Introduction to Machine Learning

Week 5 - Normal/Gaussian Distribution

Spring 2025
Instructor: Dr. Patrick Skeba

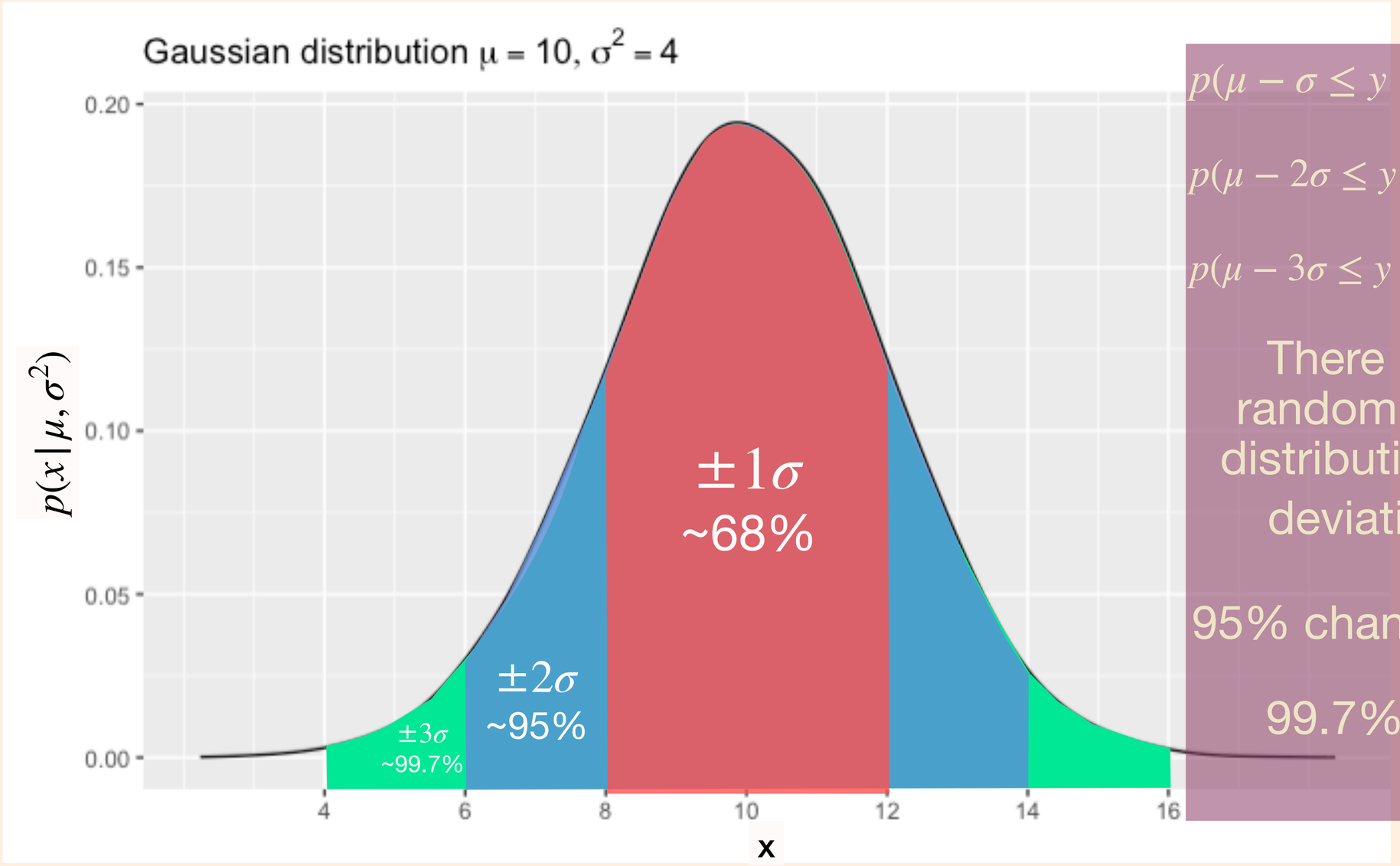
The normal or Gaussian distribution is one of the most important probability distributions in statistics and machine learning

x, μ are both unbounded continuous values, and σ is a positive real number

$$\texttt{Normal}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

- 2 parameters:
  - $\mathbb{E}[x] = \mu \in \mathbb{R}$: the expected value, or mean of the distribution. For Gaussian distributions, the mean, mode, and median are all equivalent. The value of μ can be any continuous value from negative to positive infinity
  - $\texttt{var}(x) = \sigma^2 \in \mathbb{R} > 0$: the variance of the distribution. Defines how "spread out" the distribution is about its mean. σ is known as the "standard deviation" and must be a continuous value greater than 0

The Empirical Rule - 68/95/99.7%



$p(\mu - \sigma \leq y \leq \mu + \sigma) = p(8 \leq y \leq 12) \approx 0.68$

$p(\mu - 2\sigma \leq y \leq \mu + 2\sigma) = p(6 \leq y \leq 14) \approx 0.95$

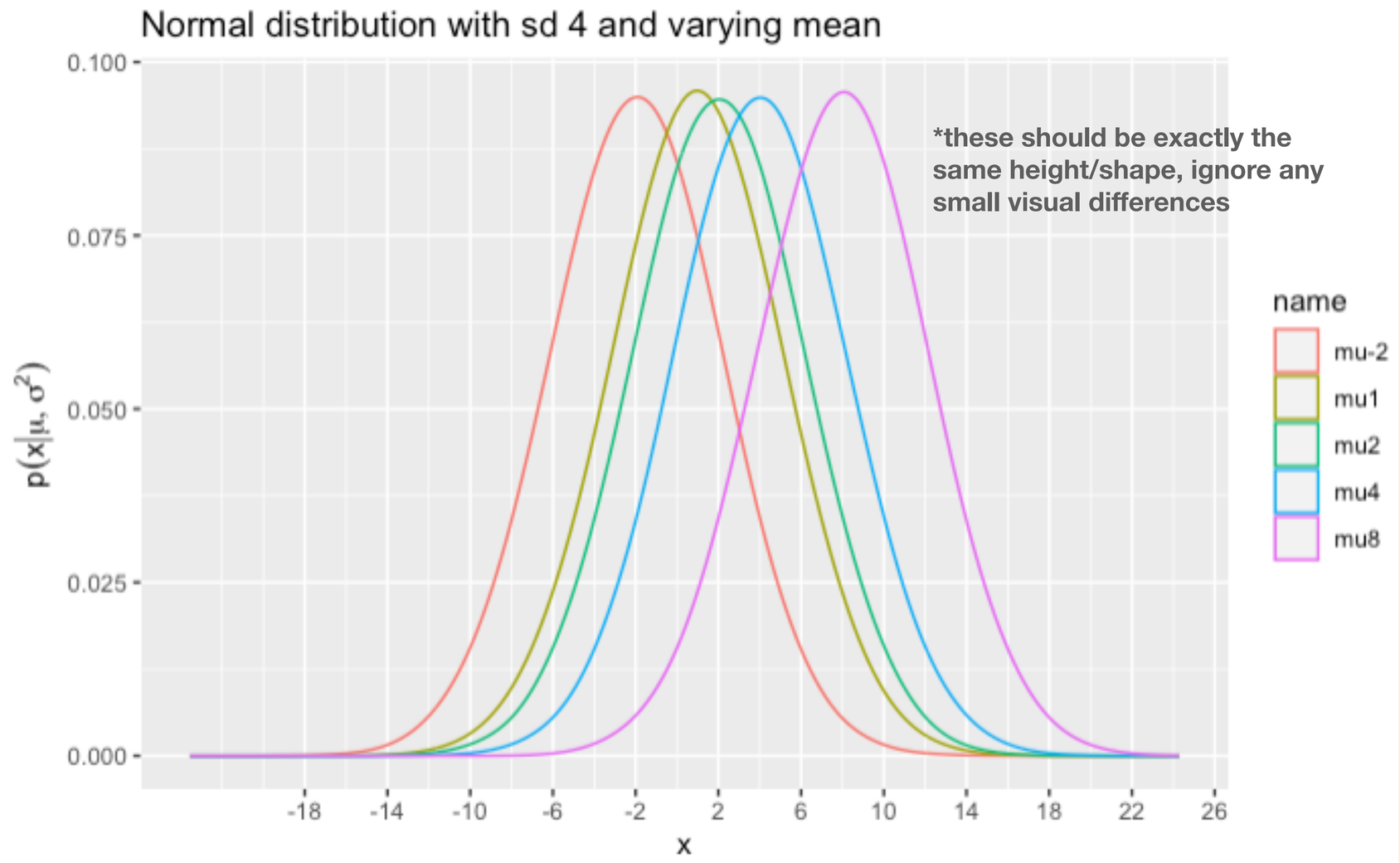
$p(\mu - 3\sigma \leq y \leq \mu + 3\sigma) = p(4 \leq y \leq 16) \approx 0.997$

There is a 68% chance that a random number drawn from this distribution is within one standard deviation of the mean ($\mu \pm \sigma$)

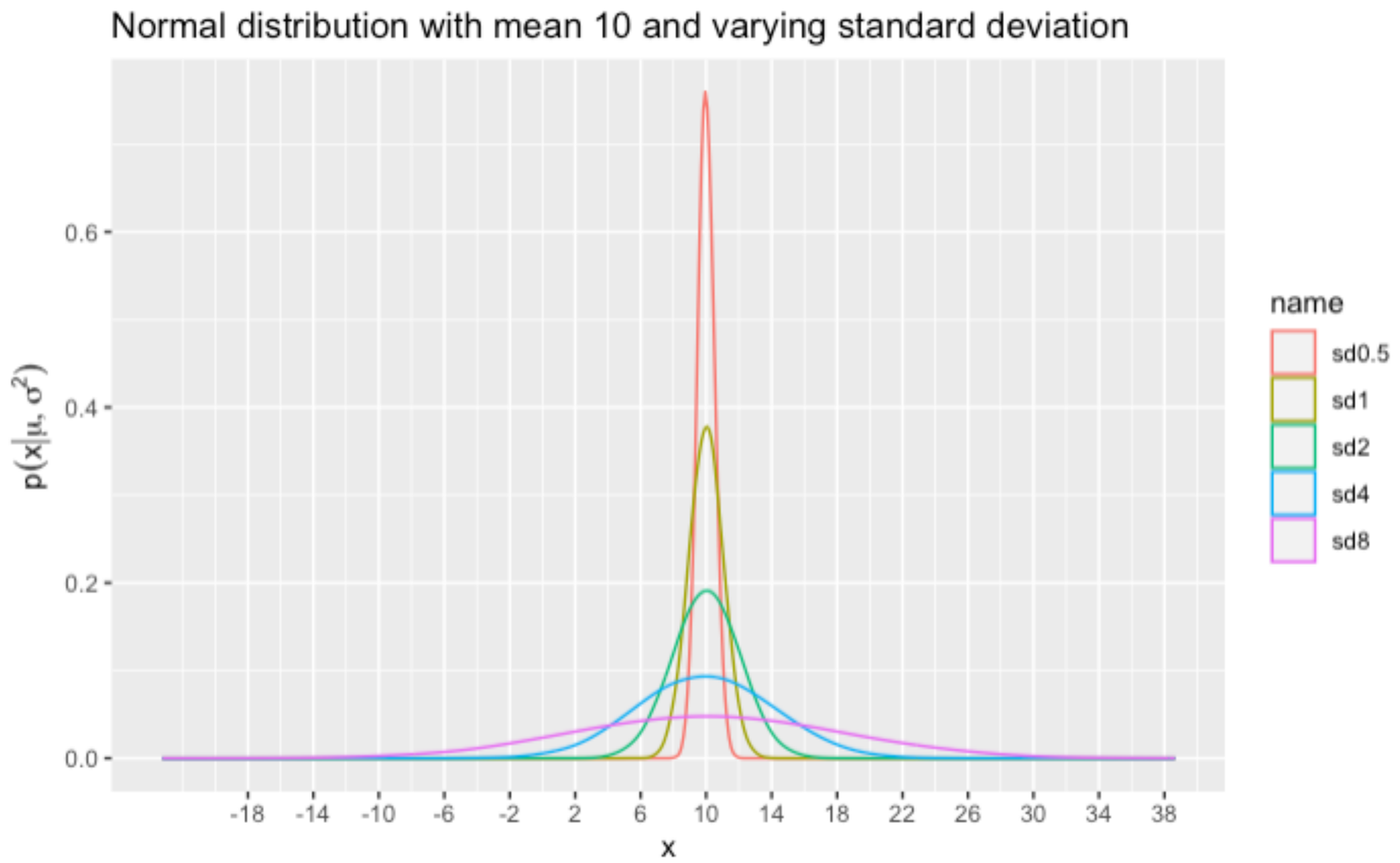
95% chance it's within 2 sd ($\mu \pm 2\sigma$)

99.7% chance it's within 3 sd ($\mu \pm 3\sigma$)

The mean μ defines the center or peak (mode) of the normal distribution. It's the *expected value*



The variance/standard deviation defines the width of the distribution, or the range of expected values



- When σ is very small, most of the probability density is concentrated around the mean.
  - For $\sigma = 0.5$, there is a 95% chance that a random x is between $10 - 2 \times (0.5) = 9$ and $10 + 2 \times (0.5) = 11$
- When σ is large, x is expected to fall in a wider range around the mean
  - For $\sigma = 8$, there is a 95% chance that a random x is between $10 - 2 \times (8) = -6$ and $10 + 2 \times (8) = 26$

For regression problems, we'll use a Gaussian Likelihood

Assume each sample is drawn from a normal distribution with mean μ and variance σ^2

- $$p(\mathbf{x}\,|\,\mu,\sigma^2) = \prod_{n=1}^{N} \texttt{Normal}(x_n\,|\,\mu,\sigma^2)$$

$$\log p(\mathbf{x}\,|\,\mu,\sigma^2) = \sum_{n=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_n-\mu)^2}{2\sigma^2}\right\}\right]$$

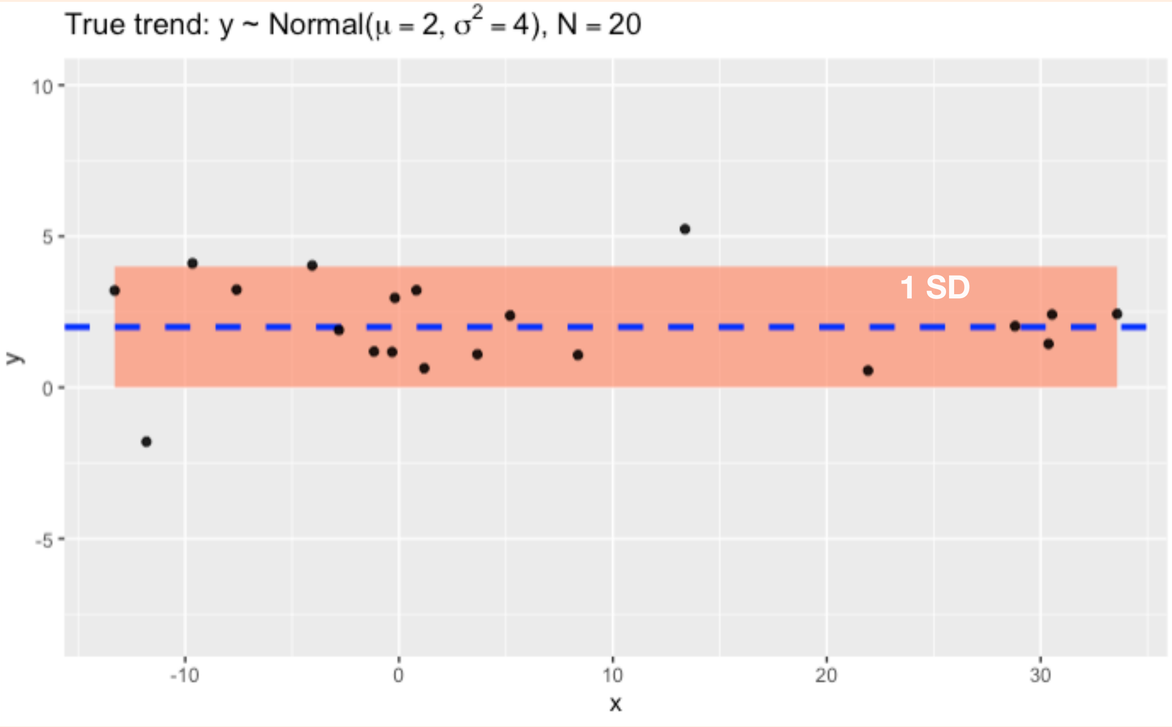
$$= \sum_{n=1}^{N}\left\{-\frac{(x_n-\mu)^2}{2\sigma^2} - \log\left[\sqrt{2\pi\sigma^2}\right]\right\}$$

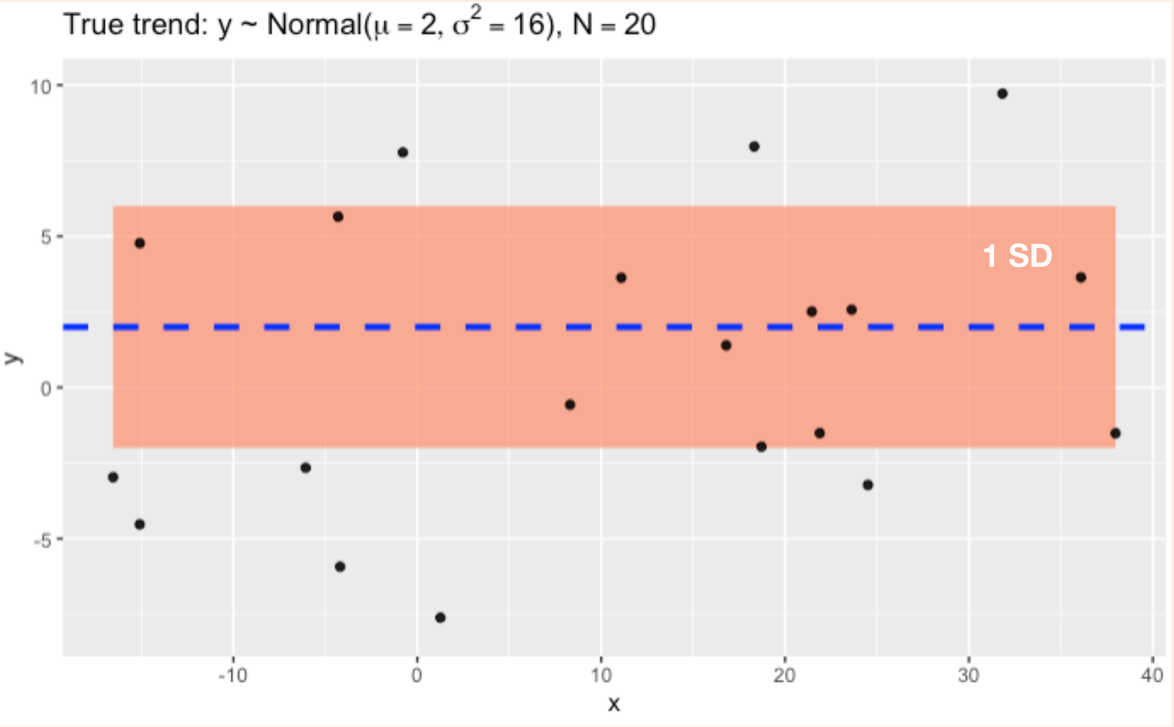
- $$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - N\log\left[\sqrt{2\pi\sigma^2}\right]$$

The two parameters of the likelihood represent two distinct pieces of information we are trying to infer about the system

μ : what is the most likely value for the response,

σ^2 : how far away from the mean we expect observations to lie (measurement noise)





- The data on the right has the same mean as the left, but a larger variance. Ultimately we will construct linear models that estimate both the mean trend and the measurement noise.
- This is NOT the confidence/predictive interval! It is a (hidden) parameter of the data which we are trying to estimate.

Let's start working with the Gaussian distribution

- Let's weigh a dumbbell N times, producing a sequence of observed weights $\mathbf{x} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$.

- Each x_n is the weight in pounds.

- We will use a 25 pound dumbbell.

We will assume a Gaussian distribution is an appropriate probability model for this situation

- Assume each observation is **conditionally independent** of the others **given** the parameters, μ and σ.

- The joint distribution of all N observations can therefore be factored into the product of N separate likelihoods:

$$p(x_1, \ldots, x_n, \ldots, x_N \mid \mu, \sigma) = p(\mathbf{x}|\mu, \sigma) = \prod_{n=1}^{N} \{p(x_n|\mu, \sigma)\}$$

Likelihood of the n-th observations

- The n-th observation is assumed to be normally distributed **conditioned** on knowing the μ and σ parameters:

$$x_n|\mu,\sigma \sim \text{normal}(x_n|\mu,\sigma)$$

What do the μ and σ represent in this context?

- μ is the mean of the population of all possible weight realizations.

- σ represents the noise or error of the process.
  - All measuring scales have some level of variation. Essentially, σ represents the **lack of** repeatability in the measurement process.
  - Think of shifting your weight around on a scale at home. Changing your balance may change the measurement.

In our dumbbell experiment

- We feel confident that the weight of a dumbbell is very close to the reported value on the weight.

- After doing some research, we find that the manufacturer of the measuring scale states that $\sigma = 1$ pound.

- **Given that we feel weighing a dumbbell is highly repeatable, let's assume that σ is known!**

How can estimate we μ?

- We will start out just as we did with the Bernoulli case by considering the Maximum Likelihood Estimate (MLE) on the unknown mean.

$$\hat{\mu} = \mu_{MLE} = \arg\max p(\mathbf{x} \mid \mu, \sigma)$$

How can estimate we μ?

- We will start out just as we did with the Bernoulli case by considering the Maximum Likelihood Estimate (MLE) on the unknown mean.

$$\hat{\mu} = \mu_{MLE} = \arg\max p(\mathbf{x} \mid \mu, \sigma)$$

- As before rather than working with the likelihood itself, we work with the log-likelihood:

$$\hat{\mu} = \mu_{MLE} = \arg\max \log[p(\mathbf{x} \mid \mu, \sigma)]$$

Log-likelihood of the N conditionally independent observations

$$\log[p(\mathbf{x} \mid \mu, \sigma)] = \sum_{n=1}^{N} \{\log[\text{normal}(x_n \mid \mu, \sigma)]\}$$

Log-likelihood of the N conditionally independent observations

$$\log[p(\mathbf{x} \mid \mu, \sigma)] = \sum_{n=1}^{N} \{\log[\text{normal}(x_n \mid \mu, \sigma)]\}$$

Write out the Gaussian pdf

$$\log[p(\mathbf{x} \mid \mu, \sigma)] = \sum_{n=1}^{N} \left(\log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_n - \mu)^2 \right) \right] \right)$$

The MLE is just the sample average!

$$\mu_{MLE} = \bar{x}$$

Where the sample average is:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N}(x_n)$$

Now let's consider the Bayesian approach to estimating the unknown μ

- Remember that the Bayesian way does not find a **point estimate**, but rather a **probability distribution**!

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Bayesian formulation for the unknown μ

- The posterior distribution on μ given $\mathbf{x}$ and σ is proportional to:

$$p(\mu|\mathbf{x},\sigma) \propto \prod_{n=1}^{N}\{\text{normal}(x_n|\mu,\sigma)\} \times p(\mu)$$

- How can we encode our prior belief on μ?

Conjugate prior to a normal likelihood

- If the prior has the same functional form as the likelihood, the posterior will be of the same distribution family as the prior!

- The conjugate prior to the normal is…

Conjugate prior to a normal likelihood

- If the prior has the same functional form as the likelihood, the posterior will be of the same distribution family as the prior!

- The conjugate prior to the normal is…a normal!

$$\mu|\mu_0,\tau_0 \sim \text{normal}(\mu \mid \mu_0,\tau_0)$$

The posterior on μ will therefore be a normal distribution!

$$p(\mu|\mathbf{x},\sigma) \propto \prod_{n=1}^{N}\{\text{normal}(x_n|\mu,\sigma)\} \times \text{normal}(\mu|\mu_0,\tau_0)$$

The posterior on μ will therefore be a normal distribution!

$$p(\mu|\mathbf{x},\sigma) \propto \prod_{n=1}^{N}\{\text{normal}(x_n|\mu,\sigma)\} \times \text{normal}(\mu|\mu_0,\tau_0)$$

How can we derive a normal from that???

Start by writing out all terms involving the unknown parameter

$$p(\mu|\mathbf{x},\sigma) \propto \prod_{n=1}^{N}\left\{\exp\left(-\frac{1}{2\sigma^2}(x_n-\mu)^2\right)\right\} \times \exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right)$$

Apply the natural log

$$\log[p(\mu|\mathbf{x},\sigma)] \propto -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{(x_n-\mu)^2\} - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2$$

We will not step through the complete derivation

- Doing so requires a fair amount of algebra and **completing the square**.

- However, let's highlight some important aspects of the derivation.

Precision

- Precision is defined as the inverse of the variance:

$$\frac{1}{\sigma^2}, \frac{1}{\tau_0^2}$$

- The posterior mean will be a **precision weighted** combination of the prior mean and the data.

Sufficient statistic

- The sample (empirical) mean $\bar{x}$ is the **sufficient statistic** for the posterior on μ.

- Sufficient represents that the "information content" of the observations can be represented by the sample mean!

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Posterior distribution on μ given $\mathbf{x}$ and σ

$$p(\mu|\mathbf{x},\sigma) = \text{normal}(\mu|\mu_N,\tau_N)$$

- The posterior standard deviation is defined in terms of the posterior precision:

$$\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2}$$

The posterior precision is the sum of the prior precision and the data precision!

Posterior distribution on μ given $\mathbf{x}$ and σ

$$p(\mu|\mathbf{x},\sigma) = \text{normal}(\mu|\mu_N,\tau_N)$$

- The posterior mean is:

$$\mu_N = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{N}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}}$$

Posterior distribution on μ given $\mathbf{x}$ and σ

$$p(\mu|\mathbf{x},\sigma) = \text{normal}(\mu|\mu_N,\tau_N)$$

- The posterior mean is:

$$\mu_N = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{N}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}}$$

All observations are aggregated into a single "sufficient observation" of the sample mean!

Posterior distribution on μ given $\mathbf{x}$ and σ

$$p(\mu|\mathbf{x},\sigma) = \text{normal}(\mu|\mu_N,\tau_N)$$

- The posterior mean is:

The weights are proportional to the precisions!

$$\mu_N = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{N}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}}$$

The posterior mean is a **weighted average** of the prior and data.

Rearrange the posterior mean

- To see the weighting or combination more explicitly:

$$\mu_N = \left(\frac{1/\tau_0^2}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}} \right) \times \mu_0 + \left(\frac{N/\sigma^2}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}} \right) \times \bar{x}$$

Rearrange the posterior mean

- The posterior mean can be written as "**adjusting**" the prior mean toward the **sample mean**:

$$\mu_N = \mu_0 + (\bar{x} - \mu_0)\left(\frac{\tau_0^2 N}{\sigma^2 + \tau_0^2 N}\right)$$

- The posterior mean can be written as the sample mean (the data) "**shrunk**" toward the **prior mean**:

$$\mu_N = \bar{x} - (\bar{x} - \mu_0)\left(\frac{\sigma^2}{\sigma^2 + \tau_0^2 N}\right)$$

What can we control?

- When we run an experiment, we can set the sample size N.
  - What happens in the limit of many samples $N \to \infty$?

- We also set our prior belief.
  - What if we test out the **sensitivity** of the posterior to our prior standard deviation, τ_0?
  - Specifically, what happens as $\tau_0 \to \infty$?

$$N \to \infty$$

Posterior mean, μ_N

$$\mu_N = \bar{x} - (\bar{x} - \mu_0)\left(\frac{\sigma^2}{\sigma^2 + \tau_0^2 N}\right)$$

$$\mu_N \to \bar{x} - (\bar{x} - \mu_0) \cdot (0) \to \bar{x}$$

Posterior standard deviation, τ_N

$$\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \Longrightarrow \frac{1}{\tau_N^2} \to \infty$$

$$\tau_N \to 0$$

$$N \to \infty$$

Posterior mean, μ_N

$$\mu_N = \bar{x} - (\bar{x} - \mu_0)\left(\frac{\sigma^2}{\sigma^2 + \tau_0^2 N}\right)$$

$$\mu_N \to \bar{x} - (\bar{x} - \mu_0) \cdot (0) \to \bar{x}$$

Posterior standard deviation, τ_N

$$\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \Longrightarrow \frac{1}{\tau_N^2} \to \infty$$

$$\tau_N \to 0$$

In the limit of infinite sample size, the posterior converges to an infinitely precise (zero variance) **Gaussian centered on the sample average!**

$$\tau_0 \to \infty$$

Posterior mean, μ_N

$$\mu_N = \bar{x} - (\bar{x} - \mu_0)\left(\frac{\sigma^2}{\sigma^2 + \tau_0^2 N}\right)$$

$$\mu_N \to \bar{x} - (\bar{x} - \mu_0) \cdot (0) \to \bar{x}$$

Posterior standard deviation, τ_N

$$\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \Longrightarrow \frac{1}{\tau_N^2} \to \frac{N}{\sigma^2}$$

$$\tau_N \to \frac{\sigma}{\sqrt{N}}$$

$$\tau_0 \to \infty$$

Posterior mean, μ_N

$$\mu_N = \bar{x} - (\bar{x} - \mu_0)\left(\frac{\sigma^2}{\sigma^2 + \tau_0^2 N}\right)$$

$$\mu_N \to \bar{x} - (\bar{x} - \mu_0) \cdot (0) \to \bar{x}$$

Posterior standard deviation, τ_N

$$\frac{1}{\tau_N^2} = \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \Longrightarrow \frac{1}{\tau_N^2} \to \frac{N}{\sigma^2}$$

$$\tau_N \to \frac{\sigma}{\sqrt{N}}$$

In the limit of an infinitely **diffuse** prior, the posterior converges to:

$$p(\mu|x, \sigma) \to \text{normal}\left(\bar{x}, \sigma/\sqrt{N}\right)$$

Both asymptotic situations reveal the link between the Bayesian formulation and the classic solution

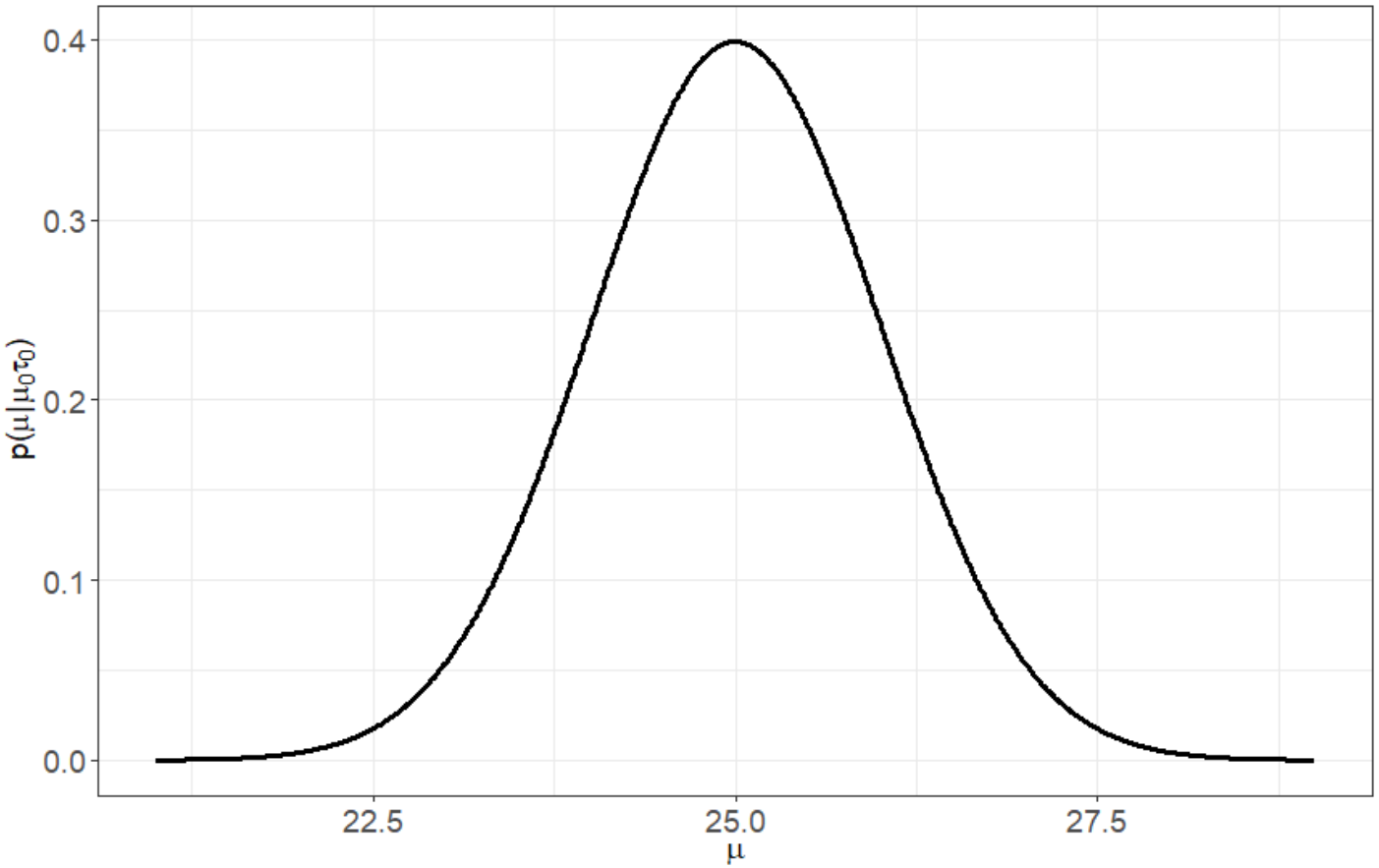
- The Bayesian result is a **compromise** between the prior and the sample average (the data).

- The data overwhelms our prior beliefs in the limit of:
  - Infinitely many observations
  - Initially, we are infinitely uncertain

Prior specification – **informative prior**

- We will use a 25-pound dumbbell in our experiment.
- A priori we feel that the manufacturer's weight should be precise.
- Define the prior to be informative around 25-pounds

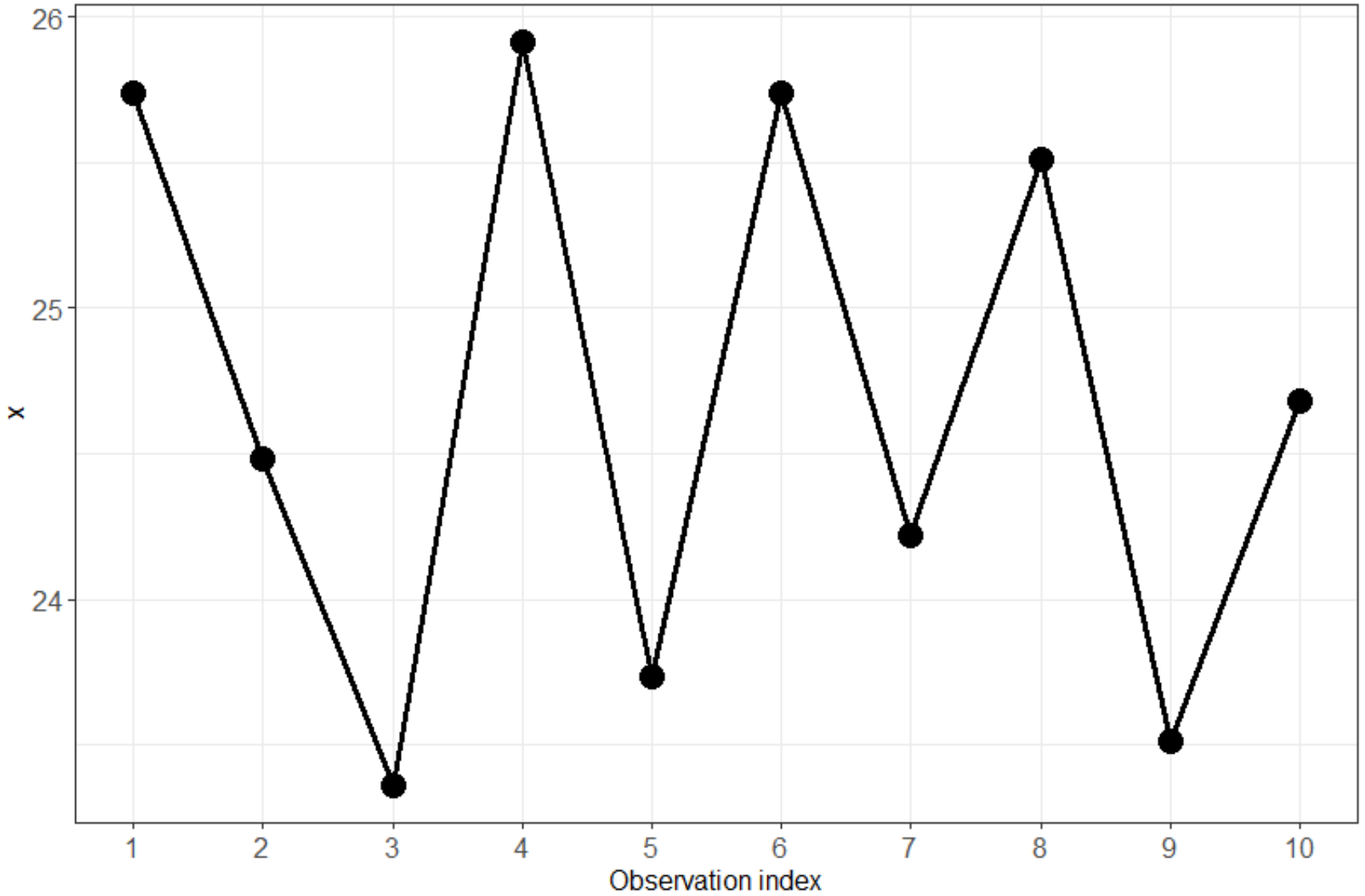
$$\mu|\mu_0, \tau_0 \sim \text{normal}(\mu|25, 1)$$



We performed the experiment measuring the weight of the dumbbell up to 1000 times!!!

- Calculate the posterior distribution on μ after each observation based on:
- Our assumed $\sigma = 1$-pound value
- The defined prior with $\mu_0 = 25$ and $\tau_0 = 1$.

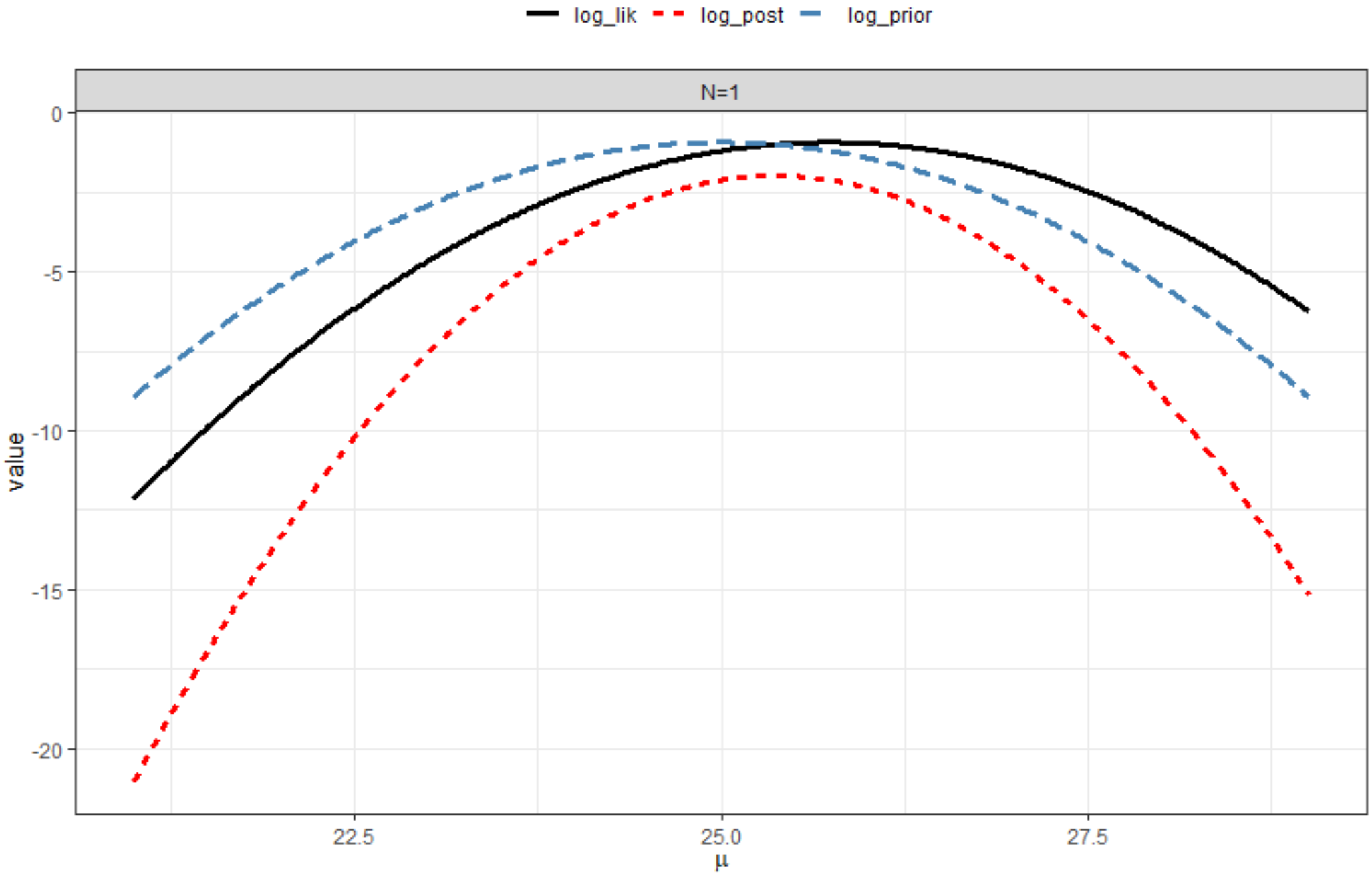
The first 10 data points displayed as a “run style” chart



How does our belief change after the first observation?

- Compare the posterior, likelihood, and prior distributions in the log-space.

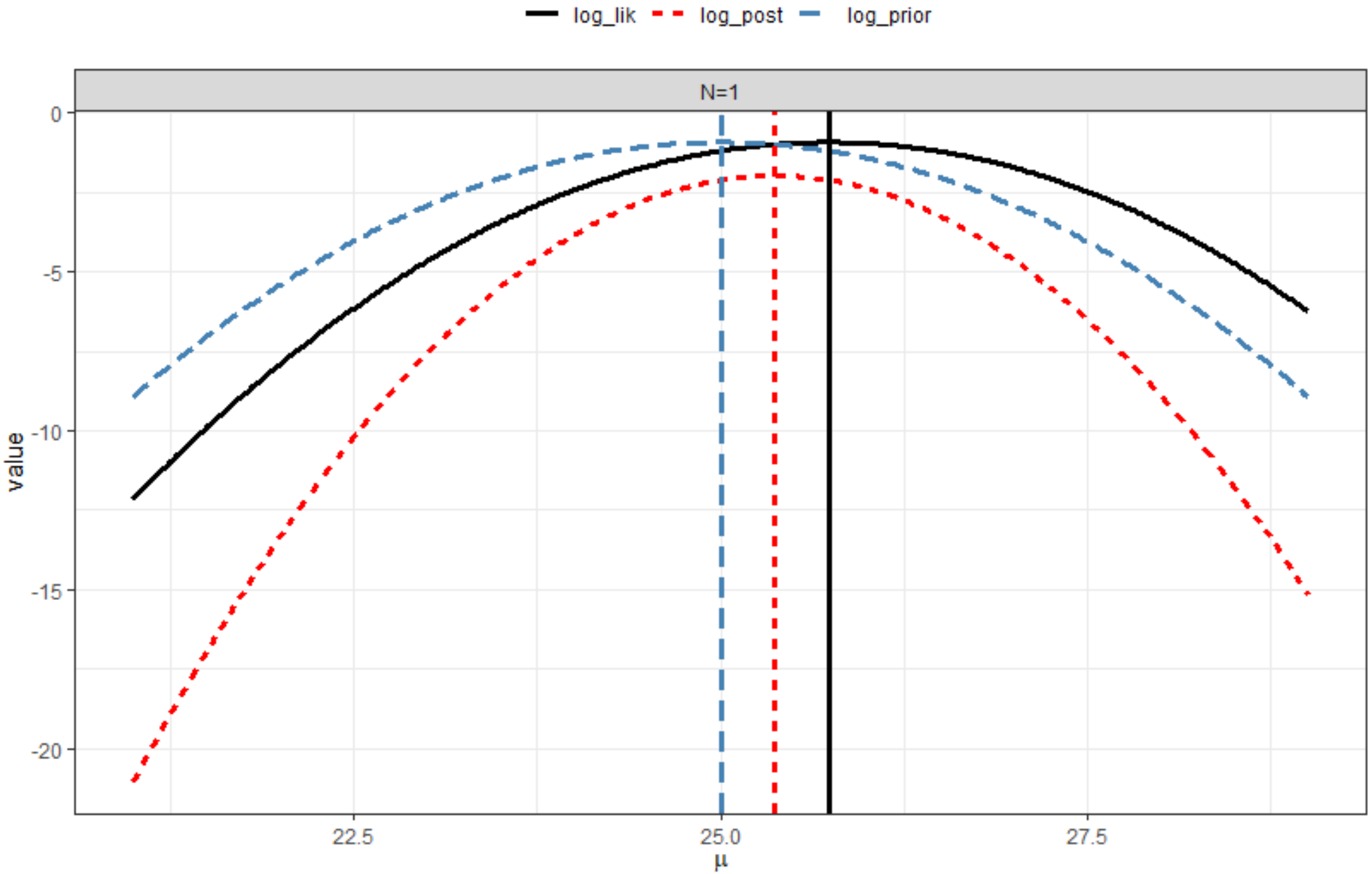
- A Gaussian is a bell curve…but in the **log-space a Gaussian is a parabola!**

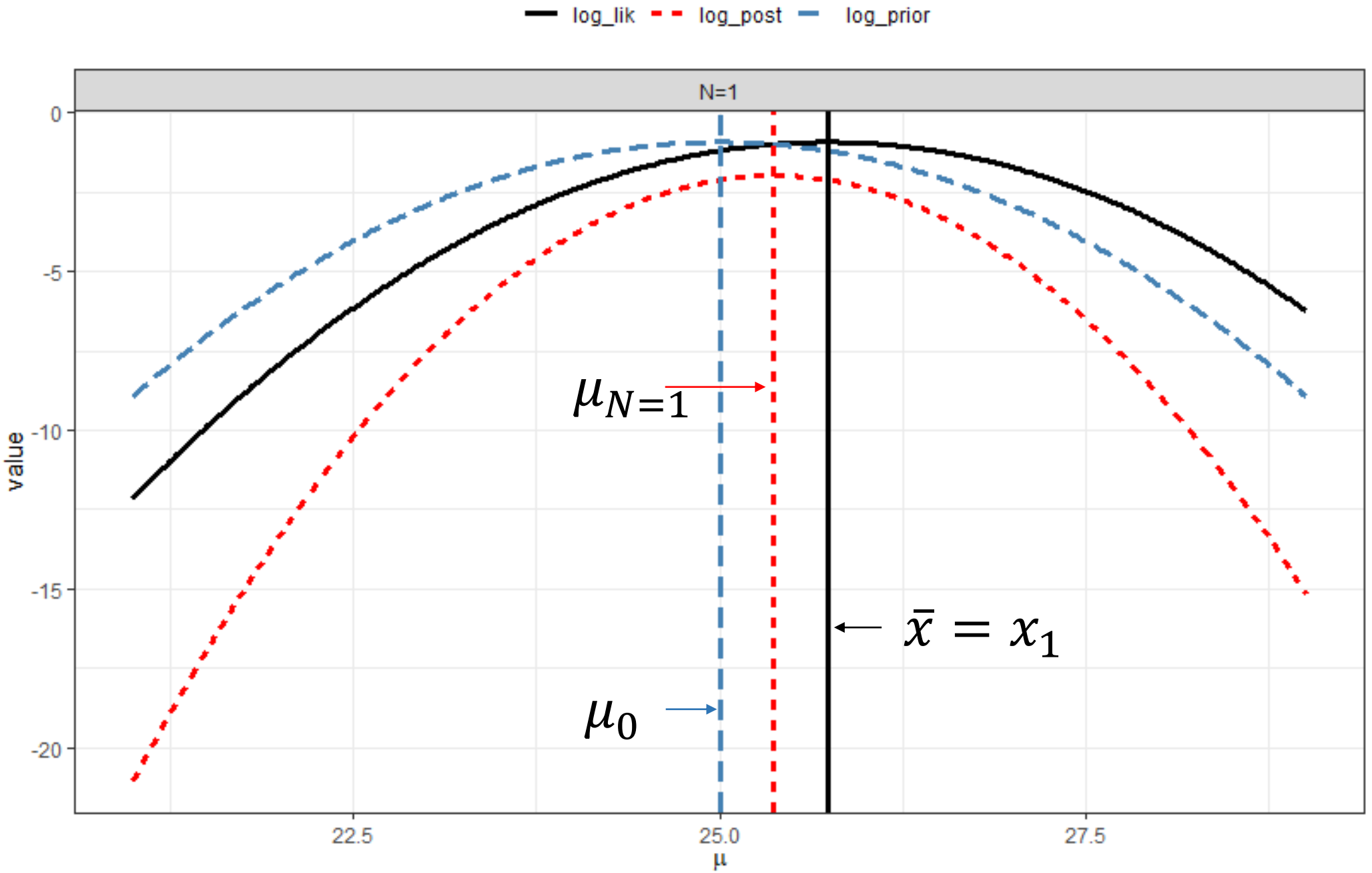

log_lik
log_post
log_prior
N=1
0
-5
-10
-15
-20
value
22.5
25.0
27.5
μ

How about the behavior of the means?

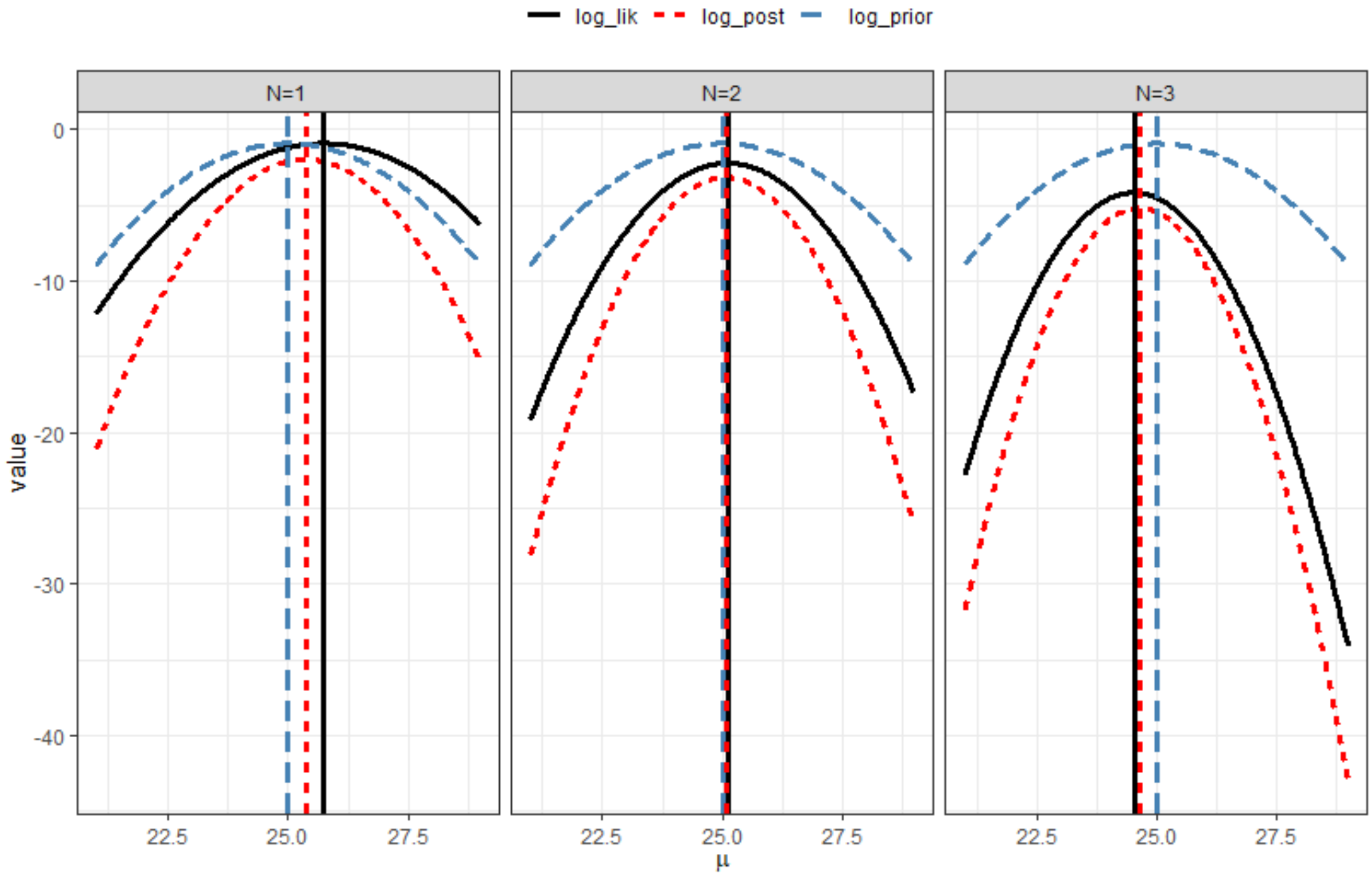
- The **posterior mean is a precision weight average** of the prior mean and the data via the sample mean.

- With 1 observation the sample mean is just the observation itself!

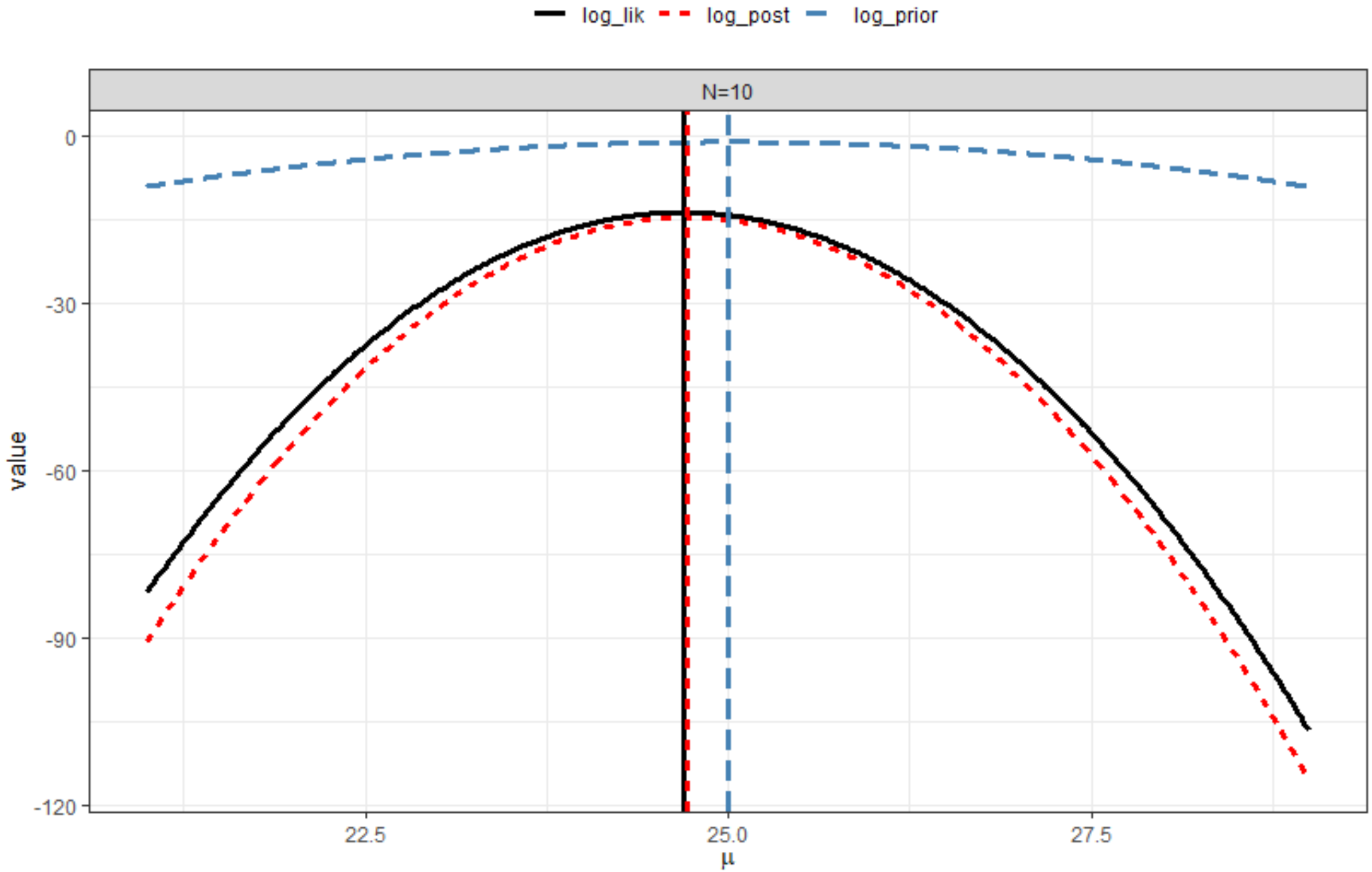

log_lik
log_post
log_prior
N=1
0
-5
-10
-15
-20
value
22.5
25.0
27.5
μ

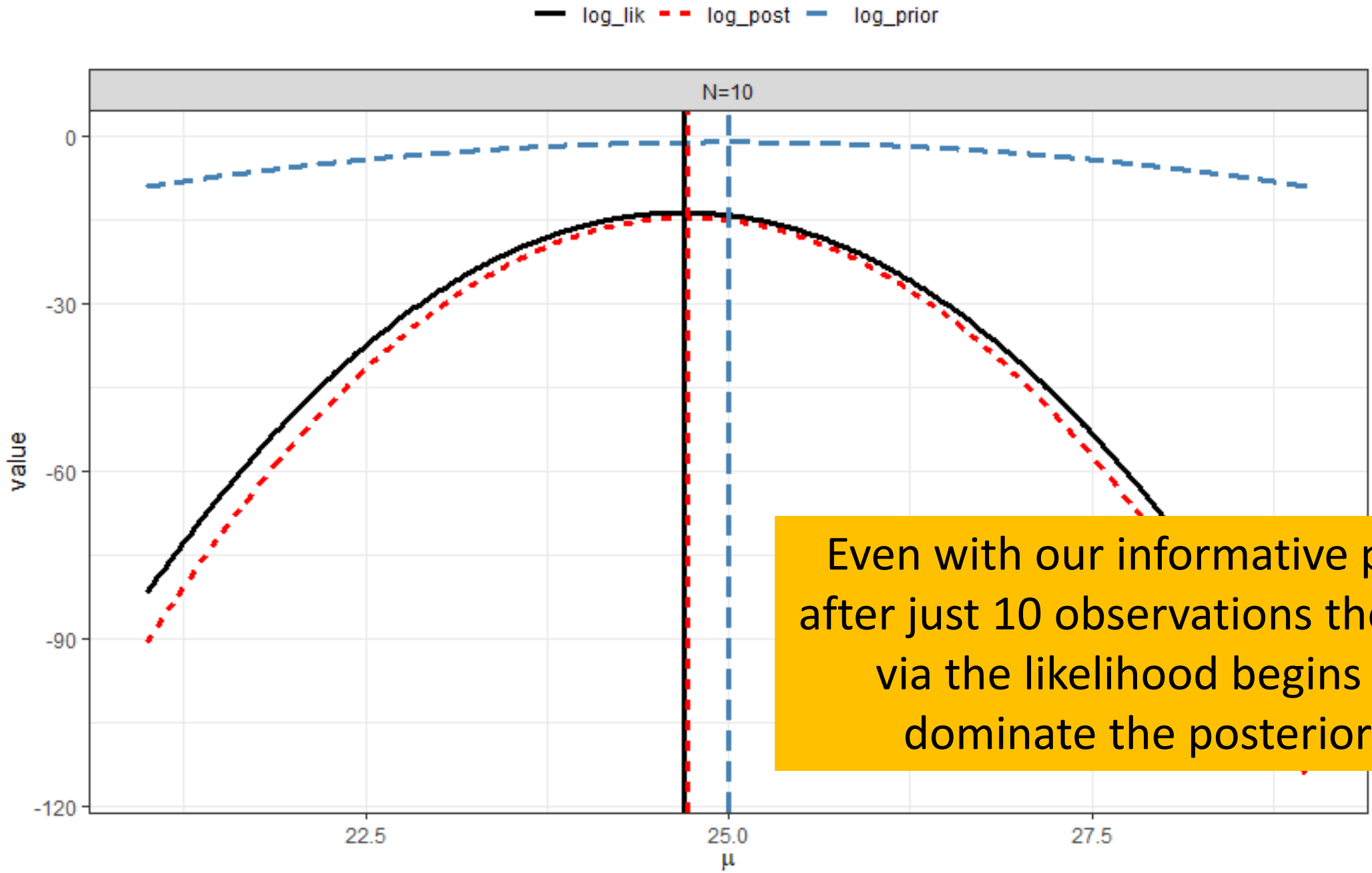

log_lik
log_post
log_prior
N=1
0
-5
-10
-15
-20
value
22.5
25.0
27.5
μ
$\mu_{N=1}$
$\bar{x} = x_1$
μ_0

What if we used the first 2 and the first 3 observations?


log_lik
log_post
log_prior
N=1
N=2
N=3
value
0
-10
-20
-30
-40
22.5
25.0
27.5
μ

What if we used the first 10 observations?

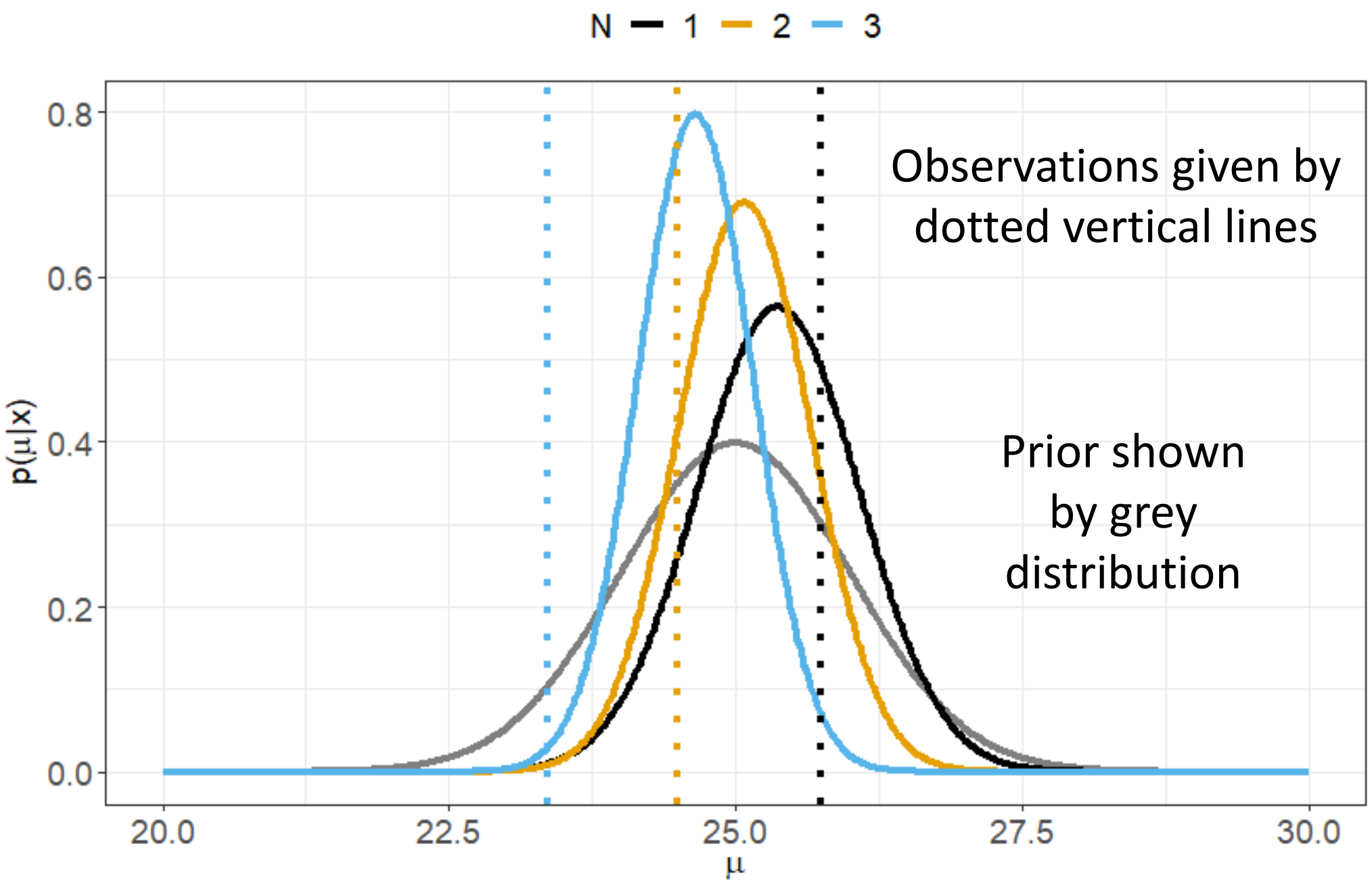

log_lik
log_post
log_prior
N=10
0
-30
-60
-90
-120
value
22.5
25.0
27.5
μ



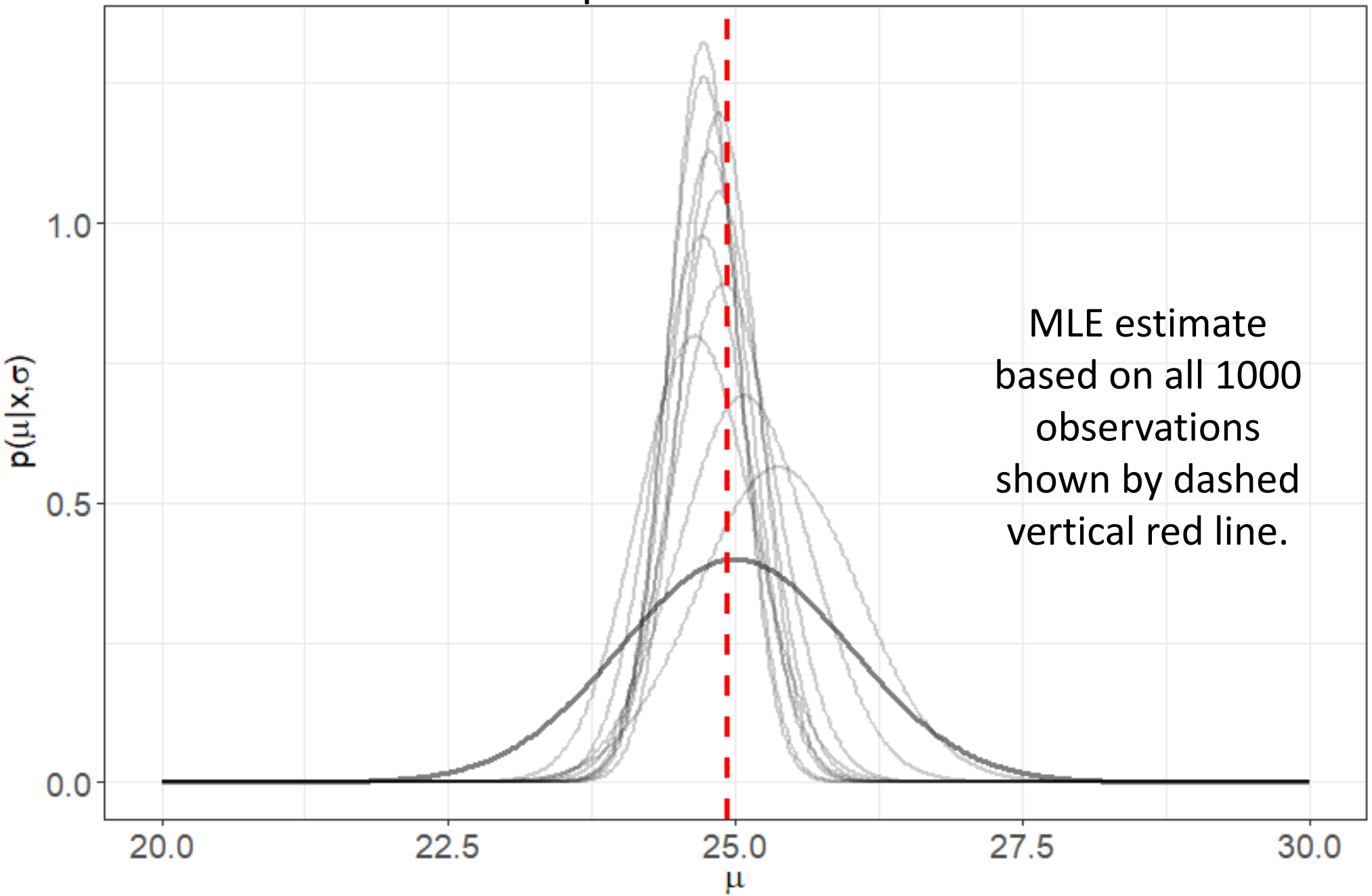
Even with our informative prior, after just 10 observations the data via the likelihood begins to dominate the posterior!

Next, compare how the posterior changes as we sequentially update

- Previous figures compared prior, likelihood, and posterior at a given number of observations.
- Following figures focus on comparing posteriors.
- Posteriors are displayed as densities rather than log-densities.

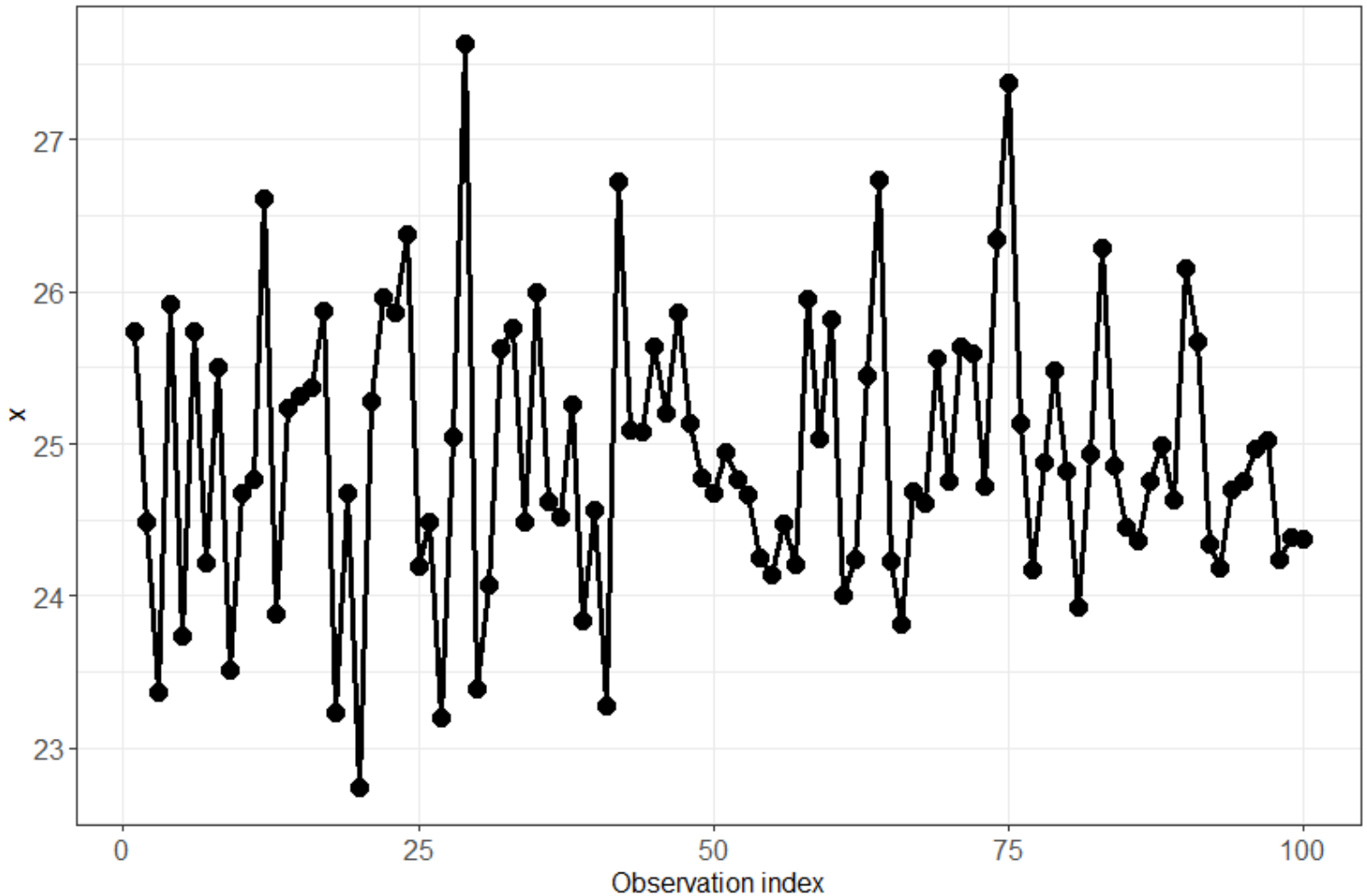

N 1 2 3
0.8
0.6
0.4
0.2
0.0
p(μ|x)
20.0
22.5
25.0
27.5
30.0
μ
Observations given by dotted vertical lines
Prior shown by grey distribution

Evolution of the posterior after 10 observations

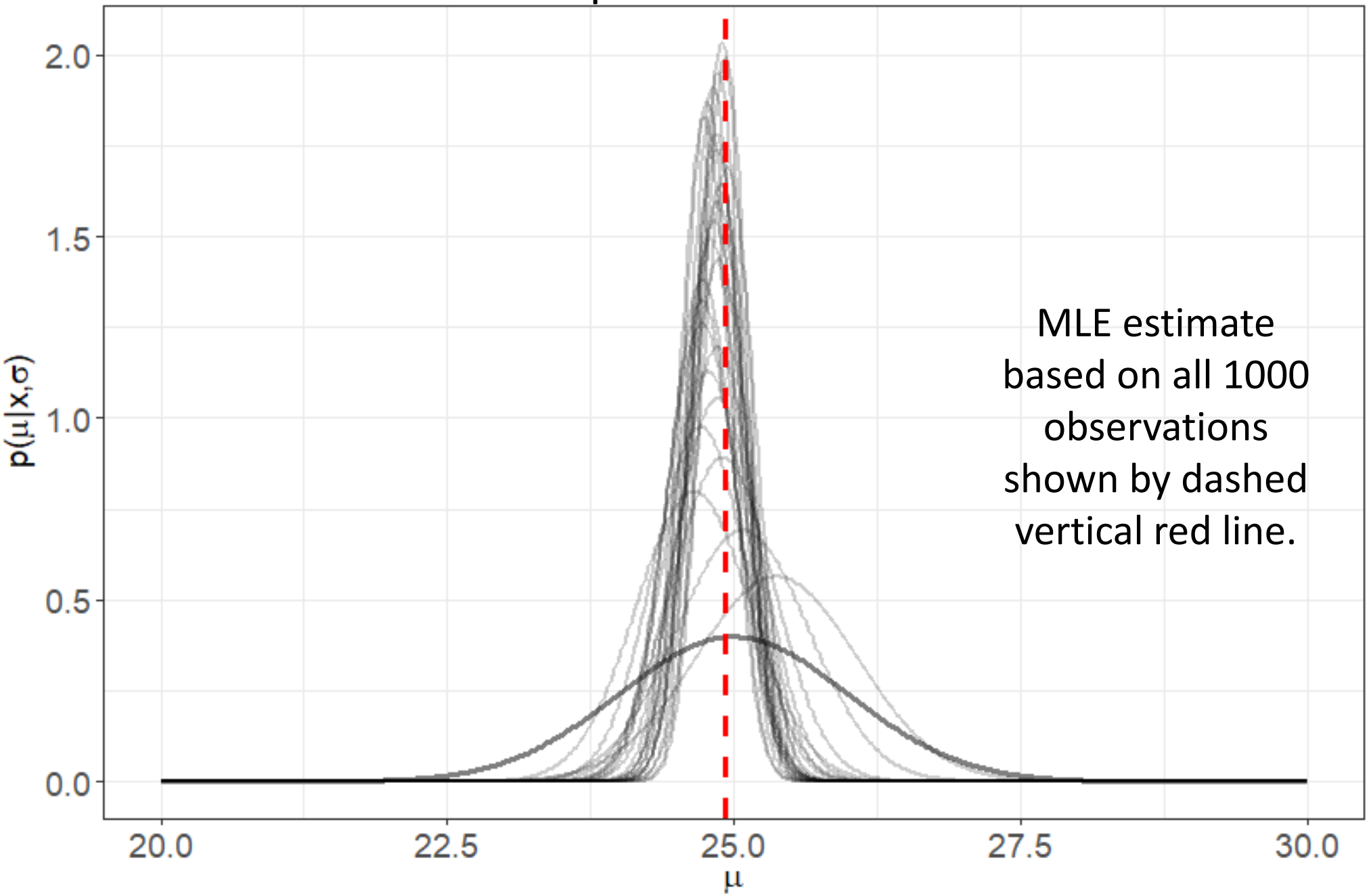


MLE estimate based on all 1000 observations shown by dashed vertical red line.

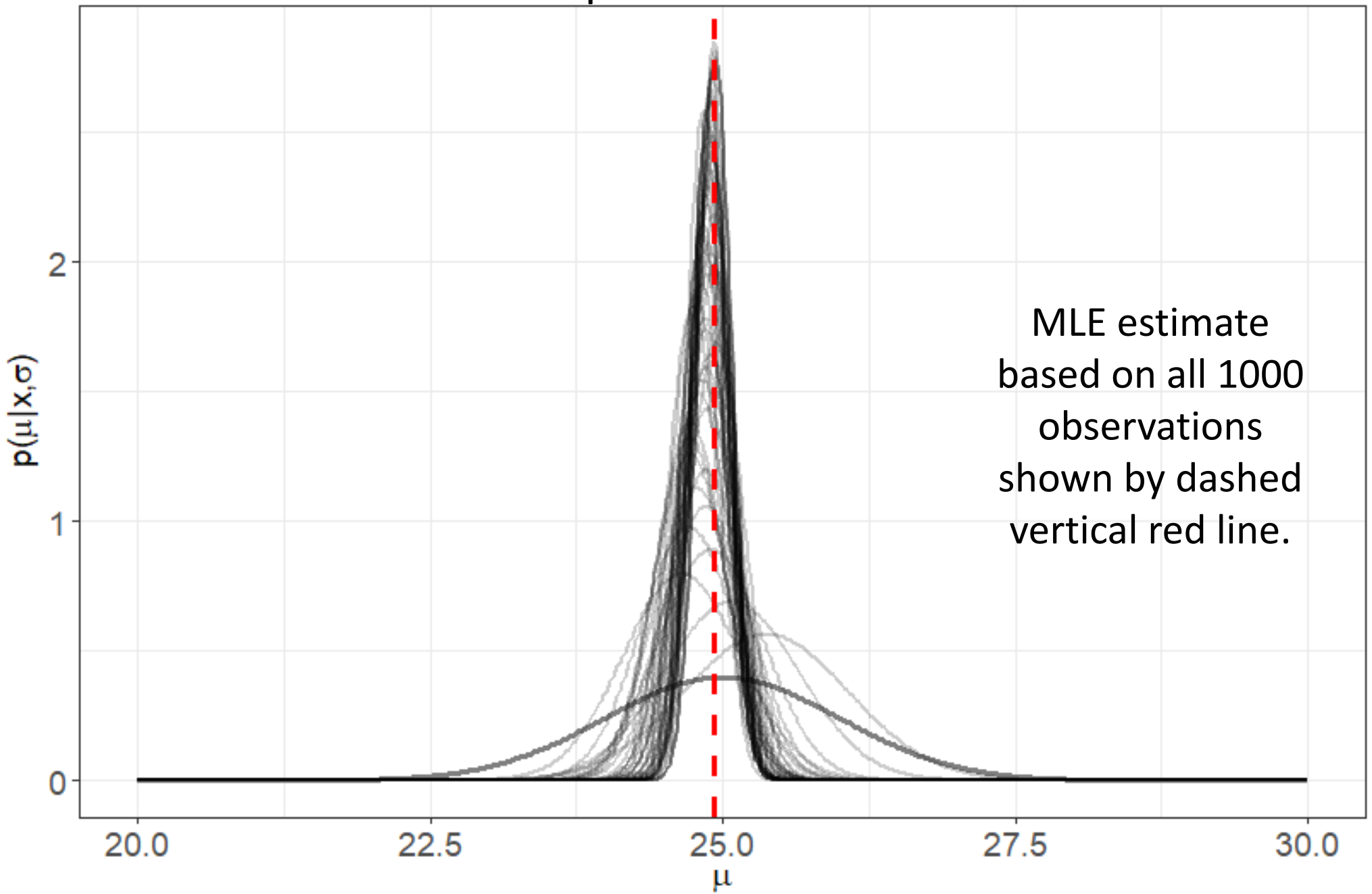
Let’s keep increasing the sample size. Run chart of the first 100 observations.



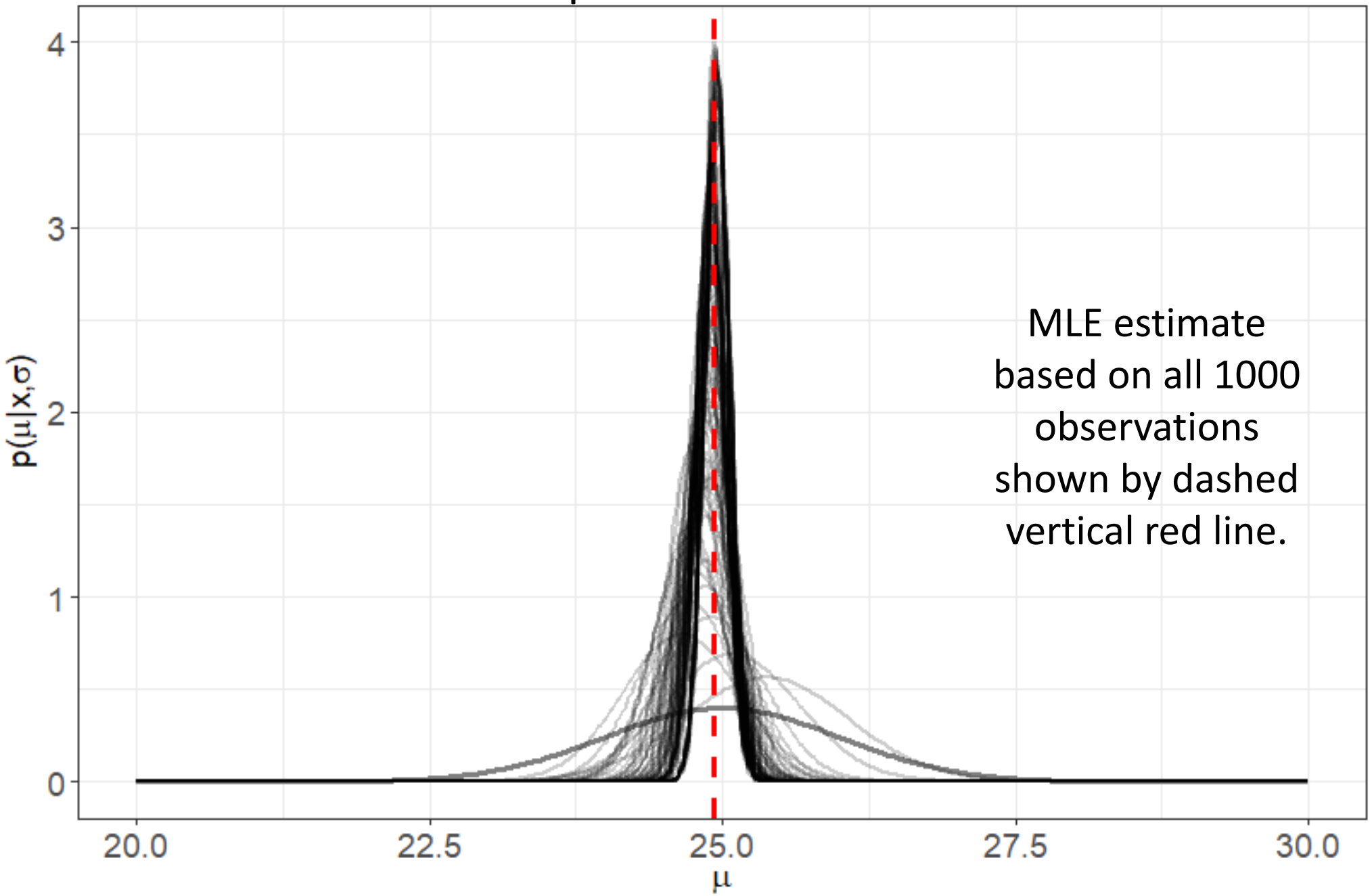
Evolution of the posterior after 25 observations



Evolution of the posterior after 50 observations

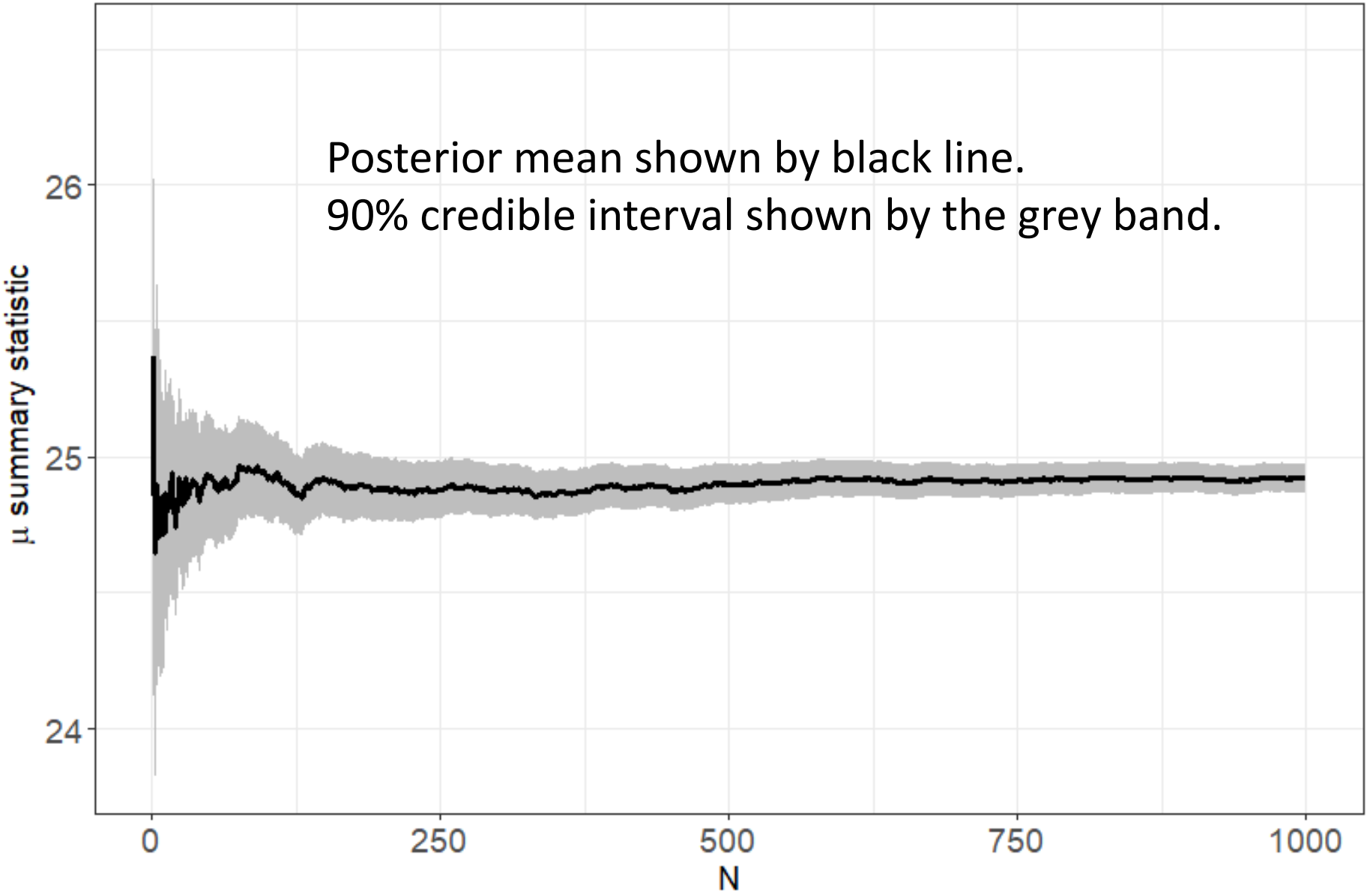


Evolution of the posterior after 100 observations



MLE estimate based on all 1000 observations shown by dashed vertical red line.

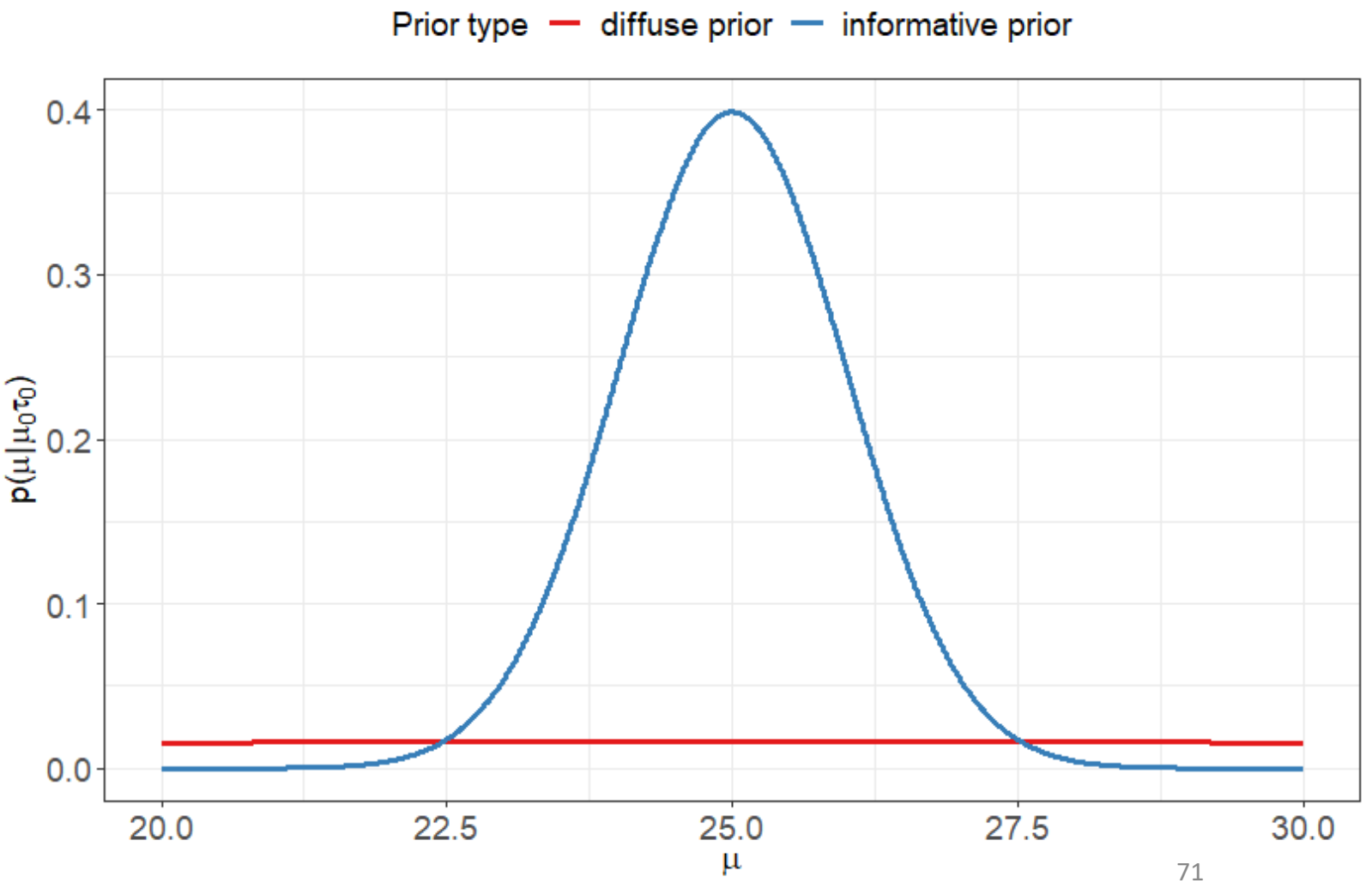
Summarize the posterior distribution and visualize summary statistics as a function of the number of observations



Repeat the procedure but now with a **diffuse prior**, with $\tau_0 = 25$-pounds

- The informative prior with $\tau_0 = 1$-pounds was constraining.

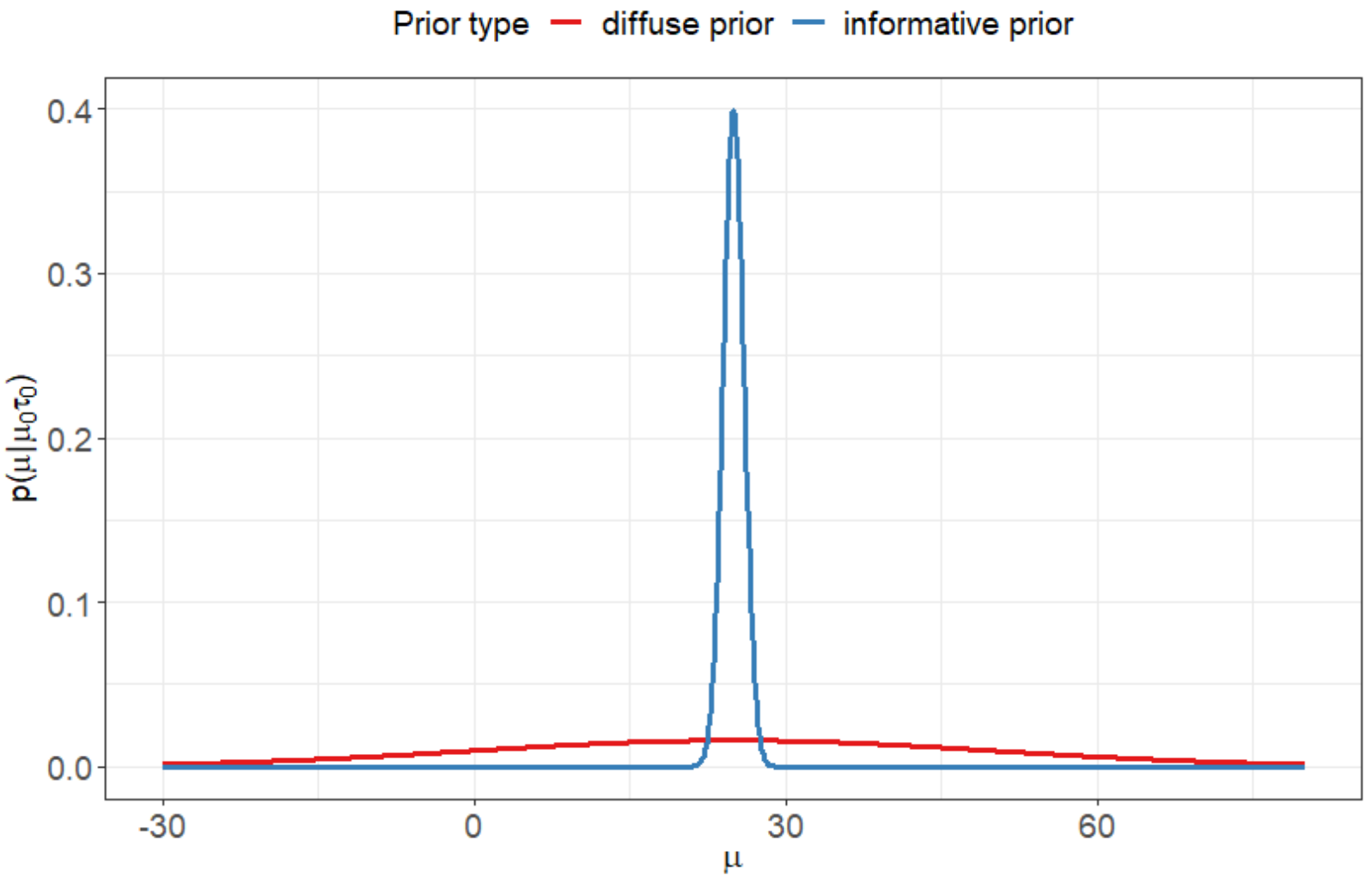
- Information prior believed observations outside 20 to 30 pounds would be considered completely unrealistic.



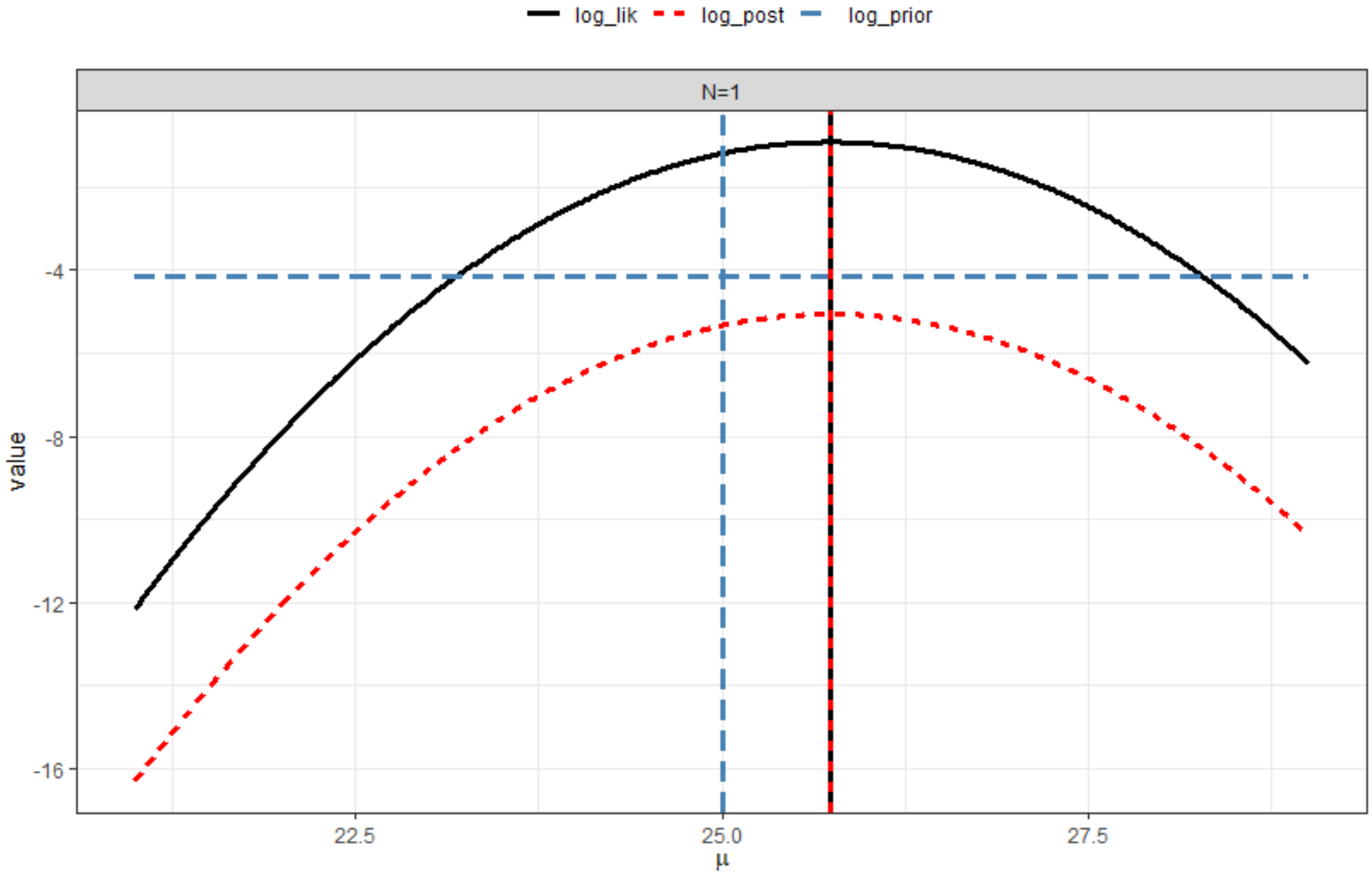
Repeat the procedure but now with a **diffuse prior**, with $\tau_0 = 25$-pounds

- The diffuse prior of $\tau_0 = 25$-pounds adds no information of its own.

- Unphysical negative observations are allowed!

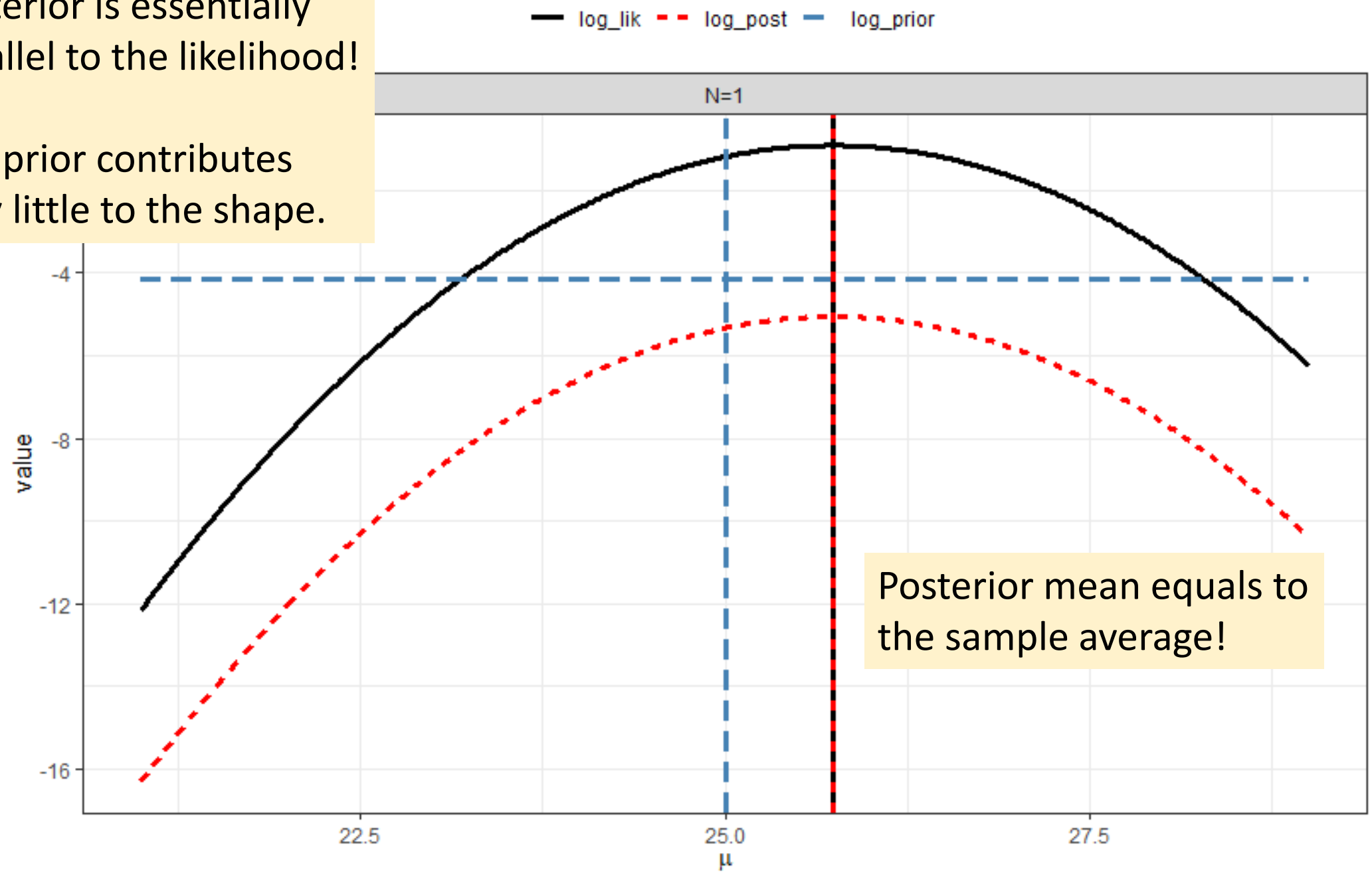


First, compare the log-posterior to the log-likelihood and the diffuse log-prior after just 1 observation.


log_lik
log_post
log_prior
N=1
-4
-8
-12
-16
value
22.5
25.0
27.5
μ

Posterior is essentially parallel to the likelihood!

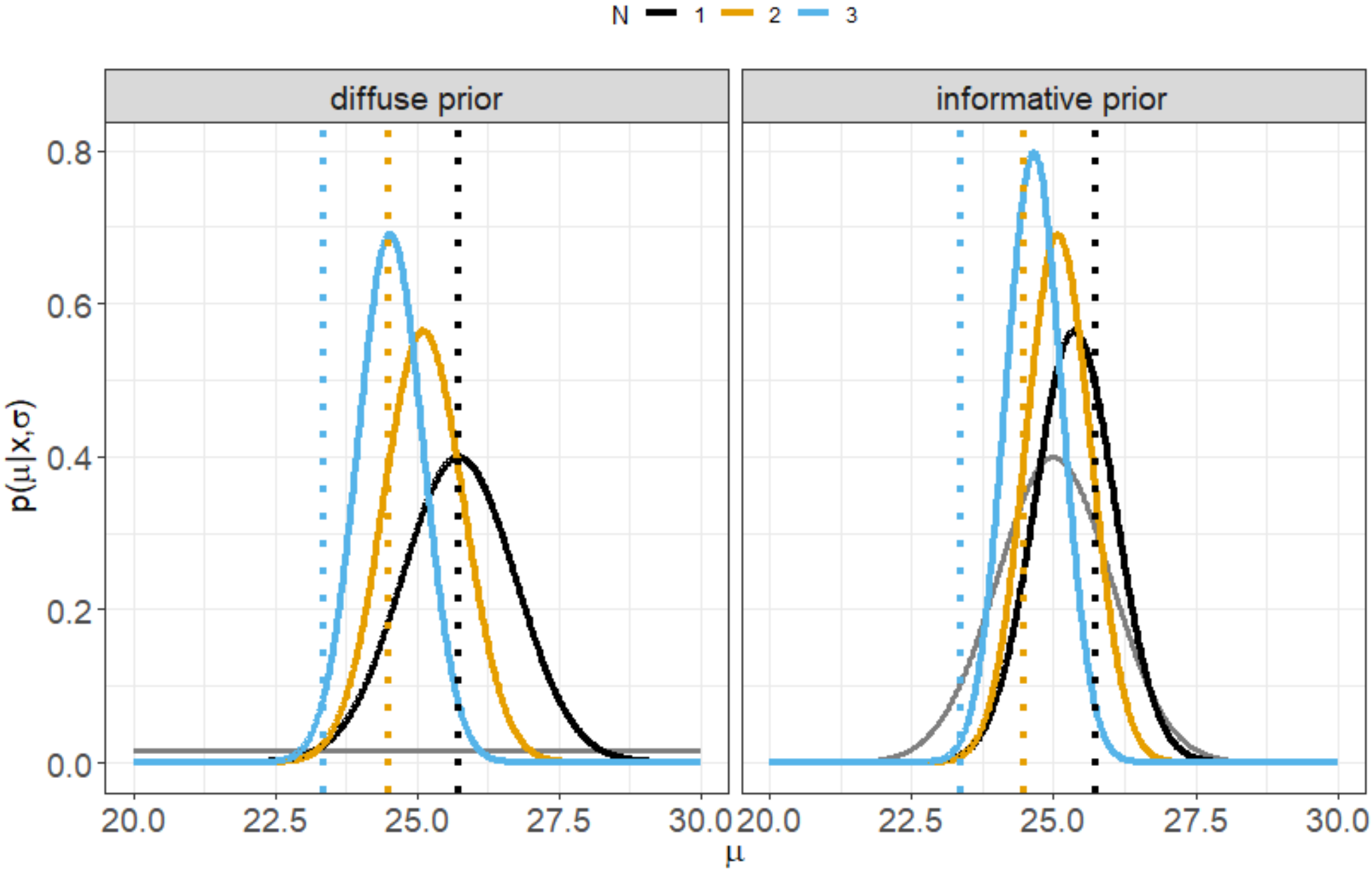
The prior contributes very little to the shape.



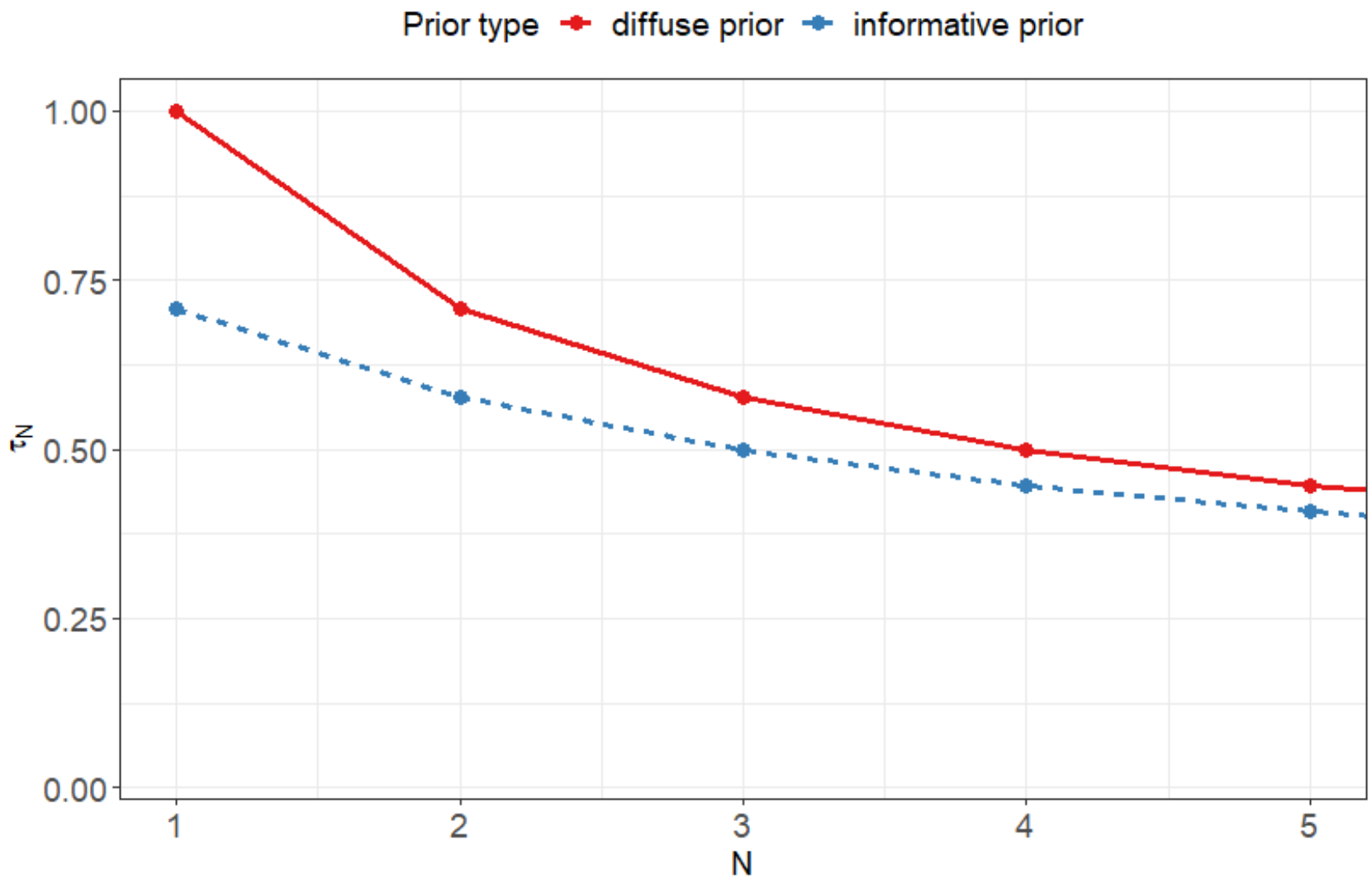
Posterior mean equals to the sample average!

- Next, compare the posterior density sequential updates between the informative prior and diffuse prior cases.

- Focus on the first three observations.


N 1 2 3
diffuse prior
informative prior
p(μ|x,σ)
0.0
0.2
0.4
0.6
0.8
20.0
22.5
25.0
27.5
30.0
μ

Compare the posterior standard deviation, τ_N, as a function of the number of observations



After 50 observations…the posterior τ_N's from the two priors are essentially identical