STA 542: Introduction to Time Series Analysis

Lecture 4: Bayesian Forecasting Under Parameter Uncertainty

Lasse Vuursteen

Duke University

Wed Sep 2, 2026

A Bayesian analysis helped guide the search for AF447

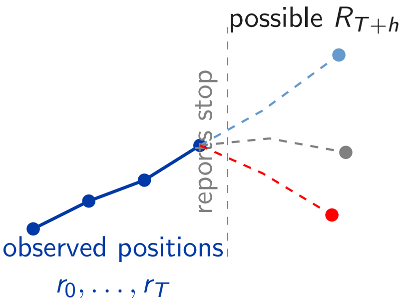
On 1 June 2009, Air France Flight 447 disappeared over the Atlantic with 228 people aboard.

The main wreckage remained missing for almost two years.

A Bayesian probability map helped guide the 2011 search that located it.

Photo: Wim Callaert, CC0, via Wikimedia Commons. Sources: BEA and Metron.

The position reports stop. Where is the aircraft?

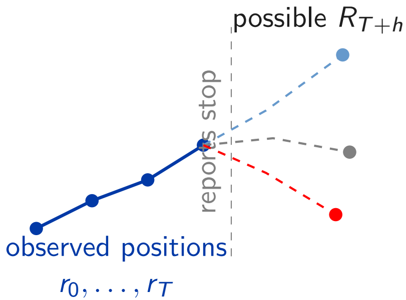


Consider the stylized model

$$R_t = R_{t-1} + \delta + Z_t, \qquad R_t, \delta \in \mathbb{R}^2.$$

- The noise law is known.
- Average speed and direction, δ, are uncertain.
- The observed positions contain information about δ.

The position reports stop. Where is the aircraft?



Consider the stylized model

$$R_t = R_{t-1} + \delta + Z_t, \qquad R_t, \delta \in \mathbb{R}^2.$$

- The noise law is known.
- Average speed and direction, δ, are uncertain.
- The observed positions contain information about δ.

How should uncertainty about δ enter a forecast of R_{T+h}?

Stylized missing-aircraft model; not a reconstruction of the AF447 analysis.

Track one coordinate first

For one coordinate, write

$$X_t = X_{t-1} + \delta + Z_t, \qquad Z_t \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Write $\mathbb{P}_\delta$ for the process law with drift fixed at δ. Under $\mathbb{P}_\delta$, Lecture 3 gives

$$X_{T+h} \mid X_{0:T} = x_{0:T} \sim N\left(x_T + h\delta,\ h\sigma^2\right).$$

Track one coordinate first

For one coordinate, write

$$X_t = X_{t-1} + \delta + Z_t, \qquad Z_t \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Write $\mathbb{P}_\delta$ for the process law with drift fixed at δ. Under $\mathbb{P}_\delta$, Lecture 3 gives

$$X_{T+h} \mid X_{0:T} = x_{0:T} \sim N\left(x_T + h\delta,\ h\sigma^2\right).$$

The known-law forecast is easy—but it changes with δ.

What should replace the unknown drift?

Two possible routes are:

1. estimate δ, plug that estimate into the known-drift forecast, and proceed as if it were known;
2. describe uncertainty about δ with a probability distribution, update that distribution using the observed path, and average forecasts.

What should replace the unknown drift?

Two possible routes are:

1. estimate δ, plug that estimate into the known-drift forecast, and proceed as if it were known;
2. describe uncertainty about δ with a probability distribution, update that distribution using the observed path, and average forecasts.

Today: the second route.

Unknown drift leaves a family of laws

The candidate process laws form the statistical model

$$\{\mathbb{P}_\delta : \delta \in \mathbb{R}\}.$$

For each fixed number δ, $\mathbb{P}_\delta$ is one complete joint law for the observed window W and future target F.

- Probabilities under that law are written $\mathbb{P}_\delta(A)$.
- Its moments are written $\mathbb{E}_\delta$ and Var_δ.

Unknown drift leaves a family of laws

The candidate process laws form the statistical model

$$\{\mathbb{P}_\delta : \delta \in \mathbb{R}\}.$$

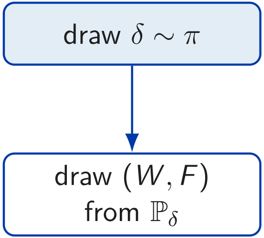
For each fixed number δ, $\mathbb{P}_\delta$ is one complete joint law for the observed window W and future target F.

- Probabilities under that law are written $\mathbb{P}_\delta(A)$.
- Its moments are written $\mathbb{E}_\delta$ and Var_δ.

The subscript selects a law; it is not a conditioning bar.

A prior creates a hierarchical law

Let π be a density for the unknown drift. The prior and the family $\{\mathbb{P}_\delta\}$ induce a new law, denoted $\mathbb{P}^\pi$:



For any event A involving W and F,

$$\mathbb{P}^\pi(A) = \int_{\mathbb{R}} \mathbb{P}_\delta(A)\pi(\delta)\, d\delta.$$

One draw of δ governs both the observed path and its future.

Three expressions answer three different questions

$\mathbb{P}_\delta(A)$	probability under one fixed-drift law,
$\mathbb{P}^\pi(A)$	probability before observing the window,
$\mathbb{P}^\pi(A \mid W = w)$	probability after observing $W = w$.

Three expressions answer three different questions

$\mathbb{P}_\delta(A)$	probability under one fixed-drift law,
$\mathbb{P}^\pi(A)$	probability before observing the window,
$\mathbb{P}^\pi(A \mid W = w)$	probability after observing $W = w$.

Posterior prediction uses the third expression when A concerns the future.

Bare $\mathbb{P}$ keeps its generic meaning from earlier lectures; it is not shorthand for $\mathbb{P}^\pi$.

The observations reweight the possible drifts

Under $\mathbb{P}_\delta$, let $p_\delta(w)$ be the density of the observed window. After $W = w$, Bayes' rule gives

$$\pi(\delta \mid w) = \frac{\overbrace{p_\delta(w)}^{\text{likelihood}}\ \overbrace{\pi(\delta)}^{\text{prior}}}{\underbrace{\int_{\mathbb{R}} p_{\delta'}(w)\pi(\delta')\,d\delta'}_{\text{makes the density integrate to one}}}.$$

The observations reweight the possible drifts

Under $\mathbb{P}_\delta$, let $p_\delta(w)$ be the density of the observed window. After $W = w$, Bayes' rule gives

$$\pi(\delta \mid w) = \frac{\overbrace{p_\delta(w)}^{\text{likelihood}} \; \overbrace{\pi(\delta)}^{\text{prior}}}{\underbrace{\int_{\mathbb{R}} p_{\delta'}(w)\pi(\delta')\, d\delta'}_{\text{makes the density integrate to one}}}.$$

Values of δ that made the observed path more plausible receive more posterior weight.

Prediction averages the known-drift forecasts

Under $\mathbb{P}_\delta$, let $p_\delta(f \mid w)$ be the conditional density of the future target. The posterior predictive density is

$$p^\pi(f \mid w) = \int_{\mathbb{R}} p_\delta(f \mid w)\, \pi(\delta \mid w)\, d\delta.$$

It averages the known-δ forecast using the posterior, not the prior.

Prediction averages the known-drift forecasts

Under $\mathbb{P}_\delta$, let $p_\delta(f \mid w)$ be the conditional density of the future target. The posterior predictive density is

$$p^\pi(f \mid w) = \int_{\mathbb{R}} p_\delta(f \mid w)\, \pi(\delta \mid w)\, d\delta.$$

It averages the known-δ forecast using the posterior, not the prior.
Under squared-error loss, the predictive mean $\mathbb{E}^\pi[F \mid W]$ is Lecture 3's best predictor under this enlarged probability model.

Four increments make the update concrete

Suppose

$$\delta \sim N(0,1), \qquad \sigma^2 = 1, \qquad X_0 = 0,$$

and the observed increments are

$$1.2, \qquad -0.4, \qquad 0.7, \qquad 0.1.$$

What are the mean observed increment $\overline{\Delta x_4}$ and the final observed level x_4?

Four increments make the update concrete

Suppose

$$\delta \sim N(0,1), \qquad \sigma^2 = 1, \qquad X_0 = 0,$$

and the observed increments are

$$1.2, \qquad -0.4, \qquad 0.7, \qquad 0.1.$$

What are the mean observed increment $\overline{\Delta x_4}$ and the final observed level x_4?

$$\overline{\Delta x_4} = 0.4, \qquad x_4 = 1.6.$$

The increments expose the drift

Define $\Delta X_t := X_t - X_{t-1}$. Then

$$\Delta X_t = \delta + Z_t.$$

Under the fixed-drift law $\mathbb{P}_\delta$,

$$\Delta X_1, \ldots, \Delta X_T \overset{\text{iid}}{\sim} N(\delta, \sigma^2).$$

The increments expose the drift

Define $\Delta X_t := X_t - X_{t-1}$. Then

$$\Delta X_t = \delta + Z_t.$$

Under the fixed-drift law $\mathbb{P}_\delta$,

$$\Delta X_1, \ldots, \Delta X_T \overset{\text{iid}}{\sim} N(\delta, \sigma^2).$$

For the observed path, write $\Delta x_t := x_t - x_{t-1}$. The posterior kernel is

$$\pi(\delta \mid x_{0:4}) \propto \exp\left[-\frac{1}{2}\left\{\delta^2 + \sum_{t=1}^{4}(\Delta x_t - \delta)^2\right\}\right].$$

Completing the square identifies the posterior

The part of the exponent involving δ satisfies

$$\begin{aligned} \delta^2 + \sum_{t=1}^{4} (\Delta x_t - \delta)^2 &= 5\delta^2 - 3.2\delta + C \\ &= 5(\delta - 0.32)^2 + C'. \end{aligned}$$

Completing the square identifies the posterior

The part of the exponent involving δ satisfies

$$\begin{aligned} \delta^2 + \sum_{t=1}^{4} (\Delta x_t - \delta)^2 &= 5\delta^2 - 3.2\delta + C \\ &= 5(\delta - 0.32)^2 + C'. \end{aligned}$$

Therefore, under $\mathbb{P}^\pi$,

$$\boxed{\delta \mid X_{0:4} = x_{0:4} \sim N(0.32, 0.20).}$$

The data pull the posterior mean from the prior mean 0 toward the sample mean increment 0.4.

Precision determines the compromise

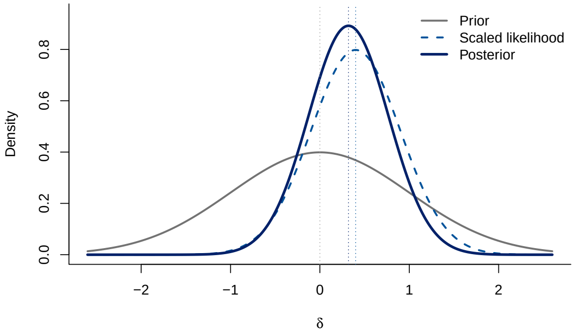
Precision is reciprocal variance. Here,

$$\underbrace{5}_{\text{posterior precision}} = \underbrace{1}_{\text{prior}} + \underbrace{4}_{\text{data}}.$$

Hence

$$0.32 = \frac{1}{5}(0) + \frac{4}{5}(0.4).$$

The more precise source receives more weight.



The same update works for any observed path

With prior $\delta \sim N(m_0, s_0^2)$ and known innovation variance σ^2, let

$$\overline{\Delta x}_T = \frac{1}{T}\sum_{t=1}^{T}(x_t - x_{t-1}).$$

Completing the same square gives, under $\mathbb{P}^\pi$,

$$\delta \mid X_{0:T} = x_{0:T} \sim N(m_T, s_T^2),$$

where

$$s_T^2 = \left(\frac{1}{s_0^2} + \frac{T}{\sigma^2}\right)^{-1}, \qquad m_T = s_T^2\left(\frac{m_0}{s_0^2} + \frac{T\overline{\Delta x}_T}{\sigma^2}\right).$$

Now forecast five steps ahead

Under $\mathbb{P}_\delta$,

$$X_9 \mid X_{0:4} = x_{0:4} \sim N(1.6 + 5\delta, 5).$$

What happens under $\mathbb{P}^\pi$, after averaging over $\delta \mid X_{0:4} = x_{0:4} \sim N(0.32, 0.20)$?

Now forecast five steps ahead

Under $\mathbb{P}_\delta$,

$$X_9 \mid X_{0:4} = x_{0:4} \sim N(1.6 + 5\delta, 5).$$

What happens under $\mathbb{P}^\pi$, after averaging over $\delta \mid X_{0:4} = x_{0:4} \sim N(0.32, 0.20)$?

$$\mathbb{E}^\pi[X_9 \mid X_{0:4} = x_{0:4}] = 1.6 + 5(0.32) = 3.2,$$

$$\mathrm{Var}^\pi(X_9 \mid X_{0:4} = x_{0:4}) = 5 + 5^2(0.20) = 10.$$

Thus, under $\mathbb{P}^\pi$, $X_9 \mid X_{0:4} = x_{0:4} \sim N(3.2, 10)$.

Plug-in prediction loses half the variance here

$$\underbrace{10}_{\text{posterior predictive variance}} = \underbrace{5}_{\text{future innovations}} + \underbrace{25(0.20)}_{\text{drift uncertainty}} .$$

Plug-in prediction loses half the variance here

$$\underbrace{10}_{\text{posterior predictive variance}} = \underbrace{5}_{\text{future innovations}} + \underbrace{25(0.20)}_{\text{drift uncertainty}}.$$

Plugging in $\hat{\delta} = 0.32$ instead gives $N(3.2, 5)$.

	Variance	Pointwise 95% interval for X_9
Posterior predictive	10	$[-3.00,\ 9.40]$
Plug-in	5	$[-1.18,\ 7.58]$

The same variance decomposition holds in general

The identity $10 = 5 + 5$ is one instance of a general result. For square-integrable F, define

$$m_\delta(W) := \mathbb{E}_\delta[F \mid W], \qquad v_\delta(W) := \mathrm{Var}_\delta(F \mid W), \qquad \bar{m}(W) := \mathbb{E}^\pi[F \mid W].$$

Then

$$\mathrm{Var}^\pi(F \mid W) = \underbrace{\int_{\mathbb{R}} v_\delta(W)\pi(\delta \mid W)\, d\delta}_{\text{future process variation}} + \underbrace{\int_{\mathbb{R}} \{m_\delta(W) - \bar{m}(W)\}^2 \pi(\delta \mid W)\, d\delta}_{\text{uncertainty about the law}}.$$

In this random walk, plug-in prediction keeps the first term exactly and omits the second.

Drift uncertainty grows faster with the horizon

In general, under $\mathbb{P}^{\pi}$, if

$$\delta \mid X_{0:T} = x_{0:T} \sim N(m_T, s_T^2),$$

then

$$X_{T+h} \mid X_{0:T} = x_{0:T} \sim N\left(x_T + hm_T, \underbrace{h\sigma^2}_{\text{future innovations}} + \underbrace{h^2 s_T^2}_{\text{drift uncertainty}}\right).$$

Drift uncertainty grows faster with the horizon

In general, under $\mathbb{P}^\pi$, if

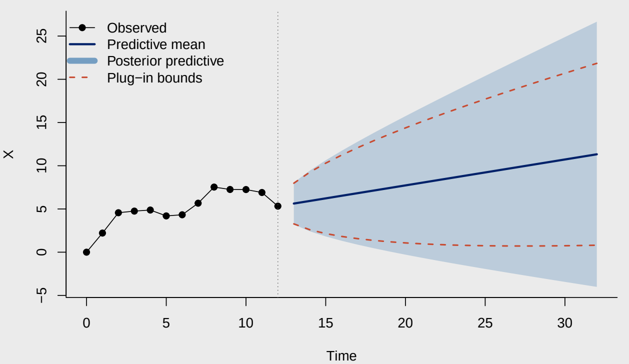
$$\delta \mid X_{0:T} = x_{0:T} \sim N(m_T, s_T^2),$$

then

$$X_{T+h} \mid X_{0:T} = x_{0:T} \sim N\left(x_T + hm_T, \underbrace{h\sigma^2}_{\text{future innovations}} + \underbrace{h^2 s_T^2}_{\text{drift uncertainty}}\right).$$

The centers agree with plug-in here only because the known-drift forecast is linear in δ.

The posterior predictive fan is wider



Blue is the pointwise 95% posterior predictive fan; red dashed lines are plug-in bounds.

Simulated random walk; seed 542. Each vertical interval concerns one horizon, not the whole path simultaneously.

How should we simulate one future path?

We need draws of the unknown drift and of future innovations.

Which procedure follows the stated model?

1. Draw a new drift at every future horizon.
2. Draw one drift and retain it for the entire future path.

How should we simulate one future path?

We need draws of the unknown drift and of future innovations.

Which procedure follows the stated model?

1. Draw a new drift at every future horizon.
2. Draw one drift and retain it for the entire future path.

One unknown drift governs the entire process: choose 2.

Draw one drift per path

To generate one posterior predictive path:

1. draw one $\delta^* \sim N(m_T, s_T^2)$;
2. draw fresh $Z^*_{T+1}, \ldots, Z^*_{T+H}$ independently;
3. set

$$X^*_{T+j} = x_T + \sum_{i=1}^{j} (\delta^* + Z^*_{T+i}).$$

Draw one drift per path

To generate one posterior predictive path:

1. draw one $\delta^* \sim N(m_T, s_T^2)$;
2. draw fresh $Z^*_{T+1}, \dots, Z^*_{T+H}$ independently;
3. set

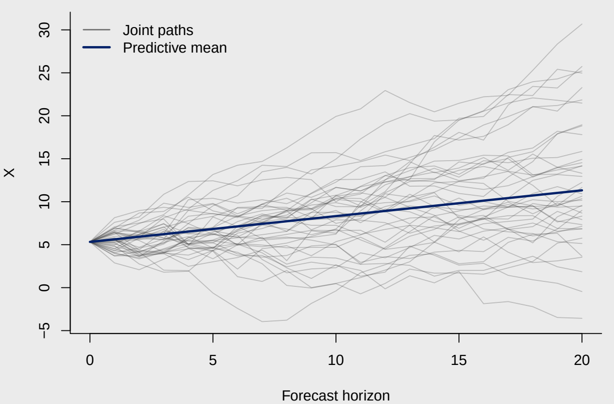
$$X^*_{T+j} = x_T + \sum_{i=1}^{j} (\delta^* + Z^*_{T+i}).$$

For two distinct future increments,

$$\mathrm{Cov}^{\pi}(\Delta X_{T+j}, \Delta X_{T+k} \mid X_{0:T} = x_{0:T}) = s_T^2.$$

A high posterior drift draw lifts the entire path.

Joint draws answer path questions



Joint draws can answer questions such as

$$\mathbb{P}^{\pi}\left(\max_{1\leq h\leq H} X_{T+h} > c \mid X_{0:T} = x_{0:T}\right).$$

Count the fraction of simulated paths that cross c at least once.

Simulated posterior predictive paths; seed 542. Each path uses one posterior drift draw.

What does posterior predictive "95%" mean?

Suppose that, for every possible observed window w, we construct $C_{\mathrm{B}}(w)$ so that

$$\mathbb{P}^{\pi}\{F \in C_{\mathrm{B}}(w) \mid W = w\} = 0.95.$$

- The observed window w is fixed.
- The unknown parameter varies according to its posterior.
- The future target varies according to its conditional law.

What does posterior predictive “95%” mean?

Suppose that, for every possible observed window w, we construct $C_{\mathrm{B}}(w)$ so that

$$\mathbb{P}^{\pi}\{F \in C_{\mathrm{B}}(w) \mid W = w\} = 0.95.$$

- The observed window w is fixed.
- The unknown parameter varies according to its posterior.
- The future target varies according to its conditional law.

Does this imply 95% coverage when the true drift is held fixed?

Holding the drift fixed asks a different question

For a fixed value δ, define

$$c_{\mathrm{B}}(\delta) := \mathbb{P}_{\delta}\{F \in C_{\mathrm{B}}(W)\}.$$

In this repeated-sample experiment:

- the same δ governs every repetition;
- a new observed window W and future target F are generated;
- the procedure recomputes $C_{\mathrm{B}}(W)$ from each new window.

This is the interval's fixed-parameter coverage function.

The prior averages fixed-drift experiments

Recall the hierarchical law $\mathbb{P}^\pi$: first draw $\delta \sim \pi$, then draw (W, F) from $\mathbb{P}_\delta$. Therefore, for any event A involving W and F,

$$\mathbb{P}^\pi(A) = \int_{\mathbb{R}} \mathbb{P}_\delta(A)\, \pi(\delta)\, d\delta.$$

Apply this continuous law of total probability to $A = \{F \in C_{\mathrm{B}}(W)\}$:

$$\mathbb{P}^\pi\{F \in C_{\mathrm{B}}(W)\} = \int_{\mathbb{R}} c_{\mathrm{B}}(\delta)\pi(\delta)\, d\delta.$$

Posterior 95% implies prior-average 95%

Conditioning instead on the observed window and using the tower property,

$$\begin{aligned}\mathbb{P}^\pi\{F \in C_{\mathrm{B}}(W)\} &= \mathbb{E}^\pi[\mathbb{P}^\pi\{F \in C_{\mathrm{B}}(W) \mid W\}] \\ &= 0.95.\end{aligned}$$

Combining the two calculations gives

$$\boxed{\int_{\mathbb{R}} c_{\mathrm{B}}(\delta)\pi(\delta)\, d\delta = 0.95.}$$

The word *average* is essential.

Must every fixed-drift coverage equal 0.95?

We know that

$$\int_{\mathbb{R}} c_{\mathrm{B}}(\delta)\pi(\delta)\, d\delta = 0.95.$$

Does it follow that

$$c_{\mathrm{B}}(\delta) = 0.95 \qquad \text{for every } \delta?$$

Must every fixed-drift coverage equal 0.95?

We know that

$$\int_{\mathbb{R}} c_{\mathrm{B}}(\delta)\pi(\delta)\, d\delta = 0.95.$$

Does it follow that

$$c_{\mathrm{B}}(\delta) = 0.95 \qquad \text{for every } \delta?$$

No. An average can be 0.95 while its ingredients differ.

One observed increment already shows the difference

This is the random-walk model in miniature: W is one observed increment, and F is the next increment.

For each fixed δ, let $\mathbb{P}_\delta$ be the law under which

$$W = \delta + Z_1, \qquad F = \delta + Z_2,$$

where $Z_1, Z_2 \overset{\text{iid}}{\sim} N(0,1)$.

Now place the prior $\delta \sim N(0,1)$. Under the resulting law $\mathbb{P}^\pi$, normal updating gives

$$\delta \mid W = w \sim N\left(\frac{w}{2}, \frac{1}{2}\right), \qquad F \mid W = w \sim N\left(\frac{w}{2}, \frac{3}{2}\right).$$

Thus a 95% posterior predictive interval is

$$C_{\mathrm{B}}(w) = \frac{w}{2} \pm 1.96\sqrt{\frac{3}{2}}.$$

Shrinkage creates uneven fixed-drift coverage

Under $\mathbb{P}_\delta$,

$$F - \frac{W}{2} \sim N\left(\frac{\delta}{2}, \frac{5}{4}\right).$$

The Bayesian interval is centered at $W/2$, because the prior shrinks the forecast toward zero.

Shrinkage creates uneven fixed-drift coverage

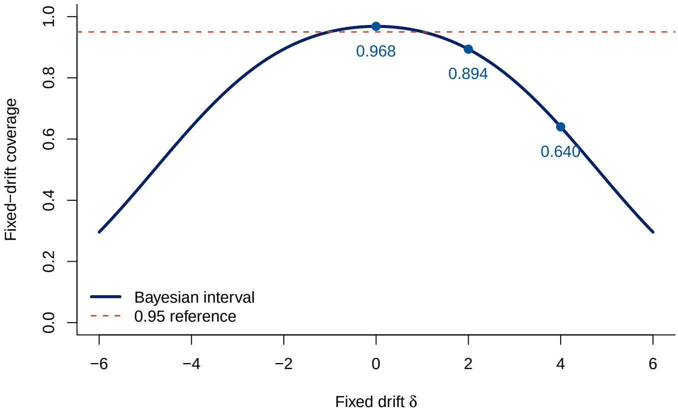
Under $\mathbb{P}_\delta$,

$$F - \frac{W}{2} \sim N\left(\frac{\delta}{2}, \frac{5}{4}\right).$$

The Bayesian interval is centered at $W/2$, because the prior shrinks the forecast toward zero.

If the fixed drift is far from the prior center, that shrinkage repeatedly pulls the interval in the wrong direction.

Prior-average 95% is not pointwise 95%



$$c_{\mathrm{B}}(0) = 0.968,$$
$$c_{\mathrm{B}}(2) = 0.894,$$
$$c_{\mathrm{B}}(4) = 0.640.$$

Yet

$$\int_{\mathbb{R}} c_{\mathrm{B}}(\delta)\pi(\delta)\,d\delta = 0.95.$$

A fixed-drift pivot gives exact coverage

Under every fixed-drift law $\mathbb{P}_\delta$,

$$F - W = Z_2 - Z_1 \sim N(0,2).$$

Therefore

$$C_{\mathrm{H}}(W) = W \ \pm \ 1.96\sqrt{2}$$

satisfies

$$\mathbb{P}_\delta\{F \in C_{\mathrm{H}}(W)\} = 0.95 \qquad \text{for every } \delta \in \mathbb{R}.$$

A fixed-drift pivot gives exact coverage

Under every fixed-drift law $\mathbb{P}_\delta$,

$$F - W = Z_2 - Z_1 \sim N(0, 2).$$

Therefore

$$C_{\mathrm{H}}(W) = W \ \pm \ 1.96\sqrt{2}$$

satisfies

$$\mathbb{P}_\delta\{F \in C_{\mathrm{H}}(W)\} = 0.95 \qquad \text{for every } \delta \in \mathbb{R}.$$

Unlike the Bayesian interval, this interval has the same guarantee under each fixed-drift law.

Honesty is a worst-case fixed-law guarantee

For a general family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, an interval rule C is honest at level 0.95 if

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta\{F \in C(W)\} \geq 0.95.$$

Honesty is a worst-case fixed-law guarantee

For a general family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, an interval rule C is honest at level 0.95 if

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta\{F \in C(W)\} \geq 0.95.$$

Neither calibration statement is model-free:

- posterior predictive probability is computed under $\mathbb{P}^\pi$;
- honesty ranges only over the stated fixed-parameter family.

In our example, both claims rely on independent Gaussian noise with known variance. Neither establishes that this model describes the world.

Posterior prediction averages fixed-law forecasts

$$p^{\pi}(f \mid w) = \int_{\Theta} p_{\theta}(f \mid w)\pi(\theta \mid w)\, d\theta.$$

The subscript in p_θ selects a fixed law; the superscript in p^π names the prior-induced law.

The predictive law carries parameter uncertainty into:

- point forecasts;
- prediction intervals and predictive densities;
- probabilities of events involving an entire future path.

No fitted value first has to be treated as the true parameter.

What Bayesian forecasting requires

The analyst must supply:

1. a family of fixed-parameter laws for the observations and future; the observed-data densities $p_\theta(w)$ supply the likelihood;
2. a prior for the unknown parameter.

What Bayesian forecasting requires

The analyst must supply:

1. a family of fixed-parameter laws for the observations and future; the observed-data densities $p_\theta(w)$ supply the likelihood;
2. a prior for the unknown parameter.

Predictions can depend materially on either choice when the observed path is weakly informative.

Bayesian updating propagates uncertainty *within* the chosen model; it does not show that the model or prior is appropriate.

Conjugacy is convenient, not essential

The normal likelihood and normal prior produced a normal posterior in closed form. This is conjugacy.

In other models, posterior predictive integrals may require:

- numerical integration;
- simulation methods such as Markov chain Monte Carlo;
- checks that the numerical approximation is accurate.

The trade-off is a unified treatment of parameter uncertainty at the price of a fuller probability model, a prior, and sometimes substantial computation.

Next class: what can one path reveal?

A finite-sample posterior can be computed without showing that it concentrates as the path grows.

Next we ask:

- Can time averages learn the mean, variance, or dependence?
- Can different targets behave differently?
- What does ergodicity guarantee?