

lec07

February 10, 2026

1 Lecture 7: Bayesian Inference

Data 145, Spring 2026: Evidence and Uncertainty

Instructors: Ani Adhikari, William Fithian

Please run the setup cell below before reading.

```
[1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

plt.style.use('fivethirtyeight')
%matplotlib inline

# Color scheme for Bayesian lectures
# Black = data (likelihood)
# Blue = beliefs (prior and posterior; prior dashed, posterior solid)
# Red = asymptotic approximations (BuM normal approx, dashed)
COLOR_LIKELIHOOD = 'black'          # Data / likelihood
COLOR_PRIOR = 'steelblue'           # Prior belief (dashed)
COLOR_POSTERIOR = 'steelblue'       # Posterior belief (solid)
COLOR_APPROX = 'firebrick'          # Asymptotic approximations (dashed)
COLOR_TRUE = '#000000'              # True parameter value

# Shades of blue for comparing multiple posteriors (prior washout plot)
MULTI_BLUES = ['#1b4f72', '#2e86c1', '#5dade2', '#85c1e9']
```

1.1 Introduction: From Decision Theory to Bayesian Inference

In Lecture 6, we studied the **decision theory** framework for evaluating estimators:

- A **loss function** $L(\theta, a)$ measures estimation error
- The **risk function** $R(\theta; T) = E_{\theta}[L(\theta, T(X))]$ averages loss over the sampling distribution
- No single estimator minimizes risk for all θ

A key result was:

The estimator minimizing **average-case risk** $\int_0^1 \text{MSE}_p(T) dp$ for the binomial model turned out to be the **posterior mean** $E[p | X] = (X+1)/(n+2)$ — Laplace’s estimator!

We arrived at Bayesian statistics through a purely frequentist argument: “which estimator has the smallest average MSE?”

1.1.1 From unweighted to weighted averages

But why did we take an *unweighted* average over $p \in [0, 1]$? Maybe we care more about performance when p is near 0.5 than when it’s near 0 or 1 — after all, values near 0.5 come up more often in practice. So we might prefer a *weighted* average: $\int_0^1 \text{MSE}_p(T) \cdot \pi(p) dp$ for some weight function $\pi(p)$.

Even more fundamentally: for a parameter like an exponential rate $\lambda > 0$ — which could be any positive number — there’s no such thing as an unweighted average over $(0, \infty)$. We *need* a weight function. This weight function π is exactly the **prior distribution**.

1.1.2 Today’s roadmap

Today, we develop the **Bayesian perspective** more fully:

1. The general framework: prior, likelihood, and posterior
2. Concrete examples: Beta-Binomial, Gamma-Exponential, and Normal-Normal
3. Conjugate priors and their interpretations
4. Why the likelihood is all that matters
5. Large-sample behavior: the prior washes out

1.1.3 Why go Bayesian?

The Bayesian framework gives us the full **posterior distribution**, not just a point estimate. With the posterior in hand, we can:

- Report a **credible interval**: an interval $[a, b]$ such that $P(\theta \in [a, b] | X) = 0.95$. This is a direct probability statement about θ — “given the data and our prior, there is a 95% probability that θ lies in this interval” — unlike a confidence interval, which is a statement about the procedure.
- Make probability statements about θ (e.g., $P(p > 0.5 | X)$)
- Make predictions about future data

1.1.4 A word of caution

Going Bayesian means layering on additional assumptions. The prior $\pi(\theta)$ is typically difficult to check — unlike a likelihood model (which we can assess with goodness-of-fit tests, residual plots, etc.), we usually only observe θ indirectly through the data, so there’s no direct way to verify whether our prior is reasonable. When the sample size isn’t very large, the choice of prior can seriously impact our inferences. We’ll come back to these questions — where do priors come from, and what does it even mean for θ to “have a distribution”? — in Lecture 8.

1.2 1. The Bayesian Framework

1.2.1 The Setup

The ingredients: - **Prior:** $\theta \sim \pi(\theta)$ — our beliefs about θ before seeing data - **Likelihood:** $X | \theta \sim f_\theta(x)$ — the data-generating process, given θ - **Posterior:** $\theta | X \sim \pi(\theta | X)$ — our updated beliefs after seeing data

1.2.2 Bayes' Rule

The posterior is computed via Bayes' rule:

$$\pi(\theta | x) = \frac{f_\theta(x) \cdot \pi(\theta)}{\int_{\Theta} f_u(x) \cdot \pi(u) du}$$

The denominator $m(x) = \int_{\Theta} f_u(x) \cdot \pi(u) du$ is the **marginal likelihood**. It doesn't depend on θ , so:

$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} : \quad \pi(\theta | x) \propto_\theta f_\theta(x) \cdot \pi(\theta)$

This proportionality (in θ) is the key trick: we can identify the posterior just by recognizing the functional form in θ , without computing the normalizing constant.

1.2.3 The Bayes Estimator

The **Bayes risk** of an estimator $T(X)$ with respect to prior π is:

$$r_\pi(T) = \int_{\Theta} R(\theta; T) \pi(\theta) d\theta = E[L(\theta, T(X))]$$

Recall from Lecture 6: we wanted an estimator that was best for every θ , but that's impossible. The Bayes approach averages over θ instead. And it turns out to have a very nice property: **we can optimize for every X value separately.**

For squared error loss $L(\theta, a) = (\theta - a)^2$:

$$r_\pi(T) = E[E[(\theta - T(X))^2 | X]]$$

The inner expectation depends on T only through $T(X)$ at the observed X , so we can minimize it **separately for each X** , and this automatically minimizes the overall Bayes risk.

By the bias-variance decomposition (with θ as the random quantity):

$$E[(\theta - T(X))^2 | X] = \text{Var}(\theta | X) + (E[\theta | X] - T(X))^2$$

The first term doesn't depend on T ; the second is minimized when:

$T^*(X) = E[\theta | X] \quad (\text{the posterior mean})$

This is remarkable: we couldn't find an estimator best for every θ , but we *can* find one that's best for every X .

What if we use a different loss function instead of squared error? The Bayes estimator will be a different summary of the posterior — on the homework, you'll work out what it is for absolute error loss and other examples.

1.3 2. Beta-Binomial Conjugacy

1.3.1 Recap: The Uniform Prior (Lecture 6)

With $p \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$ and $X | p \sim \text{Binomial}(n, p)$: - Posterior: $p | X \sim \text{Beta}(X + 1, n - X + 1)$ - Bayes estimator: $E[p | X] = (X + 1)/(n + 2)$

1.3.2 General Beta Prior

Now suppose $p \sim \text{Beta}(\alpha, \beta)$ with density $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$.

The posterior is:

$$\pi(p | x) \propto_p f_p(x) \cdot \pi(p) = \binom{n}{x} p^x (1-p)^{n-x} \cdot p^{\alpha-1} (1-p)^{\beta-1} \propto_p p^{x+\alpha-1} (1-p)^{n-x+\beta-1}$$

This is proportional (in p) to a $\text{Beta}(x + \alpha, n - x + \beta)$ density. Since two different densities can't be proportional to each other (both integrate to 1, so the proportionality constant must be 1), we conclude:

$$\boxed{p | X \sim \text{Beta}(X + \alpha, n - X + \beta)}$$

Notice what happened: we started with a Beta prior and ended up with a Beta posterior. When this occurs — when the posterior belongs to the same family as the prior — we say the prior is **conjugate** to the likelihood. We'll see more examples of this shortly.

1.3.3 Posterior Mean as Weighted Average

Recall the **pseudodata interpretation** from Lecture 6: the $\text{Beta}(\alpha, \beta)$ prior is like imagining we already observed $\alpha - 1$ successes and $\beta - 1$ failures before collecting any data. The posterior adds the real data on top of this pseudodata.

The posterior mean is:

$$E[p | X] = \frac{X + \alpha}{n + \alpha + \beta} = \underbrace{\frac{n}{n + \alpha + \beta}}_w \cdot \underbrace{\frac{X}{n}}_{\hat{p}_{\text{MLE}}} + \underbrace{\frac{\alpha + \beta}{n + \alpha + \beta}}_{1-w} \cdot \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}}$$

The posterior mean is a **weighted average** of the MLE and the prior mean. The quantity $\alpha + \beta$ acts like a **prior sample size**.

- As $n \rightarrow \infty$: $w \rightarrow 1$, so the posterior mean \rightarrow MLE (data overwhelms the prior)
- As $\alpha + \beta \rightarrow \infty$: $w \rightarrow 0$, so the posterior mean \rightarrow prior mean (strong prior dominates)

1.3.4 Visualizing Prior, Likelihood, and Posterior

In each plot below, we show the prior, the likelihood (rescaled so its peak matches the posterior, for visual comparison), and the posterior. Notice how the posterior always sits between the prior and likelihood, closer to whichever carries more information.

```
[2]: def plot_beta_binomial(ax, n, x, alpha, beta, title=None):
    """Plot prior, likelihood, and posterior for Beta-Binomial model."""
    p_grid = np.linspace(0.001, 0.999, 500)

    # Prior: Beta(alpha, beta)
    prior = stats.beta.pdf(p_grid, alpha, beta)

    # Likelihood:  $p^x (1-p)^{(n-x)}$ , rescaled
    log_lik = x * np.log(p_grid) + (n - x) * np.log(1 - p_grid)
    lik = np.exp(log_lik - np.max(log_lik)) # normalize peak to 1

    # Posterior: Beta(x + alpha, n - x + beta)
    post_a, post_b = x + alpha, n - x + beta
    posterior = stats.beta.pdf(p_grid, post_a, post_b)

    # Rescale likelihood to match posterior peak height
    lik_scaled = lik * np.max(posterior)

    # Prior: blue dashed. Likelihood: black solid. Posterior: blue solid.
    ax.plot(p_grid, prior, color=COLOR_PRIOR, linewidth=2.5, linestyle='--',
            label=f'Prior: Beta({alpha}, {beta})')
    ax.plot(p_grid, lik_scaled, color=COLOR_LIKELIHOOD, linewidth=2.5,
            label='Likelihood (rescaled)')
    ax.plot(p_grid, posterior, color=COLOR_POSTERIOR, linewidth=2.5,
            label=f'Posterior: Beta({post_a}, {post_b})')

    # Mark MLE and posterior mean
    mle = x / n
    post_mean = post_a / (post_a + post_b)
    ax.axvline(mle, color=COLOR_LIKELIHOOD, linestyle=':', alpha=0.6,
    ↪ linewidth=1.5)
    ax.axvline(post_mean, color=COLOR_POSTERIOR, linestyle=':', alpha=0.6,
    ↪ linewidth=1.5)

    if title:
        ax.set_title(title, fontsize=12, fontweight='bold')
    ax.set_xlabel('$p$', fontsize=11)
    ax.set_xlim(0, 1)
    ax.set_ylim(0, None)
    ax.legend(fontsize=9, loc='upper left')
```

```

fig, axes = plt.subplots(2, 2, figsize=(14, 10))

plot_beta_binomial(axes[0, 0], n=16, x=12, alpha=1, beta=1,
                    title='Uniform prior, n=16, X=12')
plot_beta_binomial(axes[0, 1], n=16, x=12, alpha=5, beta=5,
                    title='Beta(5,5) prior, n=16, X=12')
plot_beta_binomial(axes[1, 0], n=100, x=75, alpha=1, beta=1,
                    title='Uniform prior, n=100, X=75')
plot_beta_binomial(axes[1, 1], n=16, x=12, alpha=20, beta=20,
                    title='Strong Beta(20,20) prior, n=16, X=12')

plt.suptitle('Beta-Binomial: Prior, Likelihood, and Posterior',
             fontsize=14, fontweight='bold', y=1.02)
plt.tight_layout()
plt.show()

```

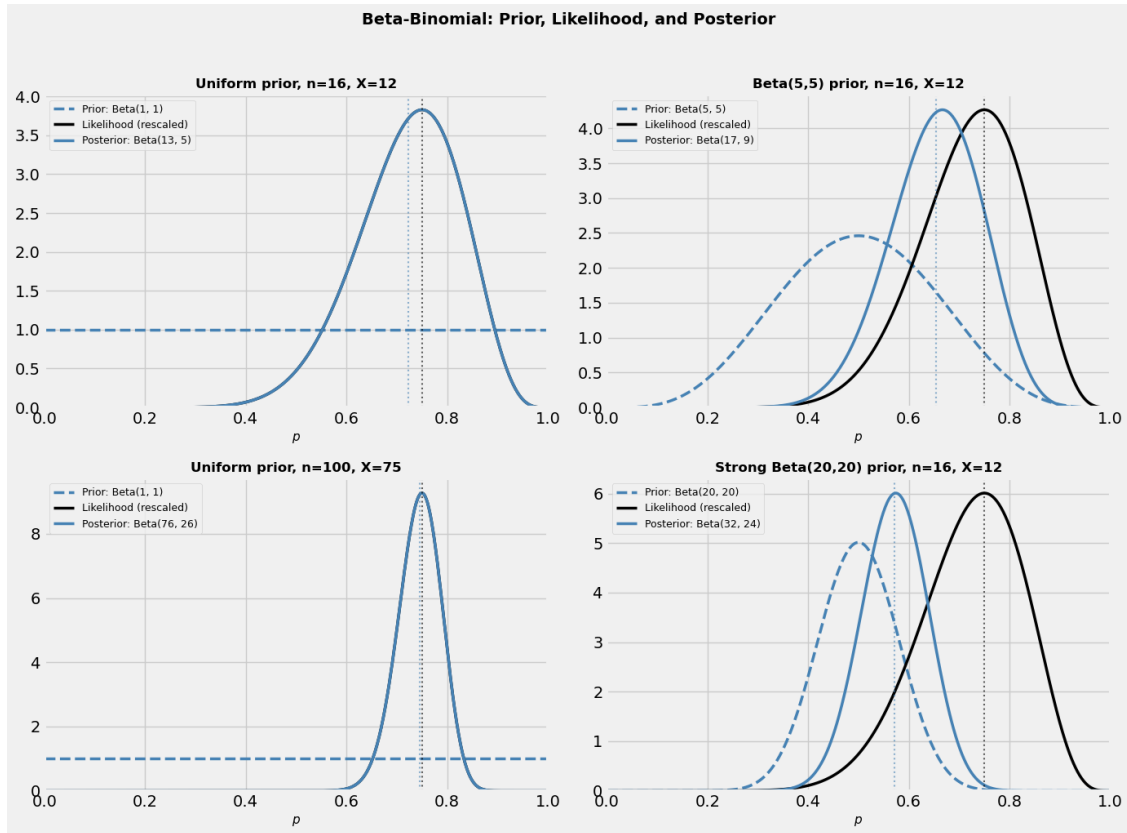


Figure 1. Prior (blue dashed), likelihood (black solid, rescaled), and posterior (blue solid) for the Beta-Binomial model under four scenarios. Top left: a weak uniform prior with moderate data — the posterior tracks the likelihood. Top right: a moderate Beta(5,5) prior pulls the posterior toward 0.5. Bottom left: with $n = 100$ observations, the posterior concentrates tightly around the MLE regardless of the prior. Bottom right: a strong Beta(20,20) prior (prior “sample size” 40)

dominates the $n = 16$ data, pulling the posterior toward 0.5. Dotted vertical lines show the MLE (black) and posterior mean (blue).

1.3.5 Observations

1. **Weak prior, moderate data** (top left): With a uniform prior ($\alpha + \beta = 2$), the prior is nearly flat, so posterior \propto likelihood \times prior \approx likelihood \times const. The posterior is *proportional* to the likelihood — not just close to it!
2. **Moderate prior, moderate data** (top right): The Beta(5,5) prior pulls the posterior toward 0.5. The posterior mean compromises between the MLE (0.75) and the prior mean (0.5), weighted by their respective “sample sizes” ($n = 16$ vs. $\alpha + \beta = 10$).
3. **Weak prior, lots of data** (bottom left): With $n = 100$ observations, the posterior is tightly concentrated around the MLE. The prior barely matters.
4. **Strong prior at wrong location** (bottom right): The Beta(20,20) prior has “prior sample size” 40, exceeding the actual sample size of 16. The prior pulls the posterior strongly toward 0.5, away from the MLE.

1.4 3. Gamma-Exponential: Back to the Earthquake Data

In Lecture 1, we modeled the interarrival times of California $M \geq 4.0$ earthquakes as $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$, where λ is the daily rate. The MLE was $\hat{\lambda} = n / \sum X_i = 1/\bar{X}$.

Recall that we checked this modeling assumption in Lecture 1 by plotting the histogram of interarrival times and comparing it to the best-fitting exponential density. The exponential model looked reasonable — so let’s take it as given and put a prior on λ .

1.4.1 The Gamma Distribution

The **Gamma**(α, β) **distribution** (shape $\alpha > 0$, rate $\beta > 0$) has density

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

with mean α/β and variance α/β^2 . The special case $\alpha = 1$ gives the $\text{Exponential}(\beta)$ distribution.

1.4.2 Choosing a Prior

We’ll try three different Exponential priors, all with $\alpha = 1$ (very weak: just one pseudo-observation), but with very different prior guesses for the rate:

Prior	Prior mean	Interpretation
Gamma(1, 1) = Exp(1)	1/day	One earthquake per day

Prior	Prior mean	Interpretation
Gamma(1, 20) = Exp(1/20)	0.05/day	One earthquake every 20 days
Gamma(1, 365) = Exp(1/365)	0.0027/day	One earthquake per year

These priors span a factor of 365 in their guesses for the rate! Yet as we'll see, with $n = 614$ observations the posteriors are virtually indistinguishable.

Unlike the likelihood model, which we could check against the data, there's no way to directly verify whether any prior is reasonable — we only observe λ indirectly. But with this much data, it doesn't matter.

1.4.3 The Posterior

The likelihood for $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$ is:

$$f_\lambda(x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum x_i}$$

The Gamma(α, β) prior has density $\pi(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$, so:

$$\pi(\lambda | x) \propto_\lambda f_\lambda(x) \cdot \pi(\lambda) = \lambda^n e^{-\lambda \sum x_i} \cdot \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{(n+\alpha)-1} e^{-(\beta+\sum x_i)\lambda}$$

This is proportional to a Gamma($n + \alpha, \beta + \sum x_i$) density:

$$\boxed{\lambda | X_1, \dots, X_n \sim \text{Gamma}(n + \alpha, \beta + \sum X_i)}$$

The Gamma prior is **conjugate** to the Exponential likelihood — the same pattern as Beta-Binomial.

1.4.4 Posterior Mean as Weighted Average

The posterior mean is:

$$E[\lambda | X] = \frac{n + \alpha}{\beta + \sum X_i} = \frac{n + \alpha}{\beta + n\bar{X}}$$

We can rewrite this as a weighted average of the MLE $\hat{\lambda} = 1/\bar{X}$ and the prior mean α/β . Dividing numerator and denominator by $\beta + n\bar{X}$:

$$E[\lambda | X] = \underbrace{\frac{n}{\beta + n\bar{X}} \cdot \frac{1}{\bar{X}}}_{\approx w \cdot \hat{\lambda}} + \underbrace{\frac{\alpha}{\beta + n\bar{X}}}_{\approx (1-w) \cdot \alpha/\beta}$$

More precisely, defining $w = n\bar{X}/(\beta + n\bar{X})$:

$$E[\lambda | X] = w \cdot \hat{\lambda}_{\text{MLE}} + (1 - w) \cdot \frac{\alpha}{\beta}$$

where $w \rightarrow 1$ as $n \rightarrow \infty$ (data overwhelms the prior). Here β plays the role of the “prior sample size” in units of $\sum X_i$.

1.4.5 Pseudodata Interpretation

The prior $\text{Gamma}(\alpha, \beta)$ is like having already observed α pseudo-events over a total pseudo-time of β , giving a prior rate of α/β . The posterior adds the real data: n observed events over total time $\sum X_i$.

1.4.6 Derived Quantities

A key advantage of the Bayesian approach is that once we have the posterior for λ , we can compute posterior distributions for *any* function of λ — for example: - **Probability of ≥ 1 earthquake in a week:** $1 - e^{-7\lambda}$ - **90th percentile of earthquakes in a year:** the 90th percentile of $\text{Poisson}(365\lambda)$

Uncertainty propagates automatically through nonlinear transformations, without needing the delta method!

```
[3]: # Load earthquake data from Lecture 1
eq_data = pd.read_csv('../demos/lec01_earthquakes/data/
    ↪california_earthquakes_declustered.csv')

# Filter to mainshocks and compute interarrival times
mainshocks = eq_data[eq_data['is_mainshock']].sort_values('time').
    ↪reset_index(drop=True)

# Parse timestamps (format='ISO8601' handles mixed fractional seconds)
timestamps = pd.to_datetime(mainshocks['time'], format='ISO8601')
interarrivals = timestamps.diff().dt.total_seconds().dropna().values / 86400

n_eq = len(interarrivals)
sum_x = np.sum(interarrivals)
xbar = np.mean(interarrivals)
mle_lambda = 1 / xbar

print(f"Number of mainshocks: {len(mainshocks)}")
print(f"Number of interarrival times: {n_eq}")
print(f"Mean interarrival time: {xbar:.2f} days")
print(f"MLE rate: {mle_lambda:.4f} per day")

# --- Three priors spanning a factor of 365 in prior mean ---
# All are Gamma(1, beta) = Exp(beta), so alpha=1 (one pseudo-observation)
priors_eq = [
    (1, 1, 'Gamma(1, 1): 1/day', MULTI_BLUES[0]),
    (1, 20, 'Gamma(1, 20): 1/20 days', MULTI_BLUES[1]),
    (1, 365, 'Gamma(1, 365): 1/year', MULTI_BLUES[2]),
]
```

```

print(f"\n{'Prior':<30s} {'Prior mean':<12s} {'Post. mean':<12s} {'w (MLE_
↳weight)'}")
print("-" * 70)
for alpha_p, beta_p, label, _ in priors_eq:
    pa = n_eq + alpha_p
    pb = beta_p + sum_x
    pm = pa / pb
    w = n_eq * xbar / (beta_p + n_eq * xbar)
    print(f"{'label':<30s} {'alpha_p/beta_p:<12.4f} {'pm:<12.4f} {'w:.6f}")

# Use the middle prior (Gamma(1,20)) as the "main" posterior for Panels 2-3
alpha_main, beta_main = 1, 20
post_alpha = n_eq + alpha_main
post_beta = beta_main + sum_x
post_dist = stats.gamma(a=post_alpha, scale=1/post_beta)

# --- Three-panel figure ---
fig, axes = plt.subplots(1, 3, figsize=(18, 5))

# Panel 1: Three priors (dashed) and posteriors (solid) for lambda
ax = axes[0]
lam_grid = np.linspace(0.025, 0.055, 500)

# First pass: compute max posterior height for rescaling priors
max_post_height = 0
for alpha_p, beta_p, label, color in priors_eq:
    pa = n_eq + alpha_p
    pb = beta_p + sum_x
    pdf = stats.gamma.pdf(lam_grid, a=pa, scale=1/pb)
    max_post_height = max(max_post_height, np.max(pdf))

# Plot priors (rescaled, dashed) and posteriors (solid)
for alpha_p, beta_p, label, color in priors_eq:
    pa = n_eq + alpha_p
    pb = beta_p + sum_x
    post_pdf = stats.gamma.pdf(lam_grid, a=pa, scale=1/pb)
    prior_pdf = stats.gamma.pdf(lam_grid, a=alpha_p, scale=1/beta_p)

    # Rescale prior so its peak is ~30% of max posterior height
    if np.max(prior_pdf) > 0:
        prior_rescaled = prior_pdf * (max_post_height / np.max(prior_pdf)) * 0.3
    else:
        prior_rescaled = prior_pdf

    ax.plot(lam_grid, prior_rescaled, color=color, linewidth=1.5,
↳linestyle='--', alpha=0.5)
    ax.plot(lam_grid, post_pdf, color=color, linewidth=2.5, label=label)

```

```

ax.axvline(mle_lambda, color=COLOR_LIKELIHOOD, linestyle=':', linewidth=1.5,
            label=f'MLE: {mle_lambda:.4f}')

# 95% credible interval (from middle prior - all are virtually the same)
ci_lo_lam = post_dist.ppf(0.025)
ci_hi_lam = post_dist.ppf(0.975)
posterior_pdf = post_dist.pdf(lam_grid)
mask = (lam_grid >= ci_lo_lam) & (lam_grid <= ci_hi_lam)
ax.fill_between(lam_grid[mask], posterior_pdf[mask], alpha=0.1,
                color=COLOR_POSTERIOR)

ax.set_xlabel(r'$\lambda$ (earthquakes per day)', fontsize=11)
ax.set_ylabel('Density', fontsize=11)
ax.set_title(r'Posterior for rate $\lambda$' + '\n(three priors)', fontsize=12,
            fontweight='bold')
ax.legend(fontsize=8)

# Panel 2:  $P(\text{earthquake in a week}) = 1 - \exp(-7 * \lambda)$ 
# Analytic PDF via change of variables:  $p = 1 - \exp(-7 * \lambda)$ ,  $\lambda = -\log(1-p)/7$ 
#  $f_P(p) = f_{\Lambda}(-\log(1-p)/7) * 1/(7*(1-p))$ 
ax = axes[1]
prob_week_mle = 1 - np.exp(-7 * mle_lambda)

prob_grid = np.linspace(0.001, 0.999, 500)
lam_of_p = -np.log(1 - prob_grid) / 7
dlam_dp = 1 / (7 * (1 - prob_grid))
prob_week_pdf = post_dist.pdf(lam_of_p) * dlam_dp

# Focus on the region where the density is nontrivial
pw_support = prob_grid[(prob_week_pdf > 1e-3)]
pw_lo, pw_hi = pw_support[0], pw_support[-1]
pw_mask = (prob_grid >= pw_lo) & (prob_grid <= pw_hi)

ax.plot(prob_grid[pw_mask], prob_week_pdf[pw_mask], color=COLOR_POSTERIOR,
        linewidth=2.5,
        label='Posterior density')
ax.axvline(prob_week_mle, color=COLOR_LIKELIHOOD, linestyle=':', linewidth=1.5,
            label=f'MLE: {prob_week_mle:.3f}')

# 95% credible interval from posterior quantiles of lambda, transformed
ci_lo_pw = 1 - np.exp(-7 * ci_lo_lam)
ci_hi_pw = 1 - np.exp(-7 * ci_hi_lam)

ci_mask = pw_mask & (prob_grid >= ci_lo_pw) & (prob_grid <= ci_hi_pw)
ax.fill_between(prob_grid[ci_mask], prob_week_pdf[ci_mask], alpha=0.15,

```

```

color=COLOR_POSTERIOR)

ax.set_xlabel('$P(\geq 1$ earthquake in a week)$', fontsize=11)
ax.set_ylabel('Density', fontsize=11)
ax.set_title('Posterior for $P$(EQ in a week)', fontsize=12, fontweight='bold')
ax.legend(fontsize=9)

# Panel 3: 90th percentile of Poisson(365*lambda) - discrete, needs MC
ax = axes[2]
np.random.seed(42)
post_samples = np.random.gamma(post_alpha, 1/post_beta, size=50000)
pct90_annual = np.array([stats.poisson.ppf(0.9, 365 * lam) for lam in
    post_samples])
pct90_mle = stats.poisson.ppf(0.9, 365 * mle_lambda)

# Since this is discrete, use a bar chart of the posterior PMF
vals, counts = np.unique(pct90_annual, return_counts=True)
pmf = counts / len(pct90_annual)
ax.bar(vals, pmf, color=COLOR_POSTERIOR, alpha=0.6, edgecolor='white', width=0.
    8,
        label='Posterior PMF')
ax.axvline(pct90_mle, color=COLOR_LIKELIHOOD, linestyle=':', linewidth=1.5,
        label=f'MLE: {pct90_mle:.0f}')

ci_lo_p90 = np.percentile(pct90_annual, 2.5)
ci_hi_p90 = np.percentile(pct90_annual, 97.5)
ax.set_xlabel('90th percentile of annual EQ count', fontsize=11)
ax.set_ylabel('Posterior probability', fontsize=11)
ax.set_title("Posterior for 90th pctile\nof annual count", fontsize=12,
    fontweight='bold')
ax.legend(fontsize=9)

plt.suptitle('Gamma-Exponential: Earthquake Rate and Derived Quantities',
        fontsize=14, fontweight='bold', y=1.02)
plt.tight_layout()
plt.show()

```

Number of mainshocks: 614
 Number of interarrival times: 613
 Mean interarrival time: 27.34 days
 MLE rate: 0.0366 per day

Prior	Prior mean	Post. mean	w (MLE weight)
Gamma(1, 1): 1/day	1.0000	0.0366	0.999940
Gamma(1, 20): 1/20 days	0.0500	0.0366	0.998808
Gamma(1, 365): 1/year	0.0027	0.0359	0.978688

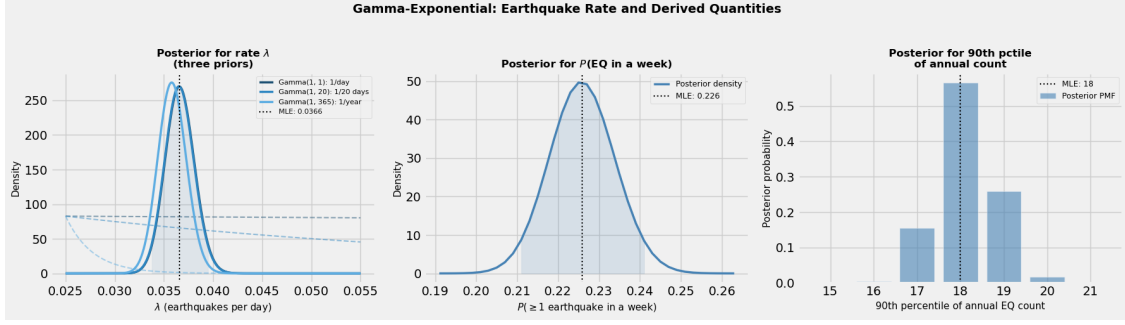


Figure 2. Posterior inference for California earthquake rates using the Gamma-Exponential conjugate model. Left: priors (dashed, rescaled) and posteriors (solid) for the daily rate λ from three different priors — $\text{Gamma}(1, 1)$ (one earthquake per day), $\text{Gamma}(1, 20)$ (one per 20 days), and $\text{Gamma}(1, 365)$ (one per year). Despite prior means spanning a factor of 365, the three posteriors are virtually indistinguishable with $n = 614$ interarrival times. Center: the posterior density of $P(\geq 1 \text{ EQ in a week}) = 1 - e^{-7\lambda}$, obtained via the change-of-variables formula. Right: the posterior distribution of the 90th percentile of the annual earthquake count (a discrete quantity, computed by Monte Carlo).

1.5 4. Normal-Normal Conjugacy

1.5.1 Setup

Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$ with σ^2 known.

Prior: $\theta \sim N(\mu_0, \tau_0^2)$

1.5.2 Deriving the Posterior

The likelihood is:

$$f_{\theta}(x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right) \propto_{\theta} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right)$$

Expanding the sum: $\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$. The first term doesn't involve θ , so:

$$f_{\theta}(x_1, \dots, x_n) \propto_{\theta} \exp\left(-\frac{n(\bar{x} - \theta)^2}{2\sigma^2}\right)$$

(This tells us the likelihood depends on the data only through \bar{x} — we'll come back to this point later.)

Now applying Bayes' rule:

$$\pi(\theta \mid x_1, \dots, x_n) \propto_{\theta} \exp\left(-\frac{n(\bar{x} - \theta)^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{(\theta - \mu_0)^2}{2\tau_0^2}\right)$$

Both factors are Gaussian in θ , so the product is also Gaussian. We need to combine the two quadratics in θ and complete the square.

Show algebra: completing the square

The exponent (ignoring the $-1/2$) is:

$$\frac{n}{\sigma^2}(\bar{x} - \theta)^2 + \frac{1}{\tau_0^2}(\theta - \mu_0)^2$$

Expanding:

$$= \frac{n}{\sigma^2}(\theta^2 - 2\bar{x}\theta + \bar{x}^2) + \frac{1}{\tau_0^2}(\theta^2 - 2\mu_0\theta + \mu_0^2)$$

Collecting terms in θ :

$$= \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right) \theta^2 - 2 \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right) \theta + \text{const}$$

Define the **posterior precision** $\frac{1}{\tau_1^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$ and complete the square:

$$= \frac{1}{\tau_1^2}(\theta - \mu_1)^2 + \text{const}$$

where $\mu_1 = \tau_1^2 \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right)$.

The result:

$$\boxed{\theta \mid X_1, \dots, X_n \sim N(\mu_1, \tau_1^2)}$$

1.5.3 The Precision Formulation

The result is cleanest in terms of **precision** ($= 1/\text{variance}$):

Posterior precision = Prior precision + Data precision

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

The posterior mean is a **precision-weighted average**:

$$\mu_1 = \underbrace{\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau_0^2}}_w \cdot \bar{X} + \underbrace{\frac{1/\tau_0^2}{n/\sigma^2 + 1/\tau_0^2}}_{1-w} \cdot \mu_0$$

This has **exactly the same structure** as the Beta-Binomial: - The posterior mean is a weighted average of the MLE (\bar{X}) and the prior mean (μ_0) - More data (larger n) \Rightarrow more weight on \bar{X} - More precise prior (smaller τ_0^2) \Rightarrow more weight on μ_0

The Normal prior is conjugate to the Normal likelihood, just as the Beta prior is conjugate to the Binomial. This pattern is **universal** across conjugate families.

```
[4]: def plot_normal_normal(ax, xbar, n, sigma2, mu0, tau0_sq, title=None):
    """Plot prior, likelihood, and posterior for Normal-Normal model."""
    # Posterior parameters
    data_prec = n / sigma2
    prior_prec = 1 / tau0_sq
    post_prec = data_prec + prior_prec
    tau1_sq = 1 / post_prec
    mu1 = tau1_sq * (xbar * data_prec + mu0 * prior_prec)

    # Grid for plotting
    all_means = [xbar, mu0, mu1]
    all_sds = [np.sqrt(sigma2 / n), np.sqrt(tau0_sq), np.sqrt(tau1_sq)]
    lo = min(m - 4 * s for m, s in zip(all_means, all_sds))
    hi = max(m + 4 * s for m, s in zip(all_means, all_sds))
    theta_grid = np.linspace(lo, hi, 500)

    prior = stats.norm.pdf(theta_grid, mu0, np.sqrt(tau0_sq))
    lik = stats.norm.pdf(theta_grid, xbar, np.sqrt(sigma2 / n))
    posterior = stats.norm.pdf(theta_grid, mu1, np.sqrt(tau1_sq))

    # Prior: blue dashed. Likelihood: black solid. Posterior: blue solid.
    ax.plot(theta_grid, prior, color=COLOR_PRIOR, linewidth=2.5, linestyle='--',
            label=f'Prior: N({mu0}, {tau0_sq})')
    ax.plot(theta_grid, lik, color=COLOR_LIKELIHOOD, linewidth=2.5,
            label=f'Likelihood: N({xbar}, {sigma2/n:.2g})')
    ax.plot(theta_grid, posterior, color=COLOR_POSTERIOR, linewidth=2.5,
            label=f'Posterior: N({mu1:.2f}, {tau1_sq:.3f})')

    ax.axvline(xbar, color=COLOR_LIKELIHOOD, linestyle=':', alpha=0.5,
    ↪linewidth=1.5)
    ax.axvline(mu1, color=COLOR_POSTERIOR, linestyle=':', alpha=0.5,
    ↪linewidth=1.5)

    if title:
        ax.set_title(title, fontsize=12, fontweight='bold')
    ax.set_xlabel(r'$\theta$', fontsize=11)
    ax.set_ylim(0, None)
    ax.legend(fontsize=9)

fig, axes = plt.subplots(1, 3, figsize=(16, 5))

plot_normal_normal(axes[0], xbar=3.0, n=10, sigma2=4.0, mu0=0.0, tau0_sq=1.0,
                    title='Balanced: prior and data\ncomparable precision')
plot_normal_normal(axes[1], xbar=3.0, n=100, sigma2=4.0, mu0=0.0, tau0_sq=1.0,
                    title='Lots of data (n=100):\nposterior near MLE')
plot_normal_normal(axes[2], xbar=3.0, n=10, sigma2=4.0, mu0=0.0, tau0_sq=0.1,
```

```

title=r'Precise prior ( $\tau_0^2=0.1$ ):' + '\nposterior near_\u
prior')

plt.suptitle('Normal-Normal: Prior, Likelihood, and Posterior',
            fontsize=14, fontweight='bold', y=1.02)
plt.tight_layout()
plt.show()

```

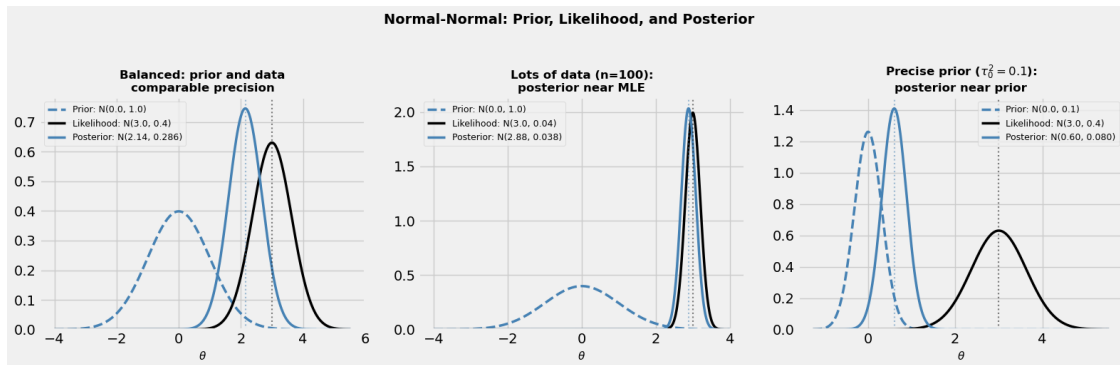


Figure 3. Prior (blue dashed), likelihood (black solid), and posterior (blue solid) for the Normal-Normal model. Left: when the prior and data have comparable precision, the posterior sits between them. Center: with $n = 100$ the data precision overwhelms the prior, so the posterior nearly coincides with the likelihood. Right: a very precise prior ($\tau_0^2 = 0.1$) dominates even $n = 10$ observations, keeping the posterior near the prior mean. In all panels, the posterior is narrower than both the prior and the likelihood — combining information always reduces uncertainty.

1.5.4 Observations

The same qualitative pattern as the Beta-Binomial:

- The posterior always sits **between** the prior and the likelihood, closer to whichever is more precise (narrower).
- The posterior is always **narrower** than both the prior and the likelihood — combining information reduces uncertainty.
- With lots of data, the posterior is essentially the likelihood.
- With a very precise prior, the posterior stays near the prior mean even if the data disagree.

1.5.5 Conjugate Priors: The Common Pattern

All three examples above share the same structure: the posterior belongs to the **same family** as the prior, with updated parameters. When this happens, we say the prior family is **conjugate** to the likelihood.

Likelihood	Conjugate prior	Posterior
Binomial(n, p)	Beta(α, β)	Beta($X + \alpha, n - X + \beta$)
Exponential(λ)	Gamma(α, β)	Gamma($n + \alpha, \beta + \sum X_i$)
Normal(θ, σ^2)	Normal(μ_0, τ_0^2)	Normal(μ_1, τ_1^2)

In every case, the posterior mean is a weighted average of the MLE and the prior mean, and the prior parameters have a **pseudodata** interpretation: imaginary observations that get combined with the real data. You'll see another conjugate pair — Gamma-Poisson — on the homework.

1.6 5. The Likelihood Is All That Matters

1.6.1 A Key Observation

In the posterior formula $\pi(\theta | x) \propto_\theta f_\theta(x) \cdot \pi(\theta)$, the data x enter **only through the likelihood function** $f_\theta(x)$ (viewed as a function of θ for fixed x).

If two data sets x and x' produce the same likelihood function — that is, $f_\theta(x) = c \cdot f_\theta(x')$ for all θ and some constant c — they lead to the **same posterior distribution**, regardless of the prior.

1.6.2 Sufficient Statistics

Sometimes a single function of the data — a statistic $T(X)$ — is all we need to compute the likelihood function (up to a multiplicative constant not depending on θ). When that happens, $T(X)$ carries all the information in the data about θ , and we call it **sufficient**.

For example, suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$. The likelihood is:

$$f_\lambda(x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum x_i}$$

The only thing about the data that matters (as a function of λ) is $\sum x_i$ — or equivalently, \bar{X} . Two data sets with the same sample mean produce the same likelihood and therefore the same posterior.

Similarly, for the Normal model the likelihood depends on the data only through \bar{x} .

We'll study sufficiency more formally in Stat 210A. For now, the key point is that for all of our conjugate models, there's a simple sufficient statistic, and the posterior depends on the data only through it.

Although we've introduced sufficiency in a Bayesian context, it's equally important for frequentist statistics. There's a *sufficiency principle* that says good estimators and other inference procedures should depend on the data only through a sufficient statistic. If you look back at some of the failed estimators from the last several weeks, a common theme is that they don't follow this principle.

1.7 6. Large-Sample Behavior: The Prior Washes Out

1.7.1 The Main Idea

As the sample size n grows: - The likelihood $f_\theta(x)$ becomes more and more **concentrated** around the MLE $\hat{\theta}$ - The prior $\pi(\theta)$ is a fixed function — it doesn't change with n . In any small neighborhood around the MLE, the prior is roughly constant (just some number $\pi(\hat{\theta})$), while the likelihood has a sharp peak whose height grows with n - So the posterior is dominated by the likelihood: **the prior washes out**

1.7.2 Asymptotic Normality of the Posterior

We can make this precise using the quadratic approximation to the log-likelihood from Lectures 4–5:

$$\ell(\theta) = \log f_\theta(x) \approx \ell(\hat{\theta}) - \frac{nI(\hat{\theta})}{2}(\theta - \hat{\theta})^2$$

So the likelihood looks like a Gaussian:

$$f_\theta(x) \approx e^{\ell(\hat{\theta})} \cdot \exp\left(-\frac{nI(\hat{\theta})}{2}(\theta - \hat{\theta})^2\right)$$

Since $\pi(\theta) \approx \pi(\hat{\theta})$ near $\hat{\theta}$:

$$\pi(\theta | x) \propto_\theta f_\theta(x) \cdot \pi(\theta) \approx \text{const} \cdot \exp\left(-\frac{nI(\hat{\theta})}{2}(\theta - \hat{\theta})^2\right)$$

This is a Normal density:

$$\pi(\theta | X) \approx N\left(\hat{\theta}_{\text{MLE}}, \frac{1}{nI(\hat{\theta}_{\text{MLE}})}\right)$$

This is called the **Bernstein–von Mises theorem** (stated here informally). The posterior variance $1/(nI(\hat{\theta}))$ is the same as the asymptotic variance of the MLE!

For large samples, **Bayesian and frequentist inference agree**: - Posterior mean \approx MLE - Posterior standard deviation \approx standard error of MLE - 95% posterior credible interval \approx asymptotic normal 95% confidence interval $\hat{\theta} \pm 1.96/\sqrt{nI(\hat{\theta})}$

Let's see this in action.

```
[5]: # Demonstrate prior washing out: multiple priors converging as n grows
true_p = 0.7
sample_sizes = [5, 20, 100, 500]

priors = [
    (1, 1, 'Uniform'),
    (2, 2, 'Beta(2,2)'),
    (0.5, 0.5, 'Jeffreys'),
```

```

    (10, 2, 'Beta(10,2)'),
]

fig, axes = plt.subplots(1, 4, figsize=(18, 4.5))
np.random.seed(42)

for ax, n in zip(axes, sample_sizes):
    x = np.random.binomial(n, true_p)
    mle = x / n
    p_grid = np.linspace(0.001, 0.999, 500)

    # Posteriors from different priors in shades of blue
    for i, (a, b, label) in enumerate(priors):
        posterior = stats.beta.pdf(p_grid, x + a, n - x + b)
        ax.plot(p_grid, posterior, color=MULTI_BLUES[i], linewidth=2,
                label=label)

    # BvM Normal approximation in red (dashed)
    if n >= 20 and 0 < mle < 1:
        fisher_var = mle * (1 - mle) / n
        normal_approx = stats.norm.pdf(p_grid, mle, np.sqrt(fisher_var))
        ax.plot(p_grid, normal_approx, color=COLOR_APPROX, linewidth=2,
                linestyle='--', label='BvM approx', alpha=0.8)

    # MLE vertical line (black, dotted)
    ax.axvline(mle, color=COLOR_LIKELIHOOD, linestyle=':', linewidth=1.5,
               alpha=0.7, label=f'MLE = {mle:.2f}')

    # True value (black, dashed)
    ax.axvline(true_p, color=COLOR_TRUE, linestyle='--', linewidth=1.5, alpha=0.
    ↪4,
               label=f'True p = {true_p}')
    ax.set_title(f'n = {n}, X = {x}', fontsize=12, fontweight='bold')
    ax.set_xlabel('$p$', fontsize=11)
    ax.set_xlim(0, 1)
    ax.set_ylim(0, None)
    if n == 5:
        ax.legend(fontsize=7, loc='upper left')

plt.suptitle('Posteriors from Different Priors Converge as n Grows',
             fontsize=14, fontweight='bold', y=1.02)
plt.tight_layout()
plt.show()

```

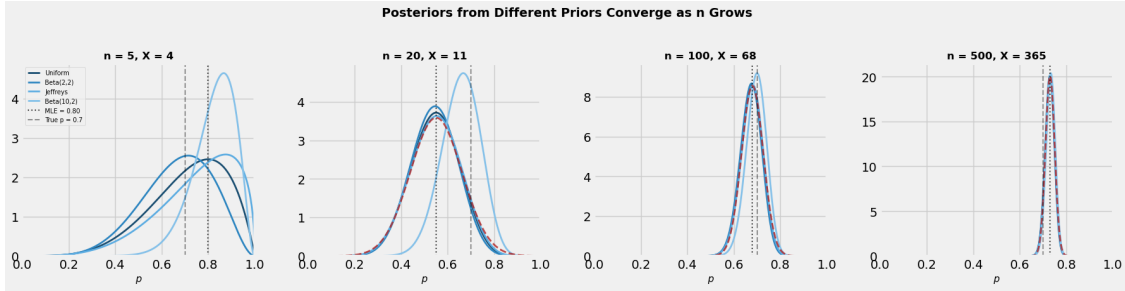


Figure 4. Posterior distributions from four different priors (shades of blue), shown for increasing sample sizes. The true parameter is $p = 0.7$ (black dashed) and the MLE $\hat{p} = X/n$ (black dotted). For small n (left), the posteriors differ substantially — the choice of prior matters. As n grows (right), all posteriors converge to the same shape, concentrated around the MLE. For $n \geq 20$, the red dashed curve shows the Bernstein–von Mises Normal approximation $N(\hat{p}, \hat{p}(1 - \hat{p})/n)$, which matches the posteriors closely.

1.7.3 Observations

- **Small n** (left): The posteriors differ substantially depending on the prior. The choice of prior matters!
- **Moderate n** (second panel): The posteriors are starting to converge. Differences are visible but shrinking.
- **Large n** (right panels): All posteriors are nearly identical — concentrated around the MLE, and well-approximated by the Bernstein–von Mises Normal distribution $N(\hat{p}, \hat{p}(1 - \hat{p})/n)$ (red dashed line).

This is the **Bernstein–von Mises phenomenon**: regardless of the prior, the posterior converges to the same Normal distribution centered at the MLE.

For large samples, we don’t need to get the prior exactly right — any reasonable prior leads to essentially the same inference. **For small samples**, the prior matters, and this is a *feature*: when data are scarce, it makes sense for prior knowledge to influence our conclusions.

1.8 7. Credible Intervals

The posterior distribution gives us more than a point estimate — it provides a complete description of our uncertainty about θ .

A $100(1 - \alpha)\%$ **credible interval** is any interval $[a, b]$ such that $P(\theta \in [a, b] \mid X) = 1 - \alpha$. The most common choice is the **equal-tailed credible interval**: the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior.

For the Beta-Binomial model with $p \mid X \sim \text{Beta}(X + \alpha, n - X + \beta)$, a 95% credible interval is the 2.5th and 97.5th percentiles of this Beta distribution.

By the Bernstein–von Mises theorem, for large n the posterior is approximately $N(\hat{p}, \hat{p}(1 - \hat{p})/n)$, so the 95% credible interval is approximately $\hat{p} \pm 1.96\sqrt{\hat{p}(1 - \hat{p})/n}$ — the same formula as the asymptotic normal 95% confidence interval. Let's see how close these two intervals are.

```
[6]: # Posterior with credible interval - larger sample
n, x = 80, 60
alpha_prior, beta_prior = 1, 1
post_a, post_b = x + alpha_prior, n - x + beta_prior
post_dist = stats.beta(post_a, post_b)

p_grid = np.linspace(0.001, 0.999, 500)
posterior = post_dist.pdf(p_grid)

# Exact Bayesian credible interval (equal-tailed)
ci_lo, ci_hi = post_dist.ppf(0.025), post_dist.ppf(0.975)
post_mean = post_a / (post_a + post_b)

# Asymptotic normal confidence interval
mle = x / n
se = np.sqrt(mle * (1 - mle) / n)
wald_lo, wald_hi = mle - 1.96 * se, mle + 1.96 * se

# BuM Normal approximation to the posterior
bvm_approx = stats.norm.pdf(p_grid, mle, se)

fig, ax = plt.subplots(figsize=(10, 6))

# Posterior (blue) with shaded credible interval
ax.plot(p_grid, posterior, color=COLOR_POSTERIOR, linewidth=2.5,
        label=f'Posterior: Beta({post_a}, {post_b})')
mask = (p_grid >= ci_lo) & (p_grid <= ci_hi)
ax.fill_between(p_grid[mask], posterior[mask], alpha=0.2, color=COLOR_POSTERIOR,
               label=f'95% credible interval: [{ci_lo:.3f}, {ci_hi:.3f}]')

# BuM Normal approximation (red dashed)
ax.plot(p_grid, bvm_approx, color=COLOR_APPROX, linewidth=2, linestyle='--',
        label=f'BvM approx: N({mle:.3f}, {se**2:.4f})')

# Mark intervals
ax.axvline(post_mean, color=COLOR_POSTERIOR, linestyle=':', linewidth=1.5,
           label=f'Posterior mean: {post_mean:.3f}')
ax.axvline(mle, color=COLOR_LIKELIHOOD, linestyle=':', linewidth=1.5,
           label=f'MLE: {mle:.3f}')

# Show asymptotic normal confidence interval as a horizontal bracket
bracket_y = ax.get_ylim()[1] * 0.05 if ax.get_ylim()[1] > 0 else 0.5
ax.plot([wald_lo, wald_hi], [bracket_y, bracket_y], color=COLOR_APPROX,
```

```

        linewidth=3, solid_capstyle='butt',
        label=f'Asymp. normal 95% conf. int.: [{wald_lo:.3f}, {wald_hi:.3f}]')
ax.plot([wald_lo, wald_lo], [bracket_y - 0.2, bracket_y + 0.2],
        color=COLOR_APPROX, linewidth=2)
ax.plot([wald_hi, wald_hi], [bracket_y - 0.2, bracket_y + 0.2],
        color=COLOR_APPROX, linewidth=2)

ax.set_xlabel('$p$', fontsize=12)
ax.set_ylabel('Posterior density', fontsize=12)
ax.set_title(f'Posterior Distribution with 95% Credible Interval\n'
            f'(n={n}, X={x}, uniform prior)', fontsize=14, fontweight='bold')
ax.legend(fontsize=10)
ax.set_xlim(0.55, 0.9)
ax.set_ylim(0, None)

plt.tight_layout()
plt.show()

```

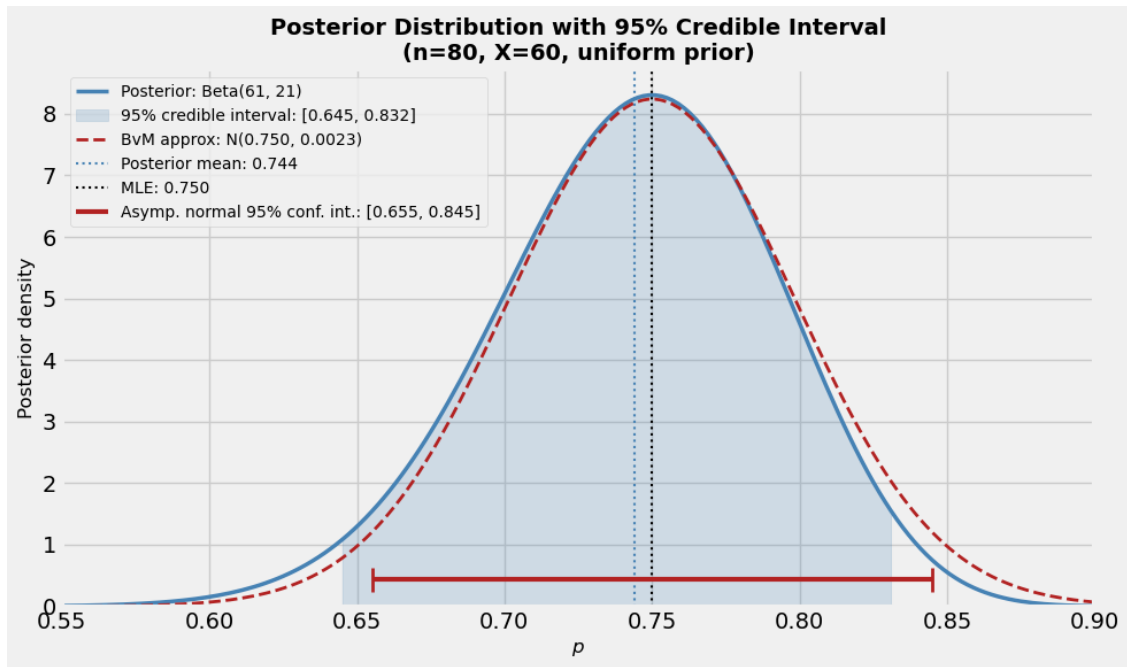


Figure 5. Posterior distribution $p \mid X \sim \text{Beta}(61, 21)$ (blue solid) after observing $X = 60$ successes in $n = 80$ trials with a uniform prior, alongside the Bernstein-von Mises Normal approximation (red dashed). The shaded blue region is the exact 95% Bayesian credible interval; the red bracket near the x-axis is the asymptotic normal 95% confidence interval $\hat{p} \pm 1.96\sqrt{\hat{p}(1 - \hat{p})/n}$. With $n = 80$, the two intervals are nearly identical.

Interpretation: Given the data ($X = 60$ successes in $n = 80$ trials) and a uniform prior, there is a 95% posterior probability that p lies in the shaded interval. This is a direct probability statement

about p itself.

Notice how close the Bayesian credible interval and the asymptotic normal confidence interval are. This is the Bernstein–von Mises theorem in action: for large n , the two approaches give essentially the same answer. The BvM Normal approximation (red dashed) is nearly indistinguishable from the exact posterior (blue solid).

1.9 8. Summary

1.9.1 Key Concepts

Concept	Definition
Prior $\pi(\theta)$	Distribution representing beliefs about θ before seeing data
Likelihood $f_\theta(x)$	Probability of data x given parameter θ
Posterior $\pi(\theta x)$	Updated beliefs: $\propto f_\theta(x) \cdot \pi(\theta)$
Bayes estimator	Posterior mean $E[\theta X]$ (for squared error loss)
Conjugate prior	Posterior stays in the same family as the prior
Sufficient statistic	A statistic that determines the likelihood (up to a constant)
Credible interval	Interval with specified posterior probability

1.9.2 What We Learned

1. **The Bayesian framework** models θ as random with a prior, and updates to a posterior via Bayes' rule: Posterior \propto Likelihood \times Prior.
2. **The Bayes estimator** for squared error loss is the posterior mean. We can optimize for every X value simultaneously, even though we can't optimize for every θ . Other loss functions lead to other summaries of the posterior (homework).
3. **Conjugate priors** yield closed-form posteriors. The posterior mean is always a weighted average of the MLE and the prior mean, with weights determined by relative "sample sizes." We saw three examples: Beta-Binomial, Gamma-Exponential, and Normal-Normal.
4. **The likelihood carries all information** from the data. A statistic that determines the likelihood is sufficient — it contains everything the data have to say about θ .
5. **For large samples**, the posterior is approximately $N(\hat{\theta}_{\text{MLE}}, 1/(nI(\hat{\theta})))$ regardless of the prior — Bayesian and frequentist inference converge. The 95% credible interval \approx asymptotic normal 95% confidence interval.

1.9.3 Next Time (Lecture 8)

- How should we choose a prior? What does it mean for θ to "have a probability"?
- Objective Bayes: Jeffreys prior and other non-informative priors
- Hierarchical Bayes: when even the prior has unknown parameters
- Computational methods for Bayesian inference