INFSCI 2595

Fall 2019

Information Sciences Building: Room 403

Lecture 04

So far we have been exposed to the following probability distributions

Discrete probability distributions

- Distribution referred to as *probability mass function* (pmf)
  - Bernoulli -> special case of the Binomial distribution

Continuous probability distributions

- Distribution referred to as *probability density function* (pdf)
  - Beta distribution -> conjugate prior to the Binomial

In this lecture we will focus exclusively on the **Gaussian** or **Normal** distribution

- Continuous probability distribution.

- Classic “bell curve”, used in a wide variety of applications and fields.

- Typically defined by two *hyperparameters* μ and σ

What do the *hyperparameters* represent?

- The expected value or mean is: $\mathbb{E}[x] = \mu$
- The variance is: $\mathrm{var}(x) = \sigma^2$
- The standard deviation is then: $\sqrt{\mathrm{var}(x)} = \sigma$

Nomenclature of the pdf

- The pdf for the continuous variable x is usually defined as:

$$x|\mu,\sigma \sim N(x|\mu,\sigma^2) \text{ or } x|\mu,\sigma \sim \mathcal{N}(x|\mu,\sigma^2)$$

- I will sometimes use a notation more in line with R's parameterization:

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

Nomenclature of the pdf

- The pdf for the continuous variable x is usually defined as:

$$x|\mu,\sigma \sim N(x|\mu,\sigma^2) \text{ or } x|\mu,\sigma \sim \mathcal{N}(x|\mu,\sigma^2)$$

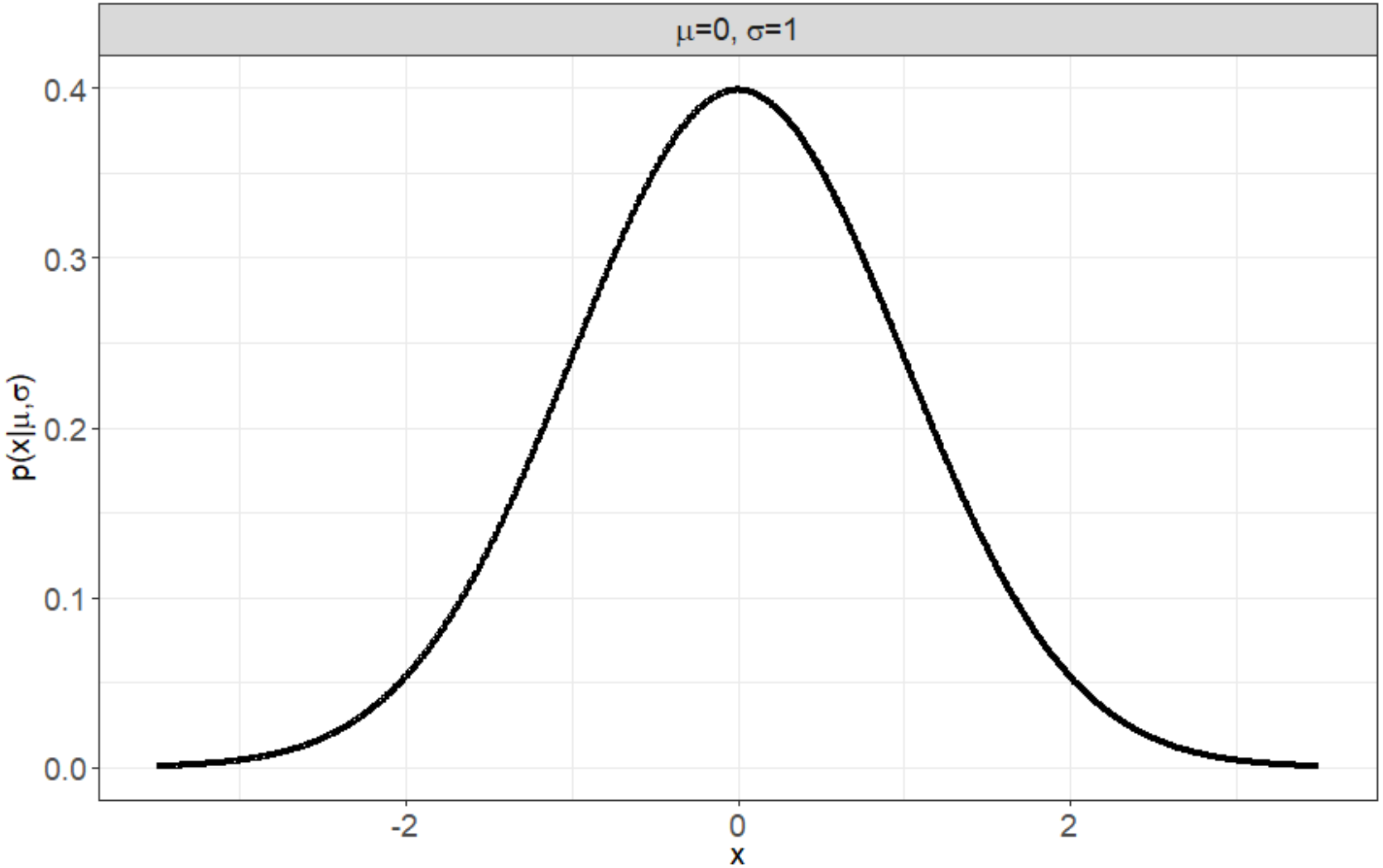
PDF written with the variance

- I will sometimes use a notation more in line with R's parameterization:

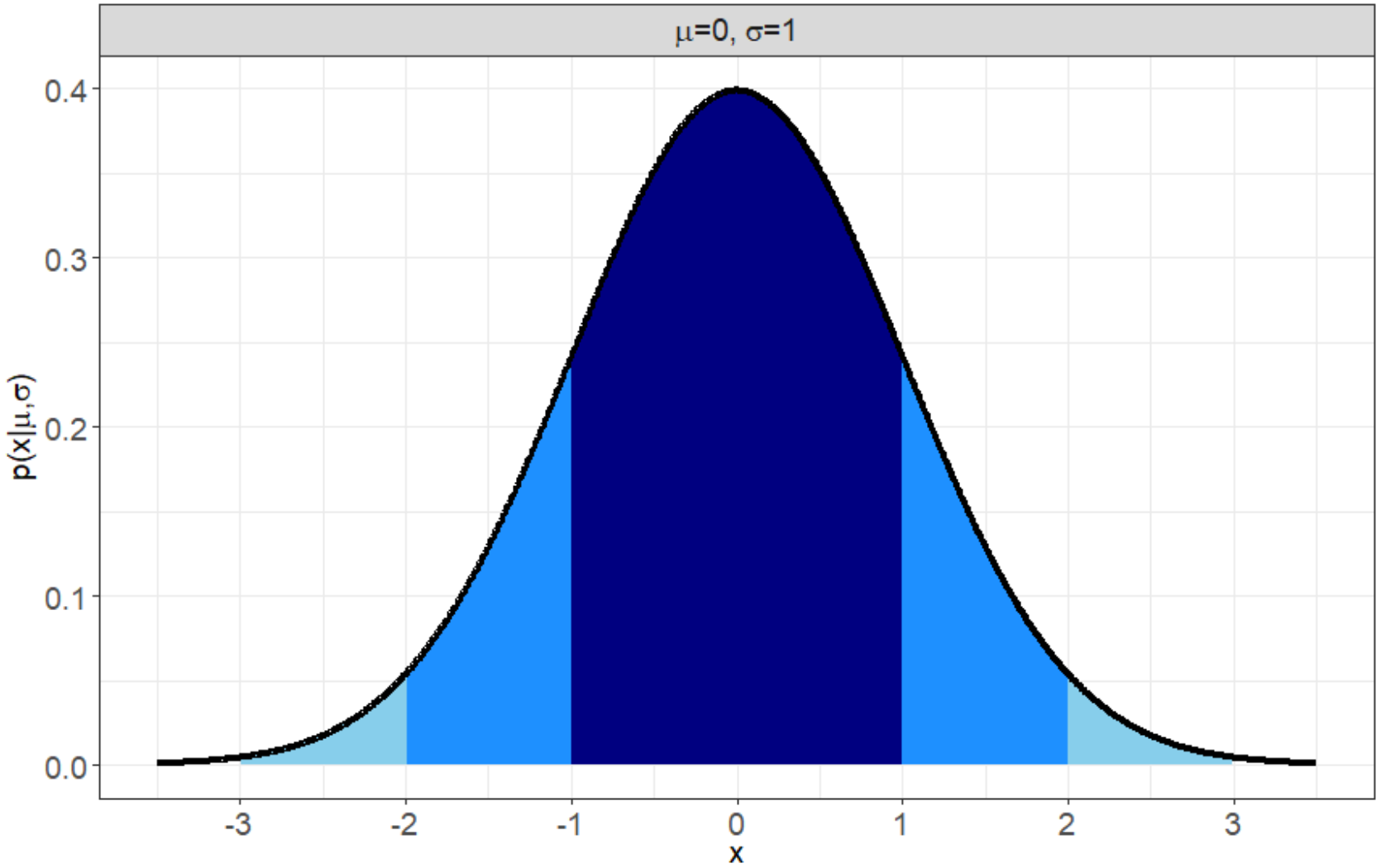
$$x|\mu,\sigma \sim \text{normal}(x|\mu,\sigma)$$

PDF written with the standard deviation

Standard normal distribution



Coverage: $\pm\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$



Gaussian pdf

$$\text{normal}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Standard normal pdf

$$\text{normal}(x|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$$

Importance of the standard normal

- Define the z-score as:

$$z = \frac{x - \mu}{\sigma}$$

- Which means we can rewrite the original variable, x as:

$$x = \sigma \cdot z + \mu$$

We can define a general Gaussian in terms of the standard normal

- The general Gaussian pdf on x:

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

- Can be equivalently defined as:

$$\text{normal}(z|0,1), x = \sigma \cdot z + \mu$$

Let's use **change-of-variables** to check if this is indeed true.

- A *one-to-one* transformation, such as $x \rightarrow z$, with $z = g(x)$ with corresponding inverse function $x = g^{-1}(z)$.

- We can write the distribution of z "based on" the distribution of x using:

$$p_z(z) = p_x\left(g^{-1}(z)\right) \cdot \left|\frac{d}{dz}\left(g^{-1}(z)\right)\right|$$

For our case of the z-score the **link** function between x and z is:

$$z = g(x) = \frac{x - \mu}{\sigma}$$

The *inverse* link function is then:

$$x = g^{-1}(z) = \sigma \cdot z + \mu$$

The change-of-variables expression looks complex…

$$p_z(z) = p_x\left(g^{-1}(z)\right) \cdot \left|\frac{d}{dz}\left(g^{-1}(z)\right)\right|$$

The change-of-variables expression looks complex…so break it up into each component

$$p_z(z) = p_x\left(g^{-1}(z)\right) \cdot \left|\frac{d}{dz}\left(g^{-1}(z)\right)\right|$$

Substitute $x = \sigma \cdot z + \mu$ into the pdf for x!

Absolute value of the derivative of x wrt z!

Substitute $x = \sigma \cdot z + \mu$ into the pdf for x!

- The pdf for x is just the normal distribution:

$$\text{normal}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

- Substitute $x = \sigma \cdot z + \mu$ and simplify:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\sigma \cdot z + \mu - \mu}{\sigma}\right)^2\right\} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

Absolute value of the derivative of x wrt z!

$$\frac{d}{dz}\left(g^{-1}(z)\right) = \frac{d}{dz}(x) = \frac{d}{dz}(\sigma \cdot z + \mu) = \sigma$$

- The standard deviation is positive: $\sigma > 0$
- $|\sigma| = \sigma$ therefore:

$$\frac{d}{dz}\left(g^{-1}(z)\right) = \sigma$$

Bring the two terms together, and simplify

$$p_z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} \cdot \sigma$$

$$p_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

Bring the two terms together, and simplify

$$p_z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} \cdot \sigma$$

$$p_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

$$p(z) = \text{normal}(z|0,1)$$

Let's start working with the Gaussian distribution

Problem statement

I will weigh myself N times, producing a sequence of observed weights $\mathbf{x} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$.

- Each x_n is my weight in pounds.

I'm not particularly interested in my "true" weight.

Rather, I'm interested in calculating the probability that I weigh less than 255 pounds.

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

- Assume each observation is **independent** of the others.
- The joint distribution of all N observations can therefore be factored:

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

- 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 parameters 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.
  - Additionally, I may shift my weight slightly differently each time I stand on the measuring scale. Distributing my weight differently may generate a different result from a previous attempt.

As an experiment, let's first weigh a dumbbell

- We feel confident that the weight of a dumbbell is very close to the reported value on the weight.
- So, let's weigh a 25-pound dumbbell.
- 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!**

Bayesian formulation for the unknown μ

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

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

- How can we encode our prior belief on μ?

Conjugate prior to a normal likelihood

- As we saw last lecture, 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

- As we saw last lecture, 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!

- Denote our prior as: $\mu|\mu_0, \tau_0 \sim \text{normal}(\mu_0, \tau_0)$

The posterior on μ will therefore be a normal distribution!

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

The posterior on μ will therefore be a normal distribution!

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

How can we derive a normal from that???

Start by writing out all terms involving the unknown parameter

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

Take 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 average!

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:

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

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

The weights are proportional to the precisions!

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:

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

What do we 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.
  - So 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

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

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

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

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

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

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

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

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

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 reveals the link between the Bayesian formulation and the classic solution

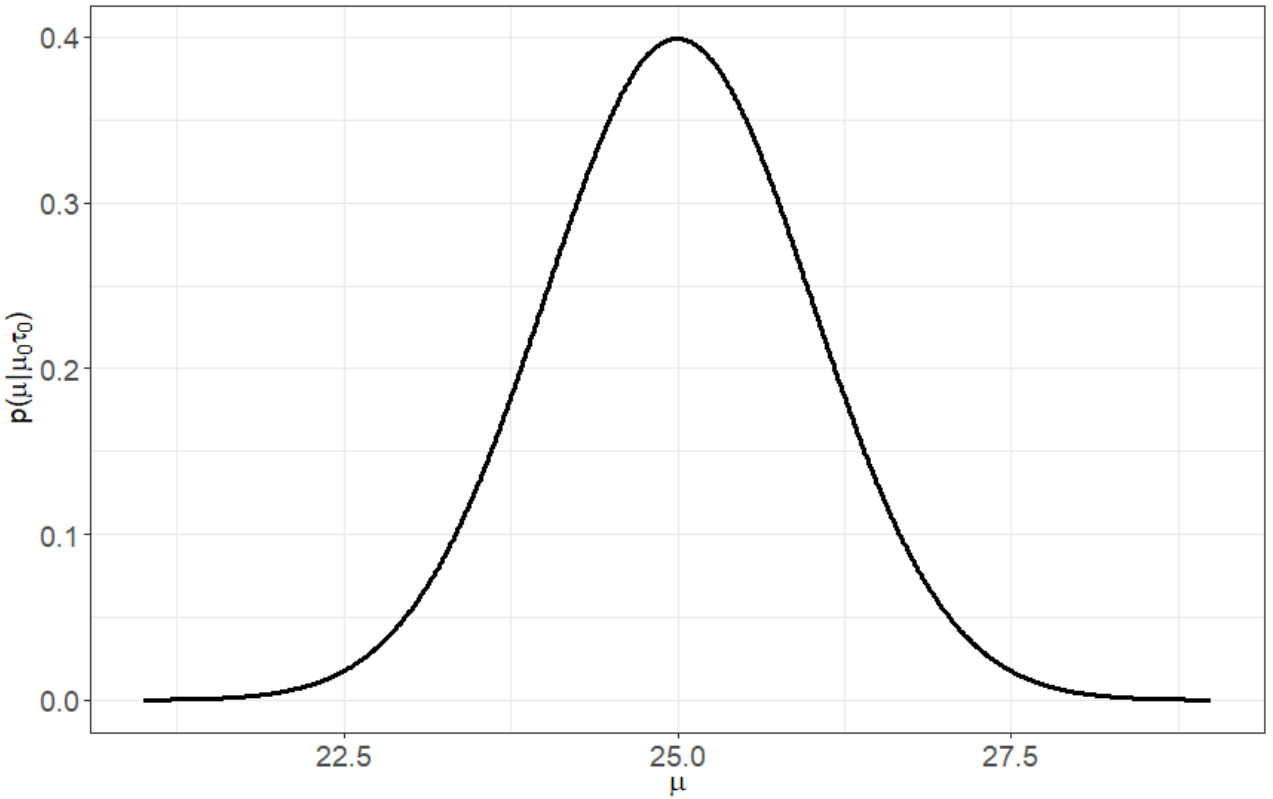
- The Bayesian result is a compromise between the prior and the sample average (the data).

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

Let's see the math in action!

- We weighed a 25-pound dumbbell in our experiment.
- A priori we felt that the manufacturer's weight should be precise.
- So the prior was defined as:

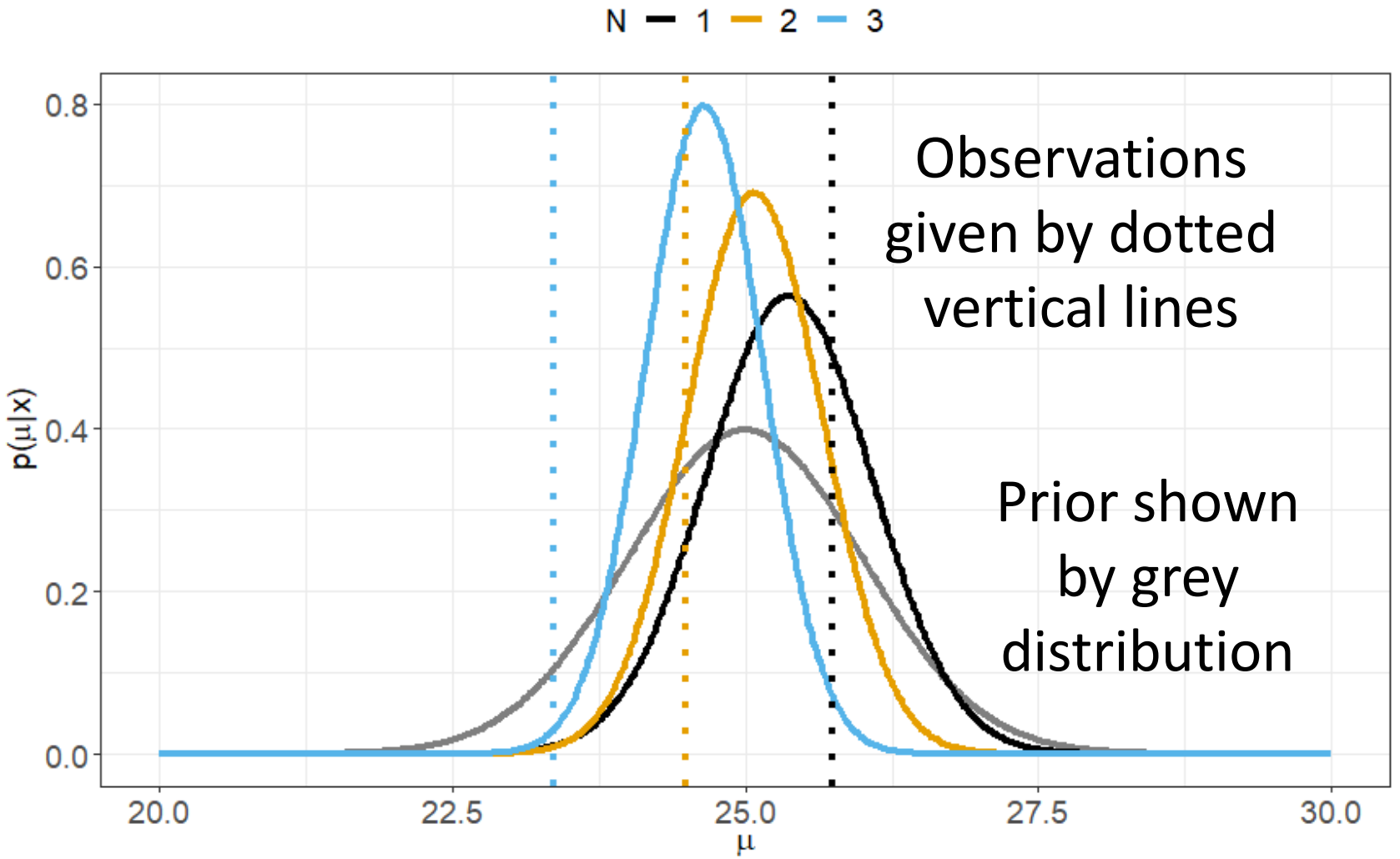
$$\mu|\mu_0, \tau_0 \sim \text{normal}(\mu|25, 1)$$



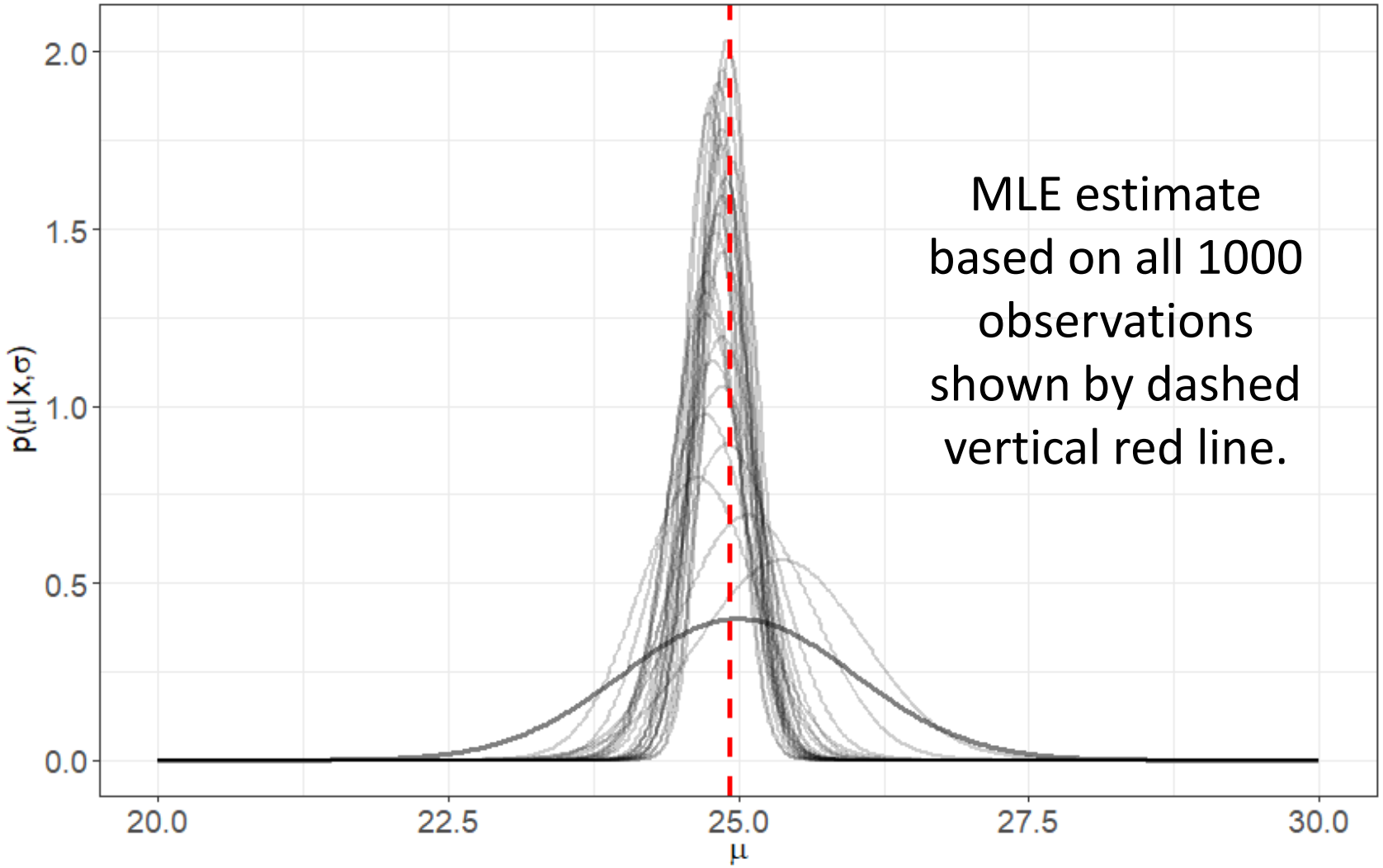
We performed the experiment measuring the weight of the dumbbell up to 1000 times!!!

- 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 $\tau_0 = 1$.

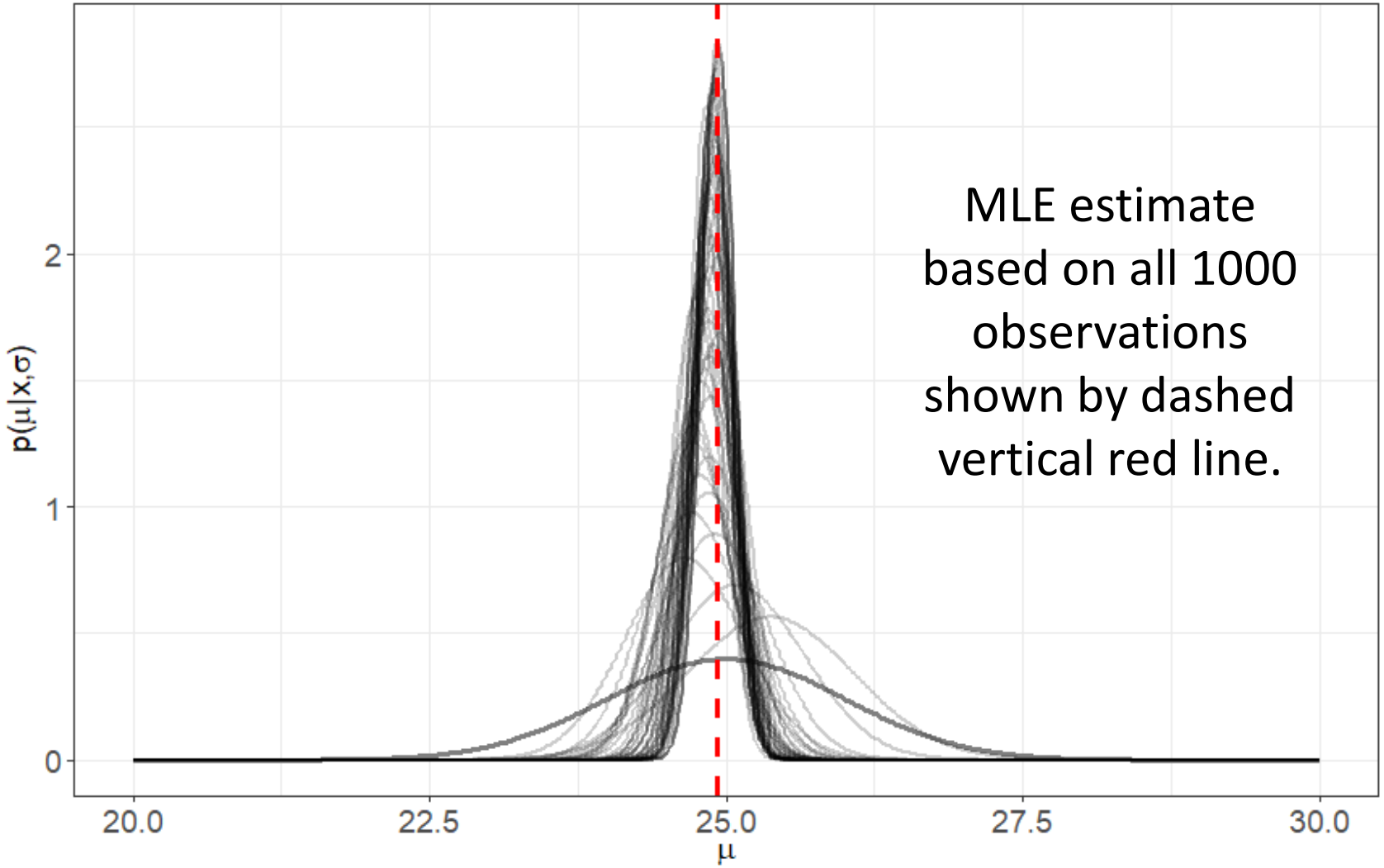
Posterior distribution after 1, 2, and 3 observations



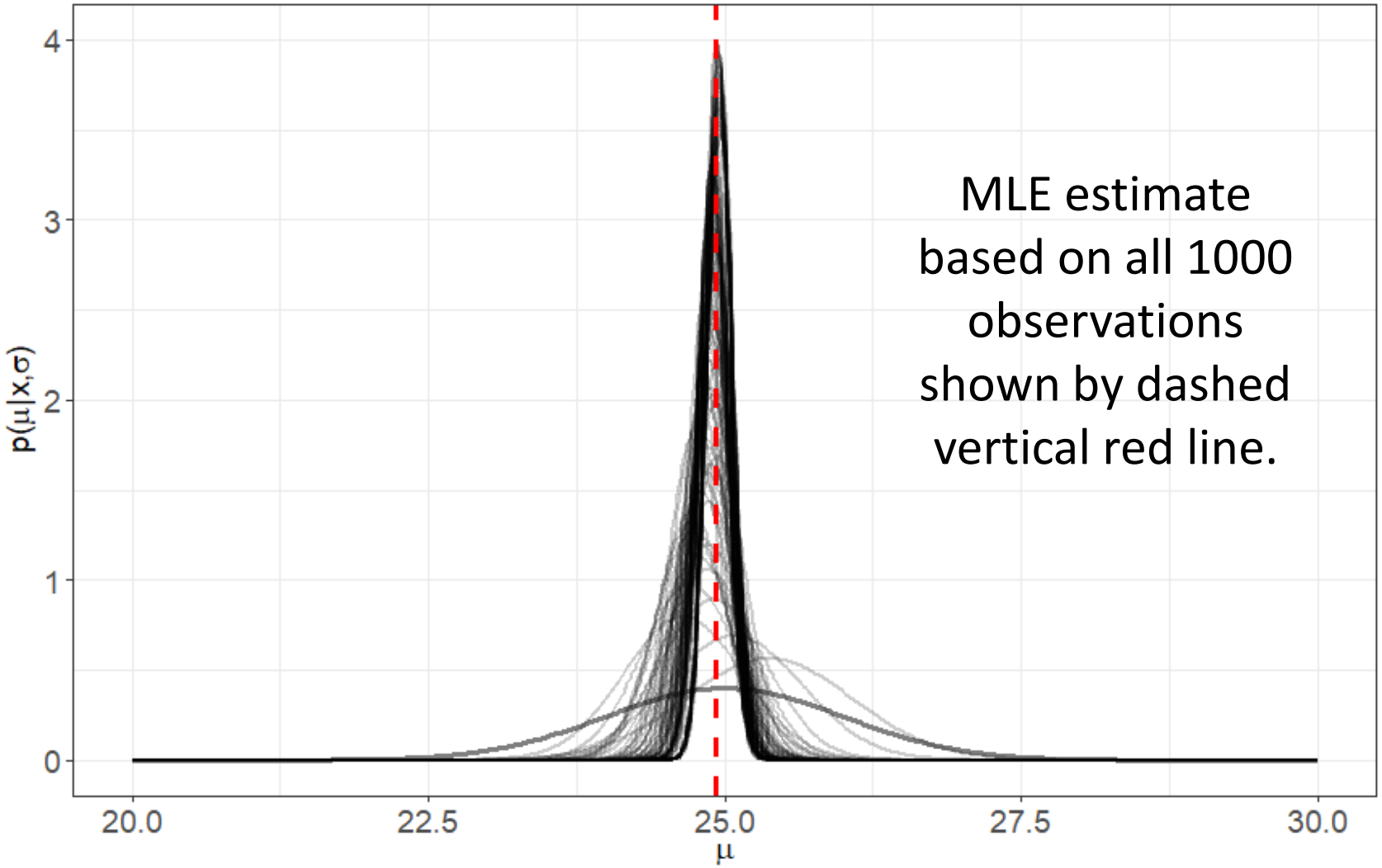
Evolution of the posterior after 25 observations



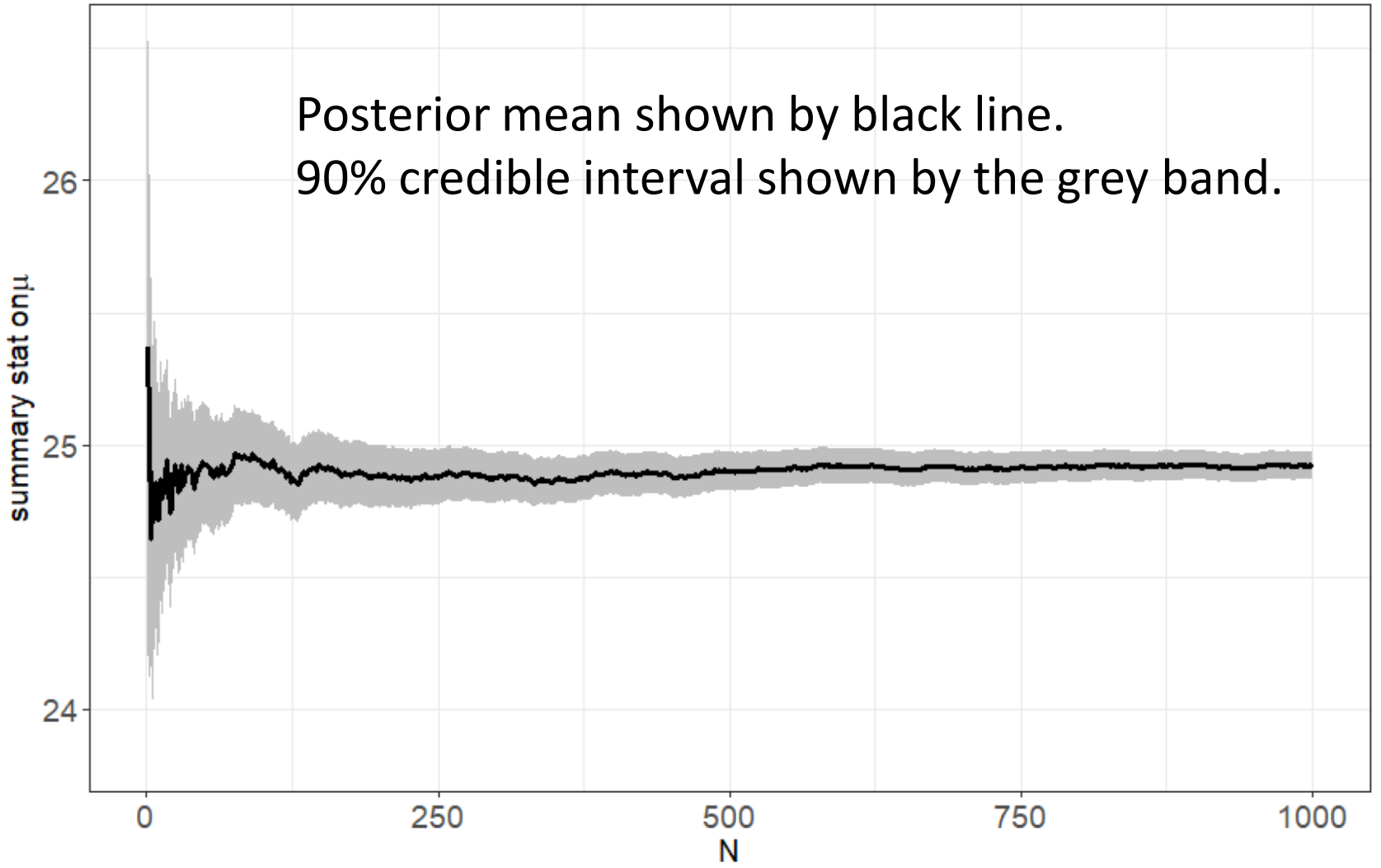
Evolution of the posterior after 50 observations



Evolution of the posterior after 100 observations



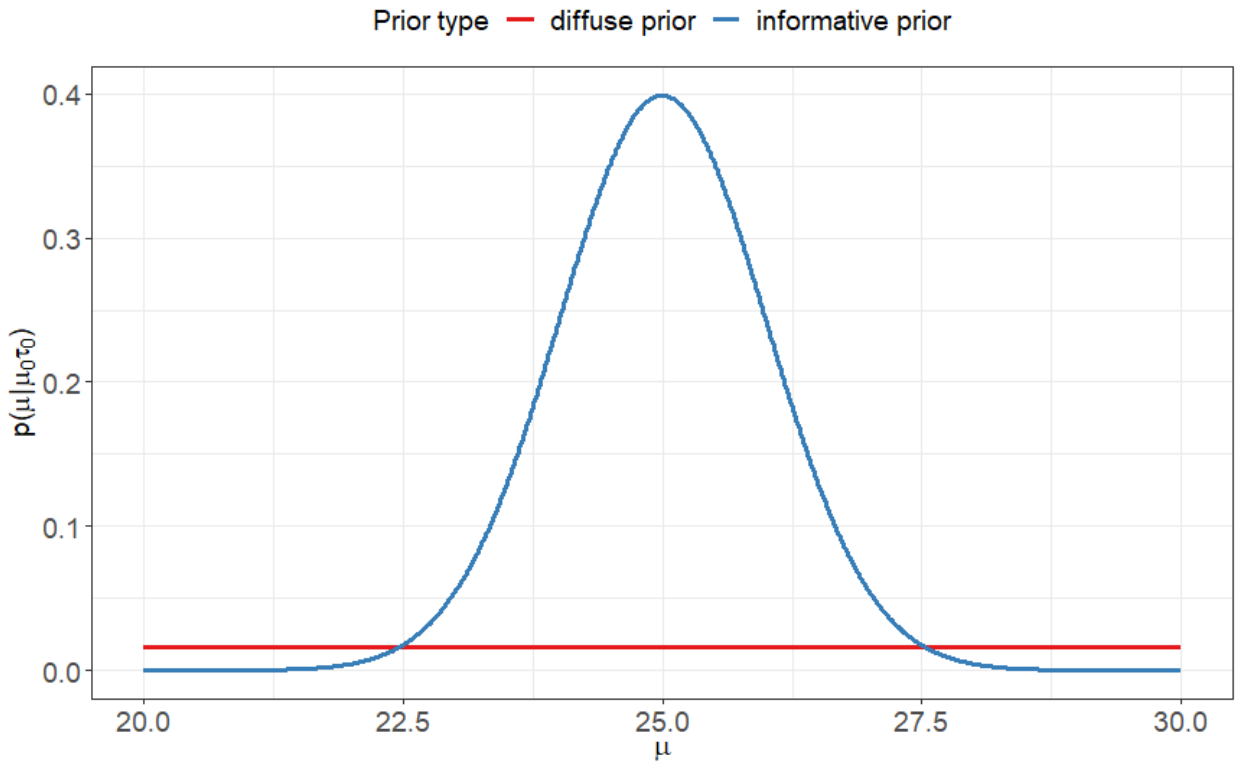
Summarize the posterior distribution and visualize summary statistics as a function of the number of observations



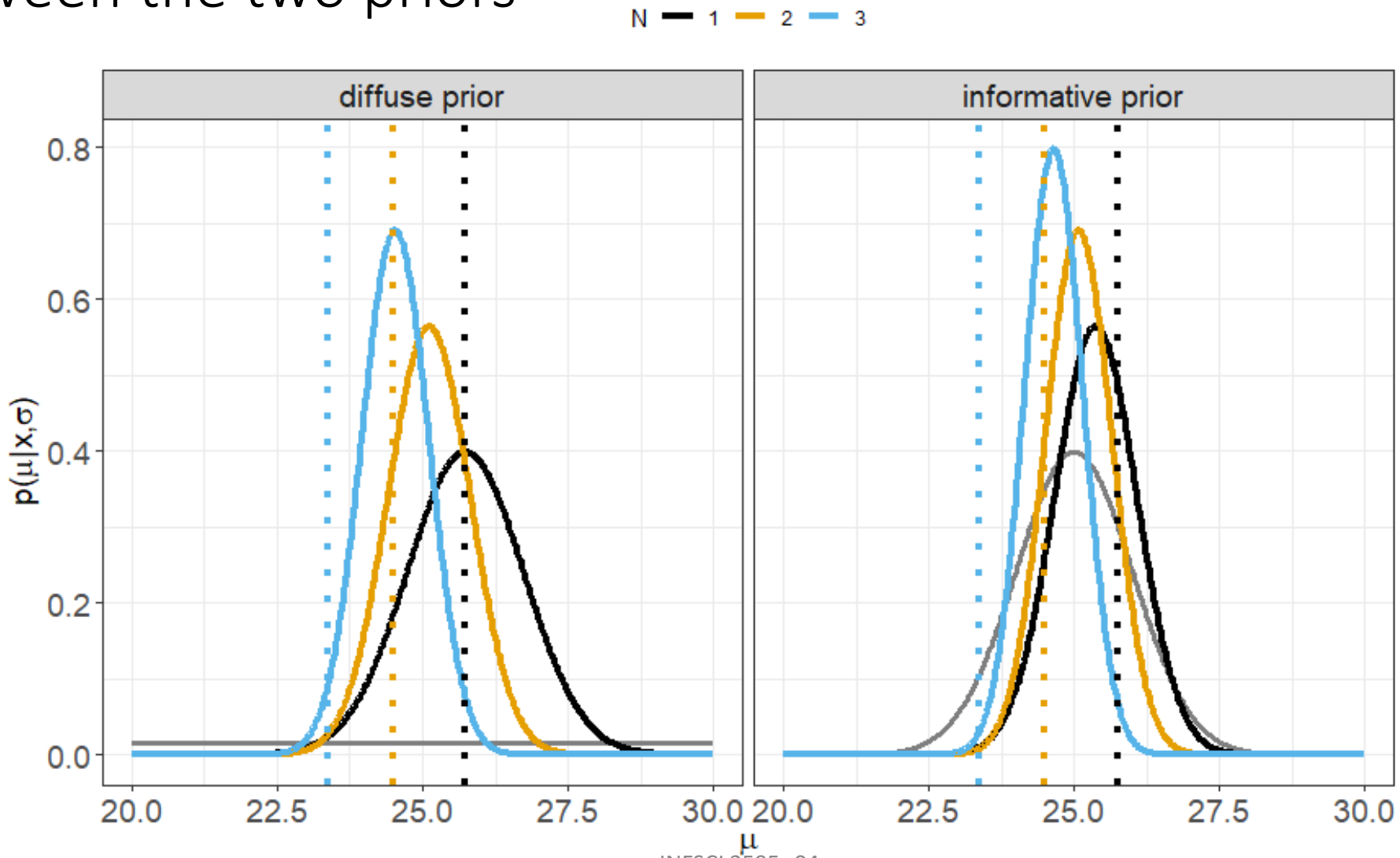
Repeat the procedure but this with a **diffuse prior**, with $\tau_0 = 25$-pounds

- The informative prior with $\tau_0 = 1$-pounds is constraining.
  - Observations outside 20 to 30 pounds would be considered completely unrealistic.

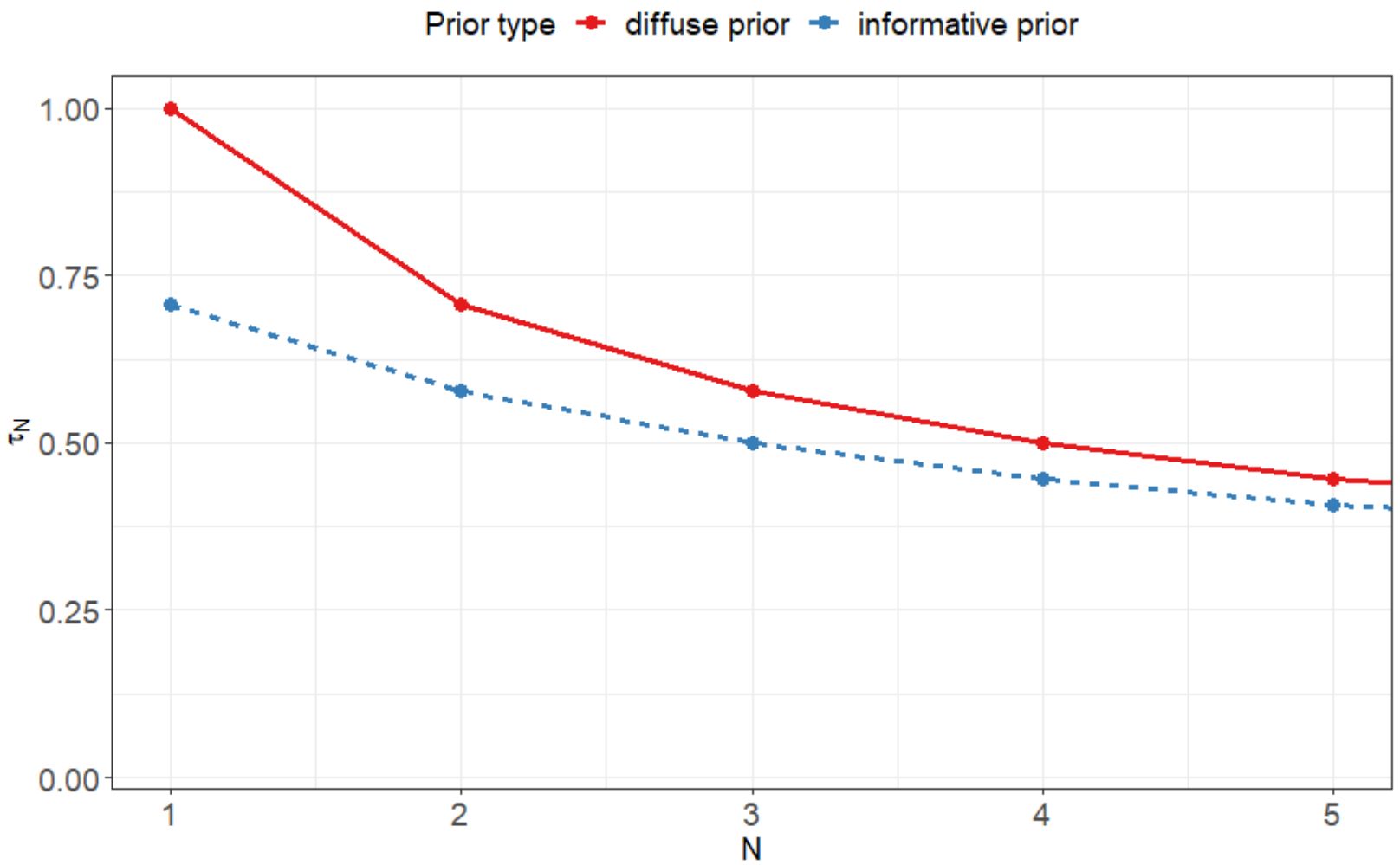
- The diffuse prior of $\tau_0 = 25$-pounds adds no information of its own.
  - Unphysical negative observations are allowed!



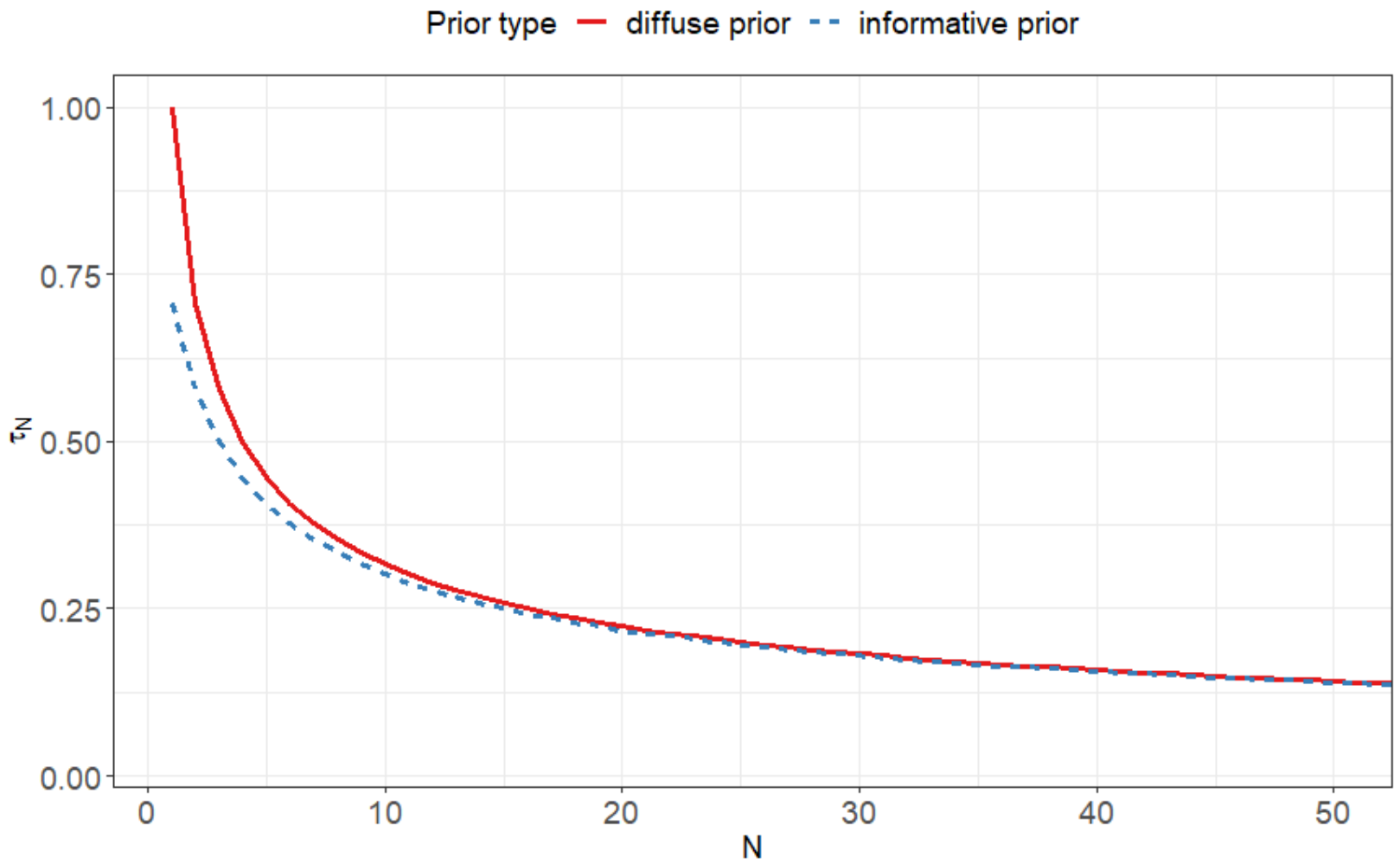
Compare posterior distributions after a few observations between the two priors



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



Now, with the basic experiment complete…**how much do I weigh?**

- In this situation, we cannot treat σ as known.

- We therefore now have 2 unknown parameters μ and σ.

Bayesian formulation with two unknowns

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

Bayesian formulation with two unknowns

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

- Most textbooks focus on factoring the prior as: $p(\mu|\sigma)p(\sigma)$.

- Allows making use of a conjugate prior on σ.

Instead, let's focus on understanding the log-posterior **surface**

- We will use independent priors:

$$p(\mu, \sigma) = p(\mu)p(\sigma)$$

- Can write the posterior then as:

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

Prior specification

- Continue to use a normal prior on μ:

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

- Use $\mu_0 = 250$ and $\tau_0 = 2$.
- ***A priori*** **I feel there's $\approx$99% probability I'm less than 255 pounds!**

Prior specification

- For σ, use a UNIFORM prior:

$$\sigma|l,u \sim \text{uniform}(\sigma|l,u)$$

- Use a lower bound, $l = 1$
- Use an upper bound, $u = 5$

Prior specification

- For σ, use a UNIFORM prior:

$$\sigma|l,u \sim \text{uniform}(\sigma|l,u)$$

- Use a lower bound, $l = 1$
- Use an upper bound, $u = 5$

So, I anticipate the noise will be higher than that for measuring a basic shape like the dumbbell.

The posterior distribution is then proportional to:

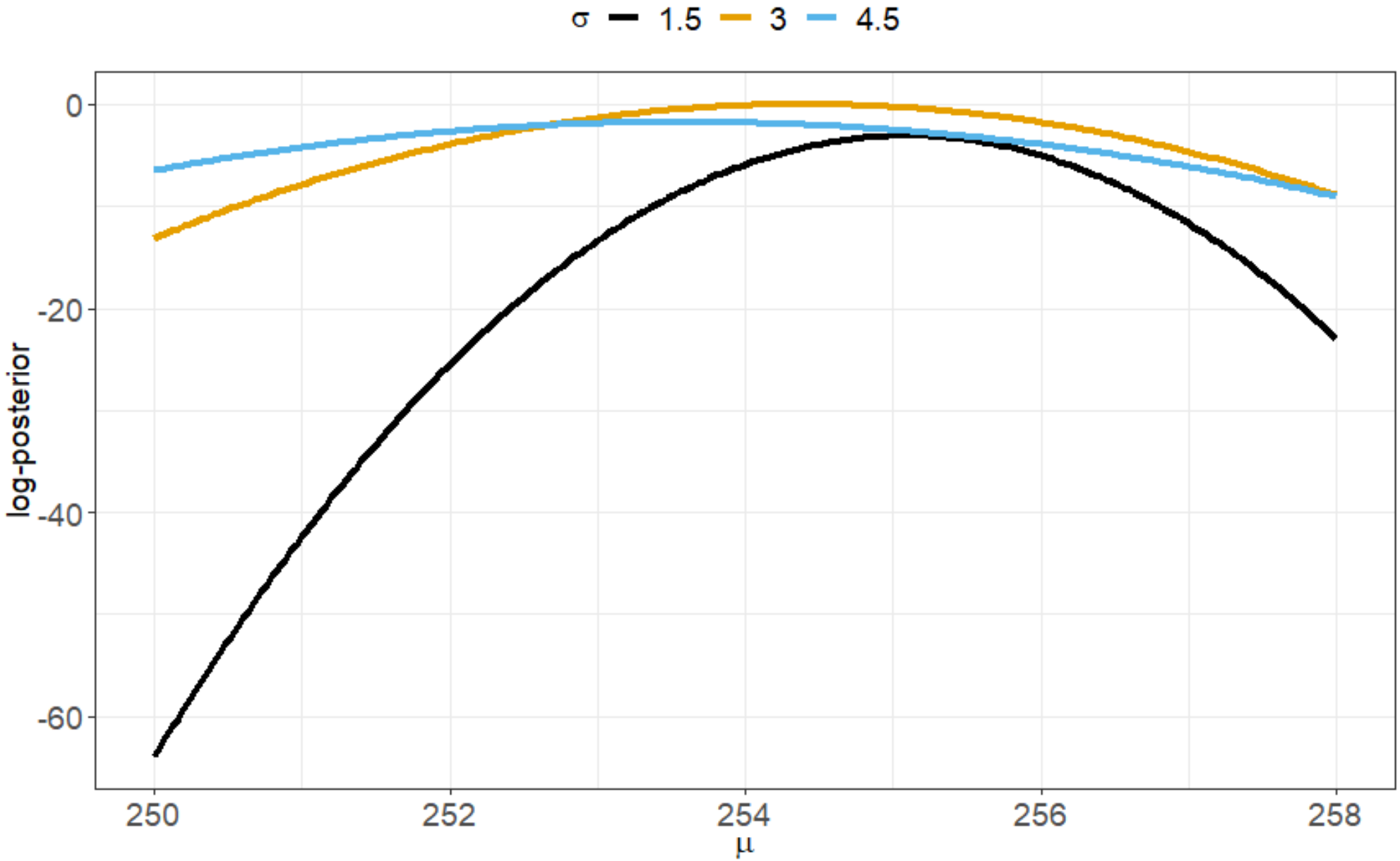
$$p(\mu, \sigma|\mathbf{x}) \propto \prod_{n=1}^{N} \{\text{normal}(x_n|\mu, \sigma)\} \cdot \text{normal}(\mu|\mu_0, \tau_0) \cdot \text{uniform}(\sigma|l, u)$$

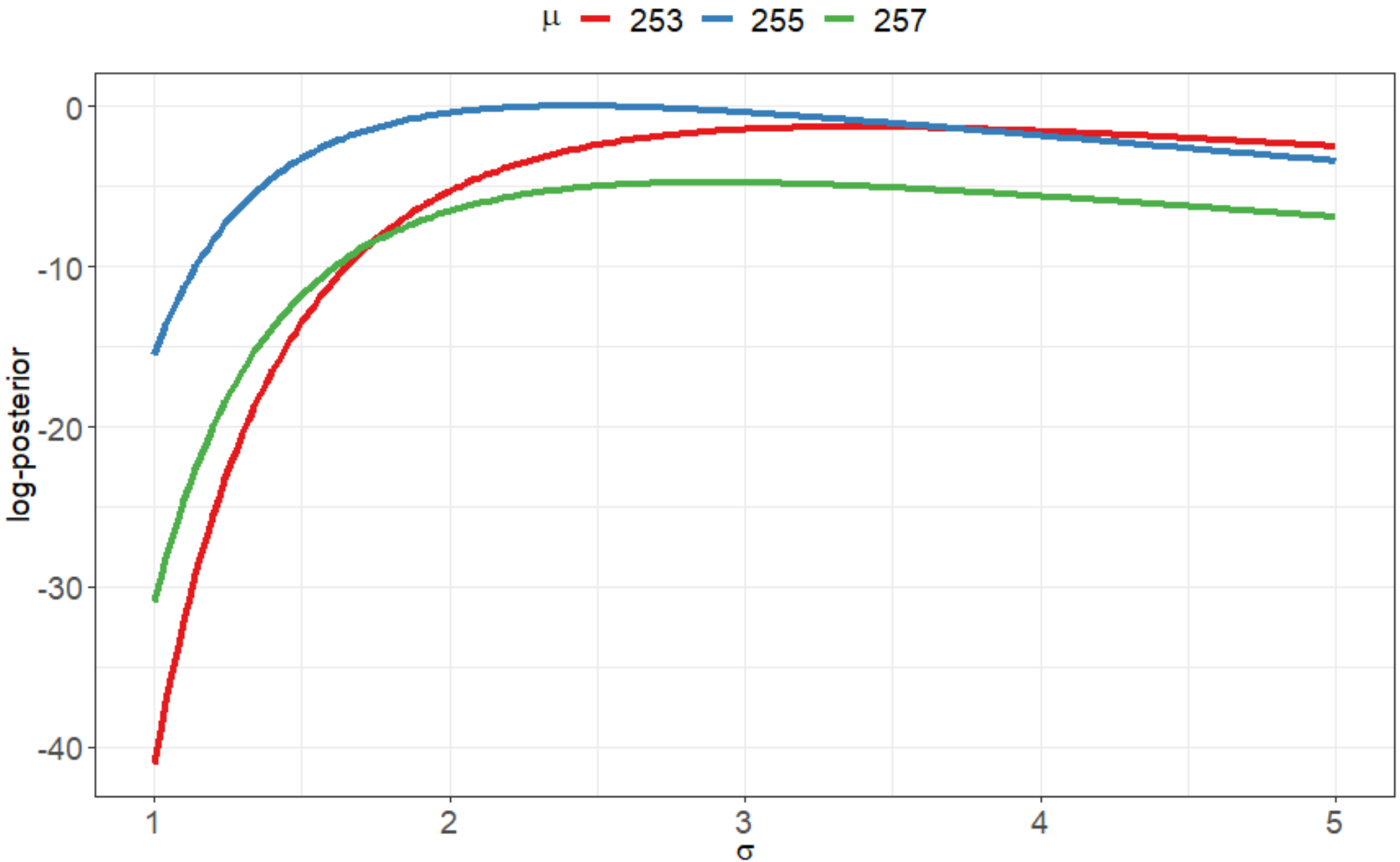
After weighing myself 10 times…what does the posterior surface look like?

Rather than working with the posterior, evaluate the log-posterior

- Plot the log-posterior as a function of μ at specific values of σ.

- Likewise, plot the log-posterior as a function of σ at specific values of μ.


σ 1.5 3 4.5
0
-20
-40
-60
log-posterior
250
252
254
256
258
μ


μ
253
255
257
0
-10
-20
-30
-40
log-posterior
1
2
3
4
5
σ

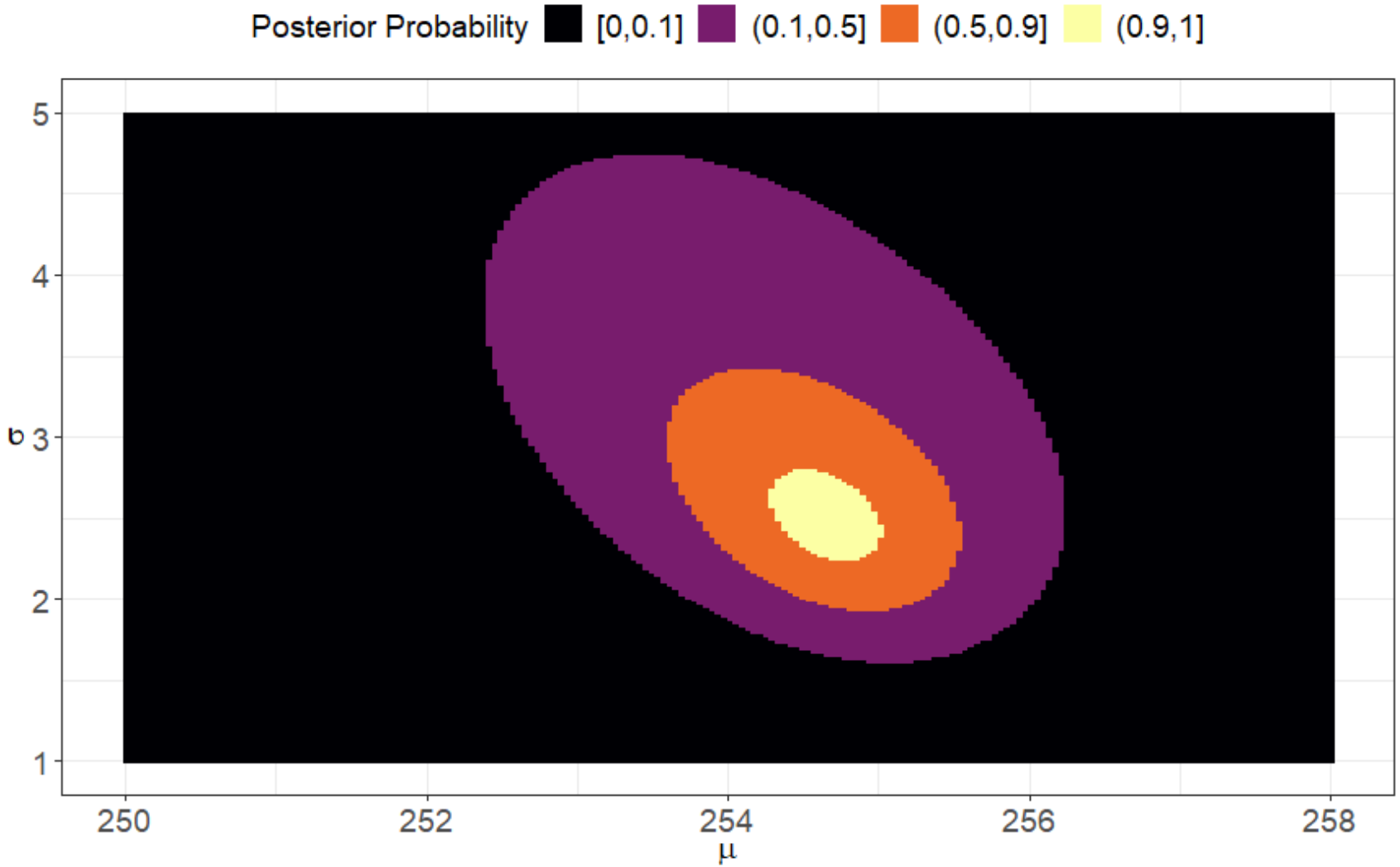
With just two unknowns, we can use the surface to directly sample from the posterior

- The previous slides visualized the trends in the log-posterior with respect to one parameter, at a few select values of the other.

- We can evaluate the log-posterior over a very fine grid of points.

- The log-posterior can then be used to estimate the posterior probability of each grid pair.

Grid-approximation or direct-sampling

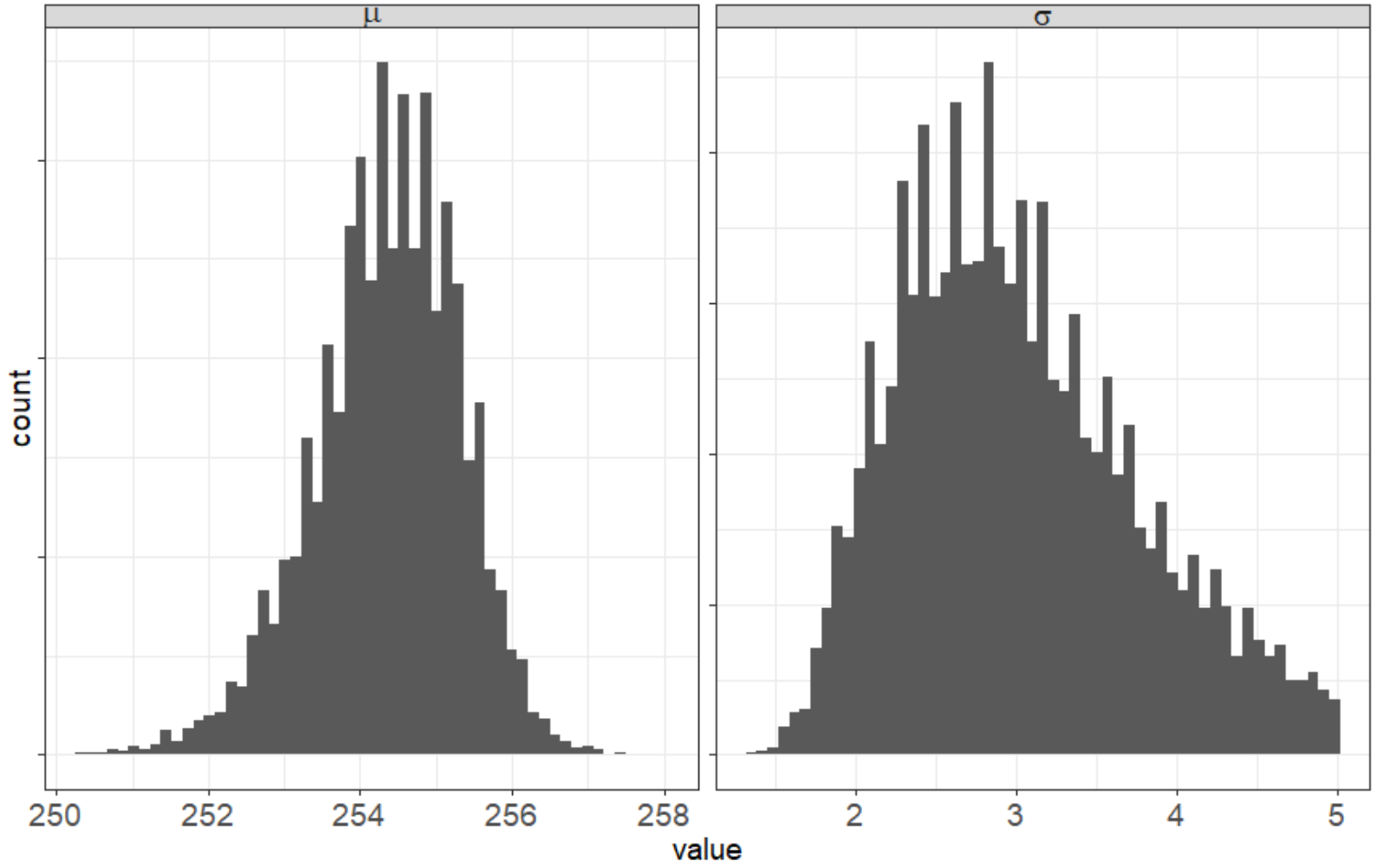


Randomly sample with replacement the grid based on the grid point posterior probability

- Can then summarize the posterior using those random samples.

- Can answer the questions of interest.

Marginal posterior distributions



Posterior probability I weigh less than 255 pounds…

- Using the grid approximate posterior samples…the posterior probability that μ is less than 255…

Posterior probability I weigh less than 255 pounds…

- Using the grid approximate posterior samples…the posterior probability that μ is less than 255…

$\approx$62% !!!!!!