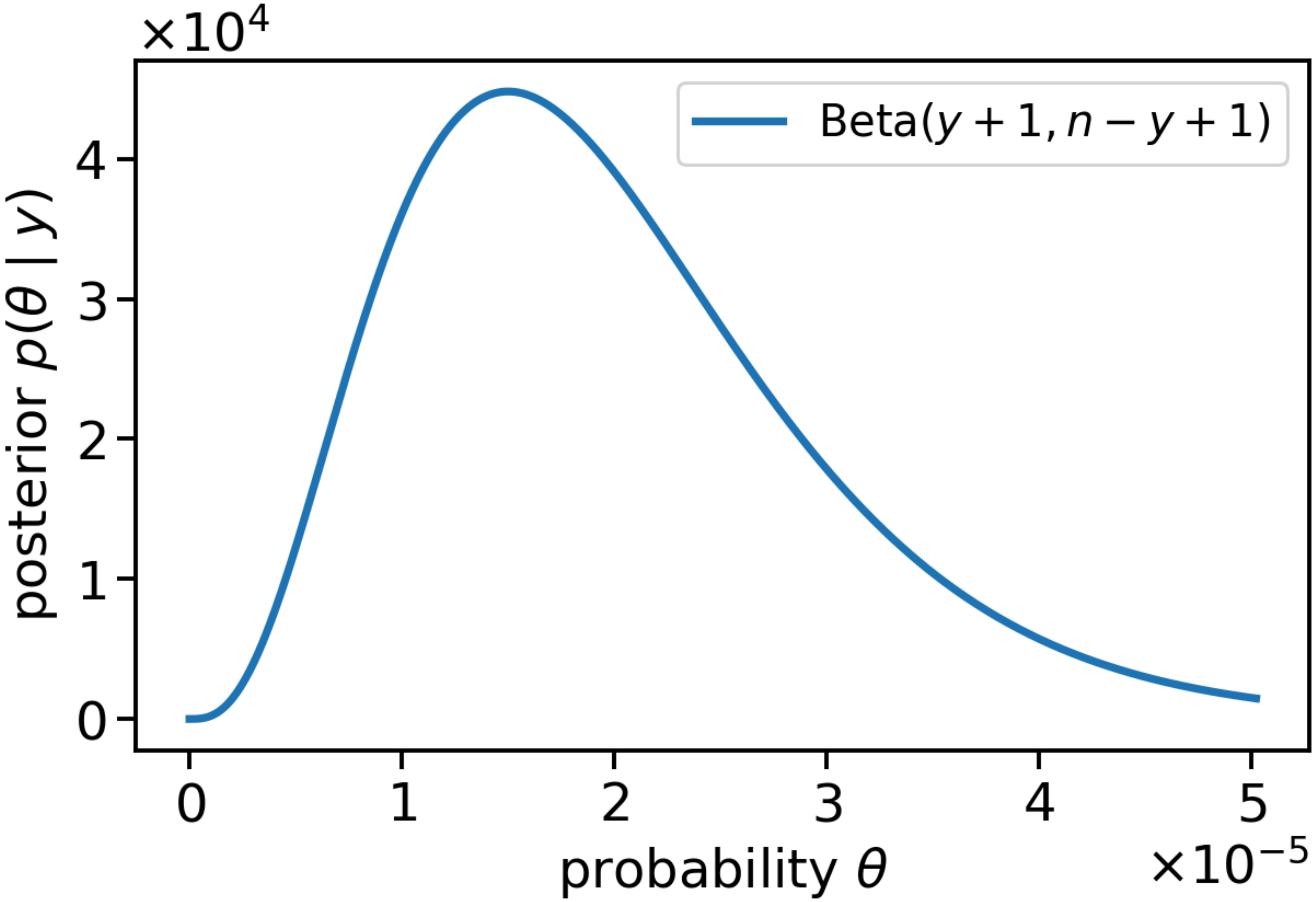
Speaker notes

- Work with your partner and put one of the distributed post-it notes on your laptop when you've finished.
- Upon completion, collect a few answers from students.

```r
1 > # Declare the data and prior hyperparameters.
2 > y <- 3
3 > n <- 200000
4 > a_0 <- 1
5 > b_0 <- 1
6 > # Evaluate posterior parameters.
7 > a_n <- a_0 + y
8 > b_n <- b_0 + n - y
9 > # Sample and report posterior mean.
10 > beta_samples <- rbeta(1000, a_n, b_n)
11 > mean(beta_samples)
12 [1] 1.974689e-05
13 >
```

Speaker notes

- Lines #2-3 declare the data, #4-5 the hyperparameters.
- #7-8 evaluate the parameters of the posterior distribution. This step is only feasible because we have used a conjugate prior.
- #10-11 draw 1,000 samples from the posterior and evaluate the posterior mean.
- Compare responses from students with reference implementation. Why might they differ? Different prior choices, implementation differences?



Speaker notes

- Because we used a conjugate prior, we can plot the posterior in closed form.
- Posterior is consistent with our expectations and is concentrated around the MLE $y/n = 1.5 \times 10^{-5}$.
- Posterior is right-skewed because mortality is bounded below.
- We next consider the same procedure (derive posterior parameters, sample from posterior, inspect posterior) for the Poisson likelihood with *rate* parameter θ.

DERIVATION OF POSTERIOR FOR POISSON LIKELIHOOD

We have the Poisson likelihood and conjugate gamma prior

$$p\left(y \mid \theta, n\right) = \frac{\left(n\theta\right)^y \exp\left(-n\theta\right)}{y!}$$
$$p\left(\theta\right) = \theta^{a_0 - 1} \exp\left(-b_0\theta\right).$$

We used $n\theta$ as the rate for the likelihood because we are interested in the per-capita mortality θ. Neglecting constants in θ, the posterior is

$$p\left(\theta \mid y, n, a_0, b_0\right) \propto \theta^{a_0 + y - 1} \exp\left(-\left[b_0 + n\right]\theta\right)$$

which has the kernel of a gamma distribution. The posterior is thus a gamma distribution with updated parameters $a_n = a_0 + y$ and $b_n = b_0 + n$.

Speaker notes

No notes on this slide.

PAIRED EXERCISE

- Identify values for hyperparameters a_0 and b_0.
- Obtain posterior parameters for $n = 200{,}000$ and $y = 3$.
- Sample from the posterior and estimate posterior mean.
- How does this compare with inference using the binomial likelihood?

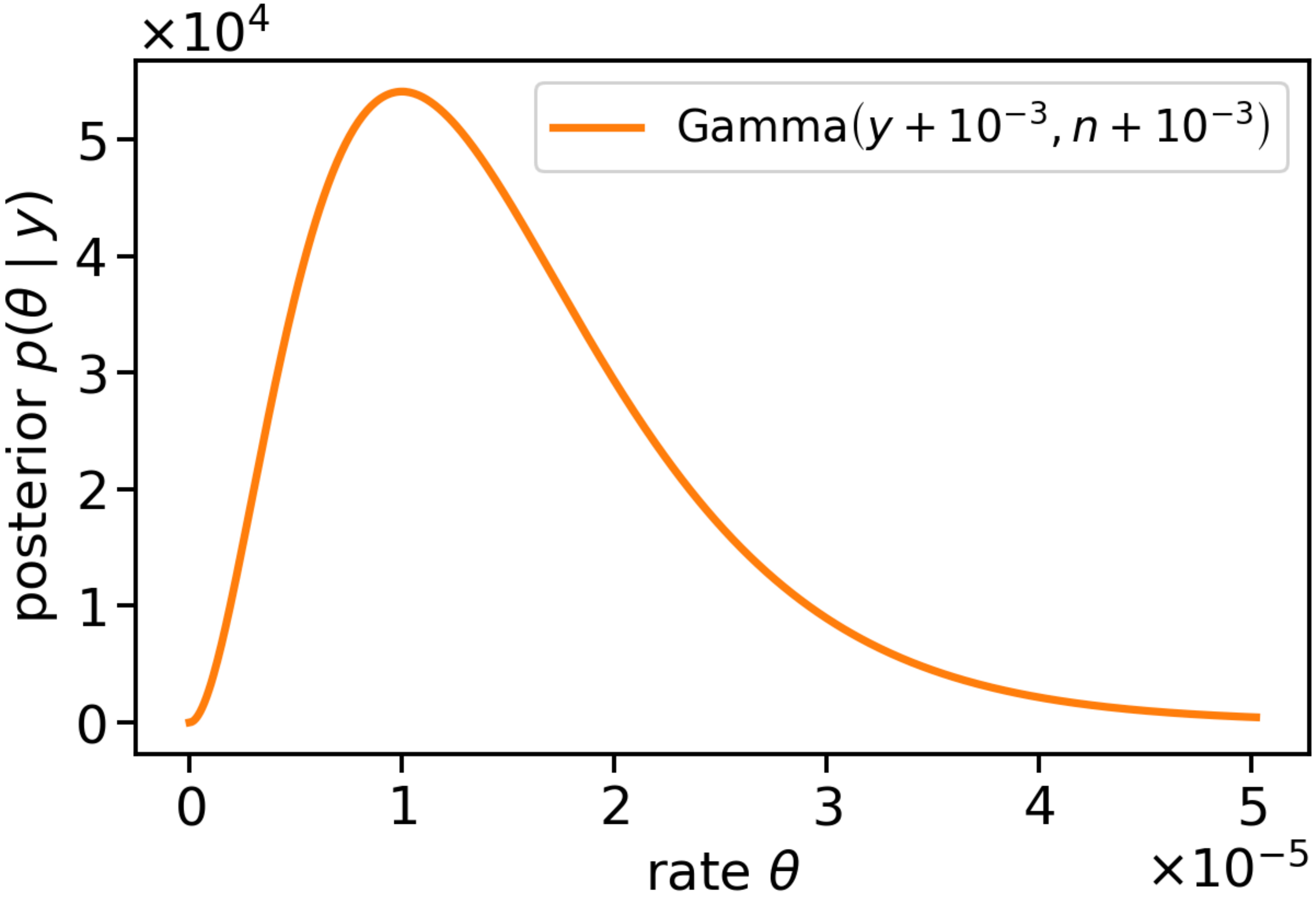
Speaker notes

- Work with your partner and put one of the distributed post-it notes on your laptop when you've finished.
- Upon completion, collect a few answers from students. How do these observations differ from our estimates using the binomial likelihood?
- Why do they differ? Did we use different priors? Is it even meaningful to compare the probability θ with the rate θ given they have different support?

```r
1 > # Declare the data and prior hyperparameters.
2 > y <- 3
3 > n <- 200000
4 > a_0 <- 0.001
5 > b_0 <- 0.001
6 > # Evaluate posterior parameters.
7 > a_n <- a_0 + y
8 > b_n <- b_0 + n
9 > # Sample and report posterior mean.
10 > gamma_samples <- rgamma(1000, a_n, b_n)
11 > mean(gamma_samples)
12 [1] 1.448339e-05
13 >
```

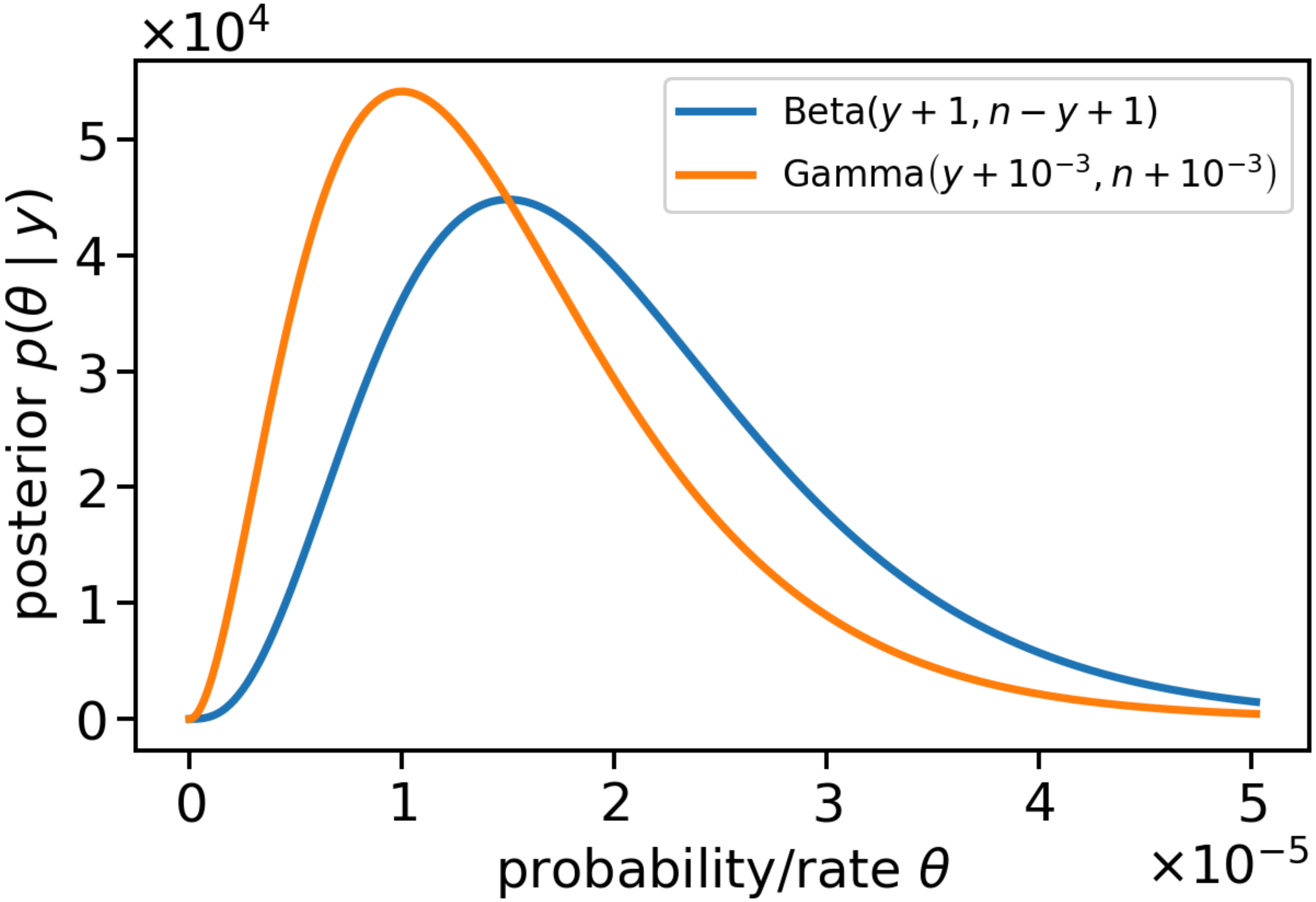
Speaker notes

- Lines #2-5 declare data and hyperparameters again.
- #7-8 evaluate parameters of the posterior.
- #10-11 draw posterior samples and evaluate posterior mean.



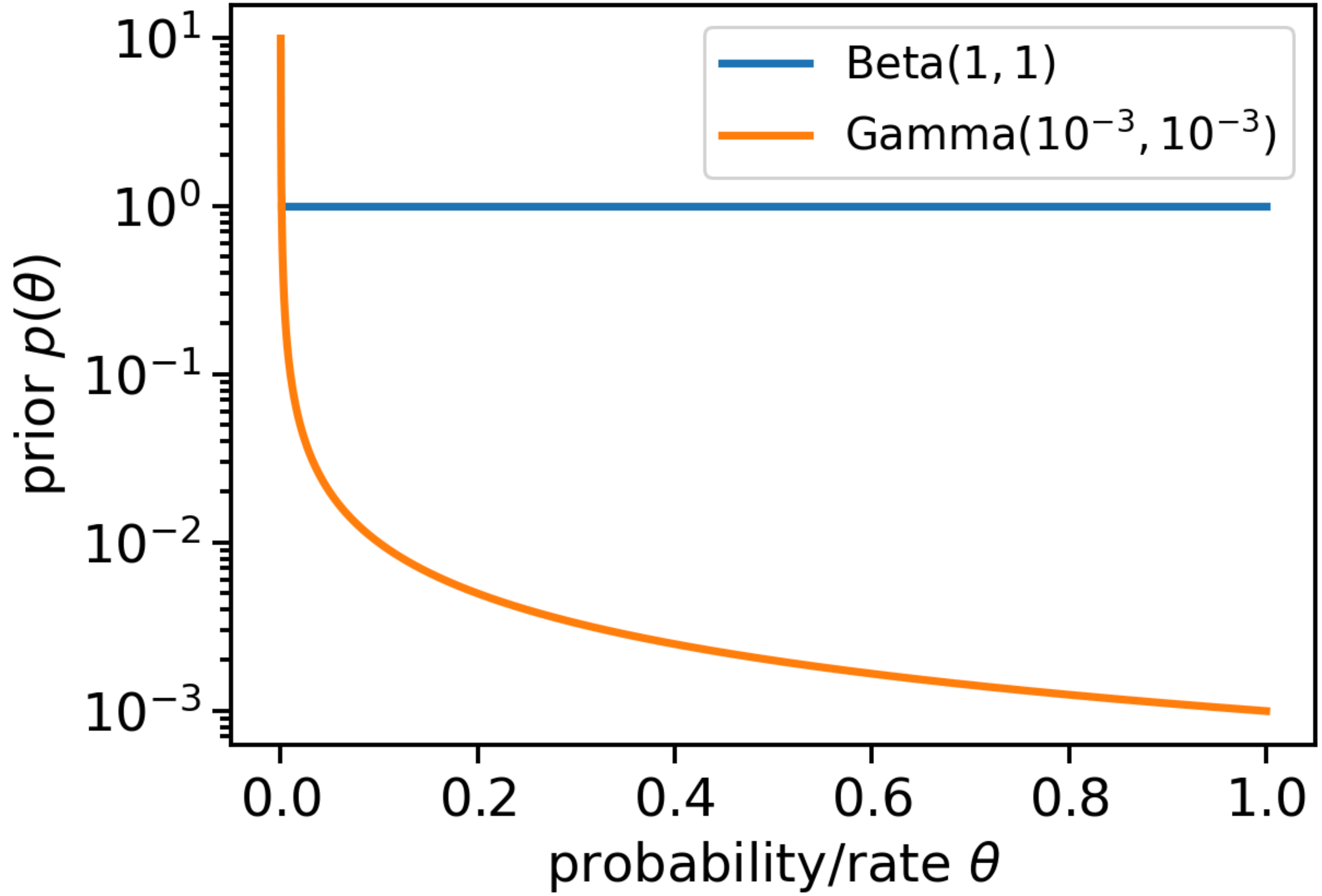
Speaker notes

- The posterior using the Poisson likelihood looks very similar and is also consistent with the MLE.



Speaker notes

- Comparing the two posteriors, they look quite different.
- Beta-binomial model: $\mathbb{E}[\theta] = 2 \times 10^{-5}$.
- Gamma-Poisson model: $\mathbb{E}[\theta] = 1.5 \times 10^{-5}$.
- Posterior mean under beta-binomial model is more than 30% larger than under gamma-Poisson model.
- ❓ Why is that?



Speaker notes

- Let's look at the priors; they are *very* different.
- Gamma prior suggests that we think the mortality is very small.
- Beta prior suggests that we think 80% of the population dying is just as likely as 0.1%.
- But they are both "uninformative". What do we mean by that?
- Really just that the posterior is dominated by the likelihood.
- It does *not* mean that the prior is uninformative in an intuitive sense.
- ❓ Which is "better"?

```r
> # Probability that theta < 1e-6 for beta prior
> # with a = b = 1.
> pbeta(1e-6, 1, 1)
[1] 1e-06
> # Probability that theta < 1e-6 for gamma prior
> # with a = b = 0.001.
> pgamma(1e-6, 1e-3, 1e-3)
[1] 0.9800547
>
```

Speaker notes

- Let's formalize the difference by using the `p[distribution name]` function in R to evaluate the cumulative distribution function of each prior.
- For the flat beta prior, we believe that the mortality is under 10^{-6} with probability 10^{-6}.
- For the gamma prior, we believe that the mortality is under 10^{-6} with probability 0.98.
- These are wildly different prior beliefs leading to different posteriors.
- ❓ Which is better?

WEAKLY INFORMATIVE PRIOR

Chose parameters a and b such that

- $p\left(\theta < 10^{-6}\right) = 0.025$
- and $p\left(\theta < 10^{-3}\right) = 0.975.$

Speaker notes

- Weakly informative priors can better encode our intuition and avoid implicit prior assumptions that affect the posterior.
- One approach to define a weakly informative prior is to match quantiles of the prior to reasonable values.
- Here, we declare that we are pretty confident that mortality is higher than $10^{-6}`. For lower mortalities, we might not see any deaths even in a city five times larger.
- Likewise, we're pretty confident that mortality is smaller than 10^{-3}. In our city, we'd expect to observe 200 deaths at that level.
- ❓ What do you expect the two priors to look like?

PRIOR HYPERPARAMETERS FROM QUANTILES

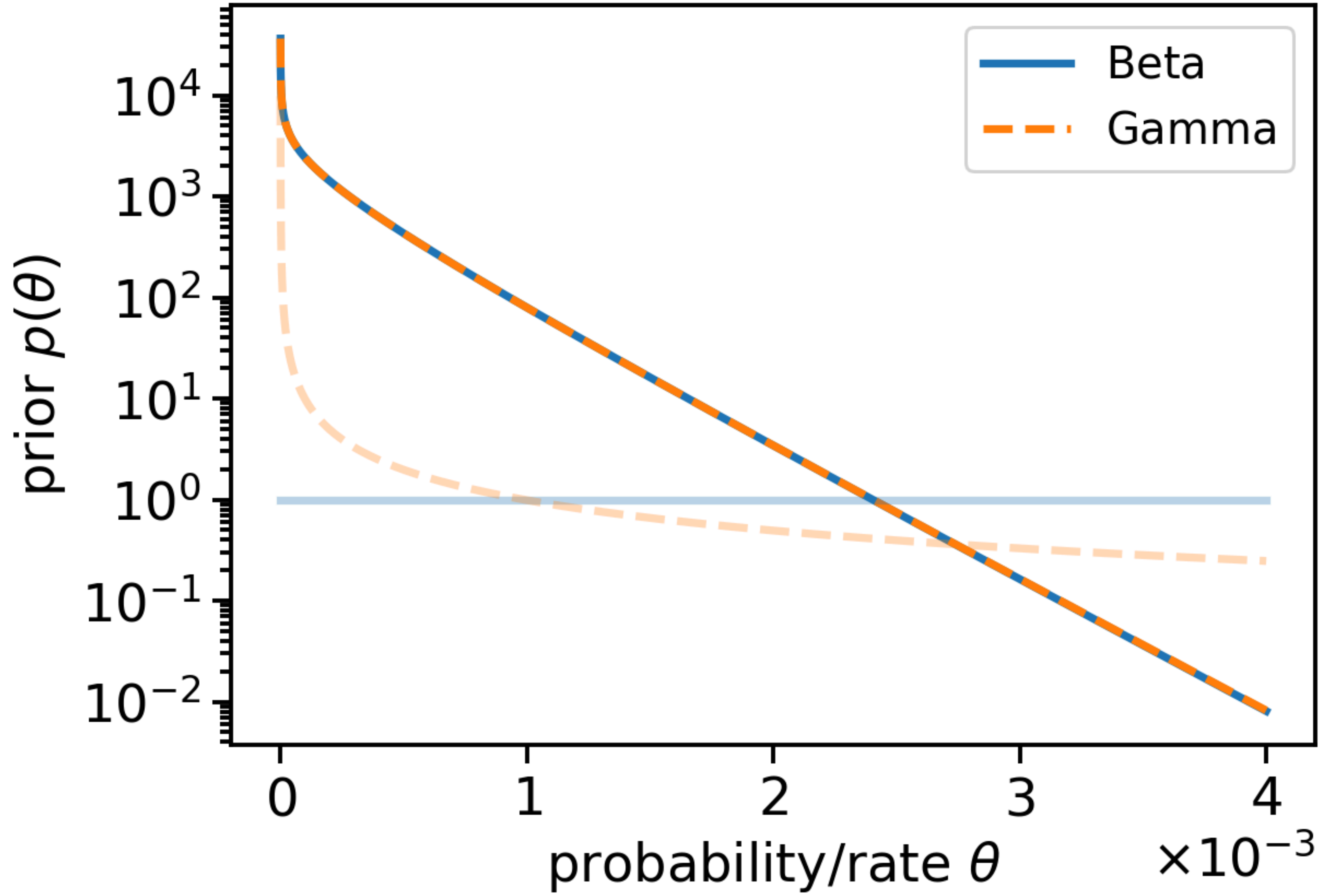
Given two parameter values $\theta_1 < \theta_2$ we seek hyperparameters a^* and b^* such that $f(\theta_1 \mid a, b) = q_1$ and $f(\theta_2 \mid a, b) = q_2$, where f is the cumulative distribution function of the prior and $0 < q_1 < q_2 < 1$. Closed form solutions to this system of equations are not generally available. We can obtain the desired parameters by optimization:

$$(a^*, b^*) = \text{argmin}_{a,b} \left[\left(f(\theta_1 \mid a, b) - q_1\right)^2 + \left(f(\theta_2 \mid a, b) - q_2\right)^2 \right].$$

See `weakly_informative_priors.R` on Canvas for an example implementation.

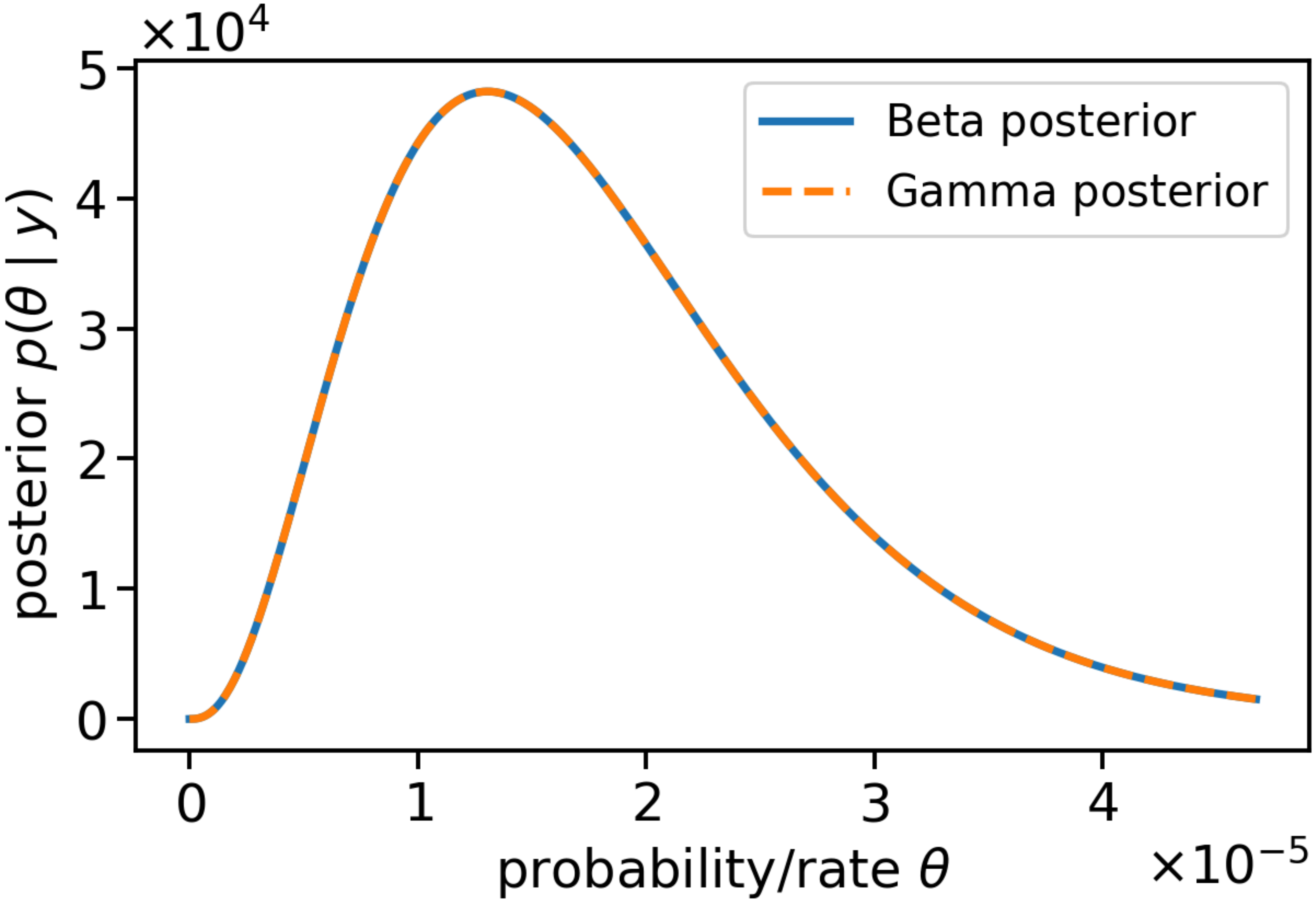
Speaker notes

No notes on this slide.



Speaker notes

- The two weakly informative priors are very similar even though one is a beta distribution and the other a gamma distribution.
- Intuitively, this makes sense because both binomial and Poisson models are suitable models for the data.
- The two “non-informative” priors are shown as semi-transparent lines for reference.



Speaker notes

- Using these priors, the posteriors are also indistinguishable.
- We have been able to resolve this conundrum by taking a formal Bayesian approach and explicitly declaring our priors.
- At 1.8×10^{-5}, the posterior means are a compromise between the two posterior means we obtained using “non-informative” priors. The posteriors remain consistent with the MLE of 1.5×10^{-5}.

RECAP

- Models depend on both data and the scientific question.
- Binomial and Poisson likelihoods have convenient conjugate priors.
- Non-informative priors are informative.
- Explicit prior elicitation can expose implicit assumptions.

NORMAL MODELS

Speaker notes

- Normal models are not just another model. They are the fundamental building blocks of many hierarchical models, state space models, and Gaussian processes for non-parametric regression.
- They can be reasonable even for complex data if they're averages due to central limit theorem.
- We implicitly use normal models whenever we use least-squares regression.
- Depending on the priors for regression parameters, ridge regression and the LASSO arise naturally from regression with normal observation errors.

NORMAL LIKELIHOOD (1 / 2)

The likelihood for mean μ and scale σ is

$$p\left(y \mid \mu, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right).$$

Speaker notes

- Normal models have two parameters: location and scale. One encodes where the distribution is centered, the other how dispersed it is.
- We will first infer each parameter assuming the other is known and then consider the common scenario where both are unknown.
- Norma models have light tails because the density decays as exponential of squared distance. This means they are not robust to outliers–just like least squares regression.

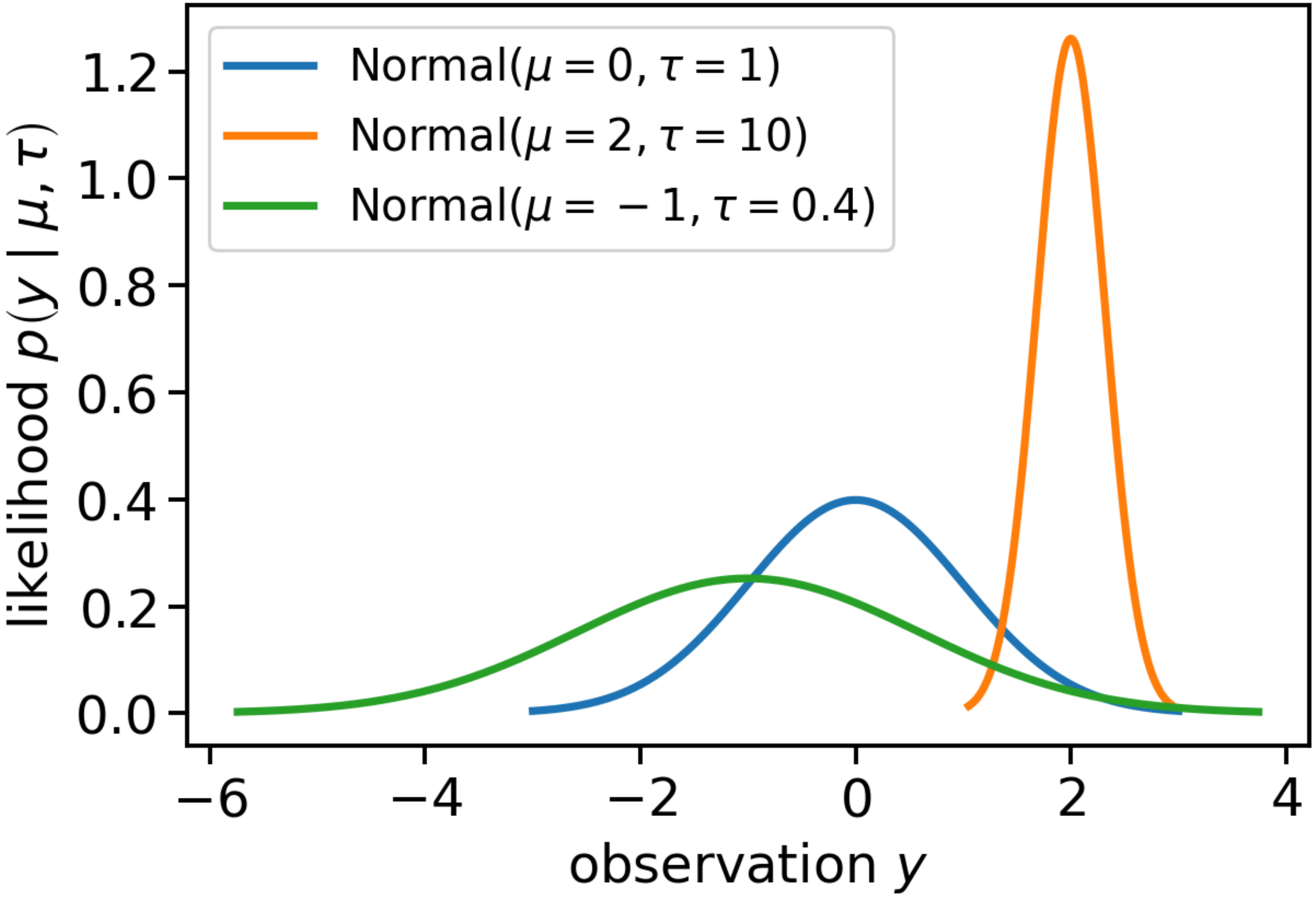
NORMAL LIKELIHOOD (2 / 2)

Algebra is *much* easier using the precision $\tau = \sigma^{-2}$, yielding

$$p\left(y \mid \mu, \tau\right) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau\left(y-\mu\right)^2}{2}\right).$$

Speaker notes

- The precision τ is just what it sounds like. It encodes how precisely observations y follow the location parameter μ.
- In an inference setting, τ quantifies how precisely data can inform the location parameter μ.



Speaker notes

- Figure shows examples of normal densities with different parameters.
- Higher precision means more concentrated densities.
- Blue is the standard normal distribution (i.e., zero mean, unit variance).

INDEPENDENT OBSERVATIONS

For n independent observations $\mathbf{y}$, the likelihood is

$$p\left(\mathbf{y} \mid \mu, \tau\right) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau \sum_{i=1}^{n} \left(y_i - \mu\right)^2}{2}\right).$$

Speaker notes

- We use lower-case bold font to denote a vector.

DERIVATION OF NORMAL LIKELIHOOD FOR I.I.D. OBSERVATIONS

The likelihood of n i.i.d. observations is the product of individual likelihoods

$$p\left(\mathbf{y} \mid \mu, \tau\right) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau\left(y_i - \mu\right)^2}{2}\right)$$

The $\sqrt{\frac{\tau}{2\pi}}$ term does not depend on the index i and contributes a constant $\left(\frac{\tau}{2\pi}\right)^{n/2}$. We express the product of exponentials as the exponential of a sum to obtain

$$p\left(\mathbf{y} \mid \mu, \tau\right) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau \sum_{i=1}^{n}\left(y_i - \mu\right)^2}{2}\right).$$

💡 Working with log probabilities is often preferable to working with probabilities directly. The latter can lead to underflows and overflows due to multiplication of many small and large numbers, respectively.

Speaker notes

No notes on this slide.

INFERRING μ FOR KNOWN τ

Speaker notes

- We may want to infer the concentration μ of a chemical with an instrument with known precision τ, e.g., the instrument manufacturer may provide the measurement error.
- To make analytic progress with inference, we next derive the conjugate prior for the location parameter μ.

KERNEL FOR μ UNDER NORMAL LIKELIHOOD WITH KNOWN τ

Consider the posterior (neglecting constants in μ)

$$\begin{aligned} p(\mu \mid \mathbf{y}, \tau) &\propto p(\mu) \exp\left(-\frac{\tau \sum_{i=1}^{n} (y_i - \mu)^2}{2}\right), \\ &\propto p(\mu) \exp\left(-\frac{\tau}{2} \sum_{i=1}^{n} \left(y_i^2 - 2\mu y_i + \mu^2\right)\right), \end{aligned}$$

where we have expanded the square in the second line. We drop the y_i^2 term and distribute the sum to obtain

$$p(\mu \mid \mathbf{y}, \tau) \propto p(\mu) \exp\left(-\frac{n\tau}{2} \left(\mu^2 - 2\mu\bar{y}\right)\right),$$

where $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ is the sample mean.

💡 The quadratic form in the exponential looks suspiciously like the kernel of a normal distribution in μ, and we use a normal prior to derive the posterior.

Speaker notes

No notes on this slide.

POSTERIOR FOR μ UNDER NORMAL LIKELIHOOD WITH KNOWN τ

The posterior given a normal prior $p\left(\mu \mid \nu_0, \kappa_0\right)$ with prior mean μ_0 and precision κ_0 is

$$p\left(\mu \mid \mathbf{y}, \tau\right) \propto \exp\left(-\frac{\kappa_0}{2}\left(\mu^2 - 2\mu\nu_0\right)\right) \exp\left(-\frac{n\tau}{2}\left(\mu^2 - 2\mu\bar{y}\right)\right),$$

where we have expanded the square in the exponential of the prior. Combining the exponentials and collecting terms in μ and μ^2 yields

$$\begin{aligned} p\left(\mu \mid \mathbf{y}, \tau\right) &\propto \exp\left(-\frac{(\kappa_0 + n\tau)\mu^2 - 2\mu(\kappa_0\nu_0 + n\tau\bar{y})}{2}\right) \\ &\propto \exp\left(-\frac{\kappa_0 + n\tau}{2}\left(\mu^2 - 2\mu\frac{\kappa_0\nu_0 + n\tau\bar{y}}{n\tau + \kappa_0}\right)\right). \end{aligned}$$

Comparing with the functional form of a normal distribution, we find that the posterior has mean $\nu_n = \frac{\kappa_0\nu_0 + n\tau\bar{y}}{\kappa_0 + n\tau}$ and precision $\kappa_n = \kappa_0 + n\tau$.

Speaker notes

No notes on this slide.

Update rules for μ posterior parameters for known precision are

$$\nu_n = \frac{\kappa_0\nu_0 + n\tau\bar{y}}{\kappa_0 + n\tau},$$
$$\kappa_n = \kappa_0 + n\tau.$$

Speaker notes

- The posterior mean ν_n is the average of the prior mean ν_0 and sample mean $\bar{y}$ weighted by the prior and likelihood precisions.
- The more data we observe (increasing n) or the more precise the observations (increasing τ), the closer the posterior mean is to the sample mean.
- For large n, the posterior variance $\kappa_n^{-1} \propto n^{-1}$, and we recover the familiar square-root scaling of the standard error.