Decision and Risk Analysis

Lecture 9: **Bayesian Simple Normal Regression**

DAEN 427 / ISEN 427 / ISEN 627 - Fall 2026

Alexander Abuabara



Learning goals

1. Identify when Normal regression is appropriate.
2. Interpret coefficients, residual variation, and priors.
3. Use posterior simulations for inference and prediction.
4. Apply predictive uncertainty to risk-based decisions.
5. Assess model limitations before acting.

The business question: tomorrow bike demand

Capital Bikeshare wants enough bikes available to serve customers, without paying for excessive spare capacity.

- Outcome Y: number of rides on a day.
- Predictor X: temperature (the data use `temp_feel`).
- Modeling question: how does typical ridership vary with temperature?
- Decision question: how much capacity should we prepare given a forecast of 75°F?

R: fit the model

```
library(tidyverse)
library(bayesrules)
library(rstanarm)
data(bikes)

glimpse(bikes)

summary(bikes[, c("rides", "temp_feel")])

nrow(bikes)

summary(lm(rides ~ temp_feel,
           data = bikes))
```

Crucial/pivotal: estimating the mean is not the same as forecasting one unpredictable day.

Model choice comes before computation

1. What kind of response is Y?
   Rides are counts, but daily values can be large.

2. Is a Poisson model reasonable?
   A Poisson imposes $\text{Var}(Y) = \mathbb{E}(Y)$;
   historical bike demand is/can be much more variable.

   - Bike demand may fluctuate by hundreds of rides because of weather, weekends, holidays, and other factors.
   - If the observed variance exceeds the mean, we call this overdispersion.

3. A Normal model may be a useful approximation to the broad, roughly bell-shaped variation.

   - Temperature explains some, but not all, of that variation.
   - Start with Normal regression, then check:
     - Whether the residuals are approximately Normal.
     - Whether variability changes with temperature.
     - Whether predictions are realistic, especially for unusually high demand.

From one overall mean to a temperature-specific mean

Modeling decision: start with Normal regression; later test fit and check whether predictions behave sensibly (especially in tails).

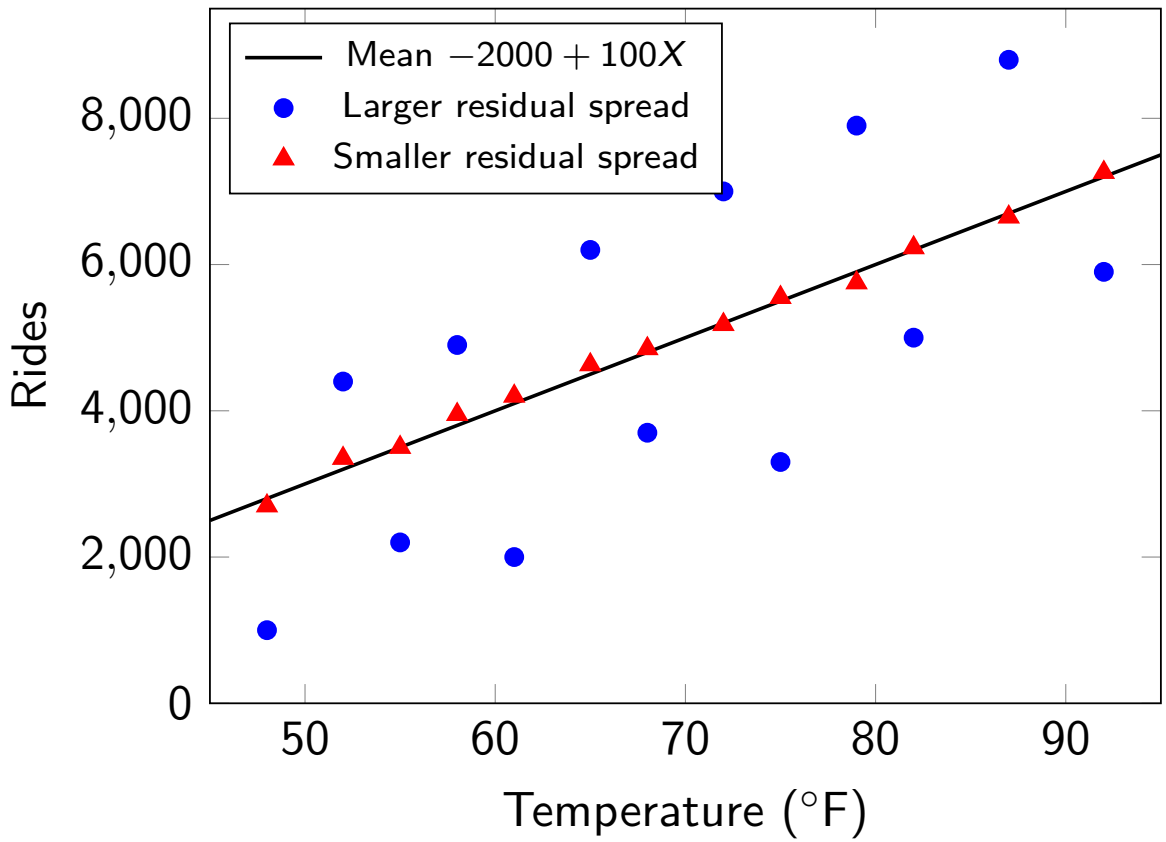
$$Y_i \mid \mu, \sigma \sim N(\mu, \sigma^2) \qquad \Longrightarrow \qquad Y_i \mid \beta_0, \beta_1, \sigma \sim N(\mu_i, \sigma^2)$$

$$\boxed{\mu_i = \beta_0 + \beta_1 X_i}$$

- β_0: fitted mean at $X = 0$ (possibly outside a meaningful range).
- β_1: expected additional rides per one-degree increase.
- σ: day-to-day standard deviation *at a fixed temperature*.

Example: $\mu = -2000 + 100X$ predicts 5000 rides at 70°F. The negative intercept is an extrapolation, not negative actual demand.

Same trend, different risk



Synthetic points: both have the same slope; larger σ makes tomorrow demand harder to predict.

What assumptions are we choosing?

1. **Conditional independence:** after accounting for X, observations do not depend on each other.
2. **Linear mean:** $\mathbb{E}(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$.
3. **Normal errors:** departures from the line are approximately Normal.
4. **Constant spread:** the same σ applies at every X.

Could weekends, holidays, seasonality, or weather patterns break these assumptions? Why might a forecast be risky on an unusually hot day?

Bayesian regression = likelihood + priors

$$
\begin{aligned}
Y_i \mid \beta_0, \beta_1, \sigma &\overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2), \\
\beta_0 &\sim N(m_0, s_0^2), \\
\beta_1 &\sim N(m_1, s_1^2), \\
\sigma &\sim \text{Exp}(\lambda).
\end{aligned}
$$

These priors describe plausible parameter values *before* observing the modeled dataset. An Exponential prior keeps $\sigma > 0$.

Choice: tune priors to real knowledge; avoid priors implying absurd ridership changes or implausible fluctuations.

- Bayesian regression uses the likelihood to describe the data, priors to express initial uncertainty, and the posterior to update our knowledge about regression parameters.
- Important: Priors should be checked against real-world knowledge, ideally using prior predictive simulations before fitting the model.

Centering gives the intercept practical meaning

Define $X_c = X - 70$; then

$$\mu = \beta_{0c} + \beta_1(X - 70), \qquad \beta_{0c} = \beta_0 + 70\beta_1.$$

- β_{0c} means typical rides at 70°F.
- β_0 means a hypothetical average at 0°F.
- Both parameterizations imply *the same line* when transformed consistently.

Prior intuition: 5,000 rides at a typical temperature is easier to assess than $-2,000$ at freezing temperatures.

Informed priors

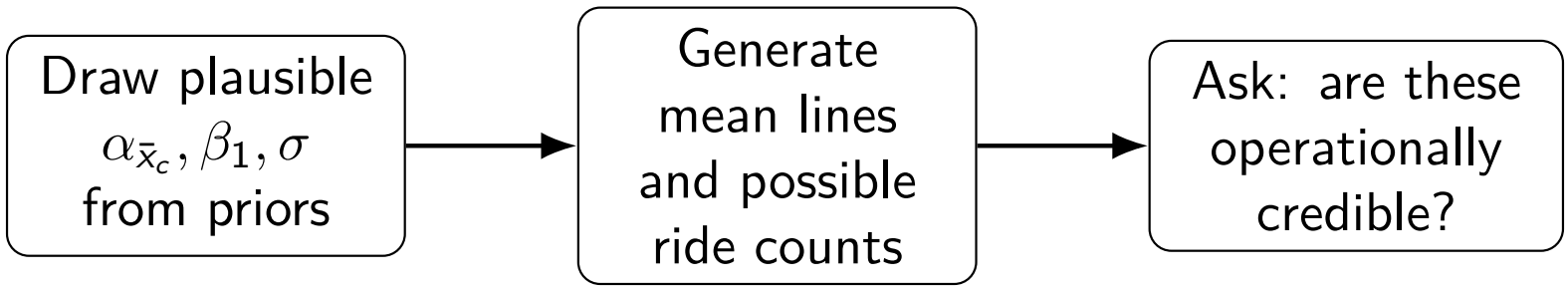
$$\alpha_{\bar{x}_c} \sim N(5000, 1000^2), \quad \beta_1 \sim N(100, 40^2), \quad \sigma \sim \text{Exp}(0.0008).$$

$$\alpha_{\bar{x}_c} = \beta_{0c} + \beta_1 \bar{X}_c, \qquad \bar{X}_c = \overline{(X - 70)}.$$

- Baseline at the sample mean temperature: about 3,000–7,000 rides (roughly 95% prior range).
- Slope: around 100 additional rides per 1°F, with appreciable uncertainty.
- Residual SD: prior mean $1/0.0008 = 1250$ rides.

Prior predictive check: simulate lines *and whole datasets*. Do these imply believable high-demand days, low-demand days, and uncertainty?

Prior predictive checks prevent poor decisions



A prior can look reasonable for each parameter separately yet produce unbelievable combinations.

Action: revise the prior or the model *before* interpreting posterior probabilities as business risks.

Updating with 500 observed days

The example uses a sample of 500 days with `rides` and `temp_feel`.

$$p(\beta_{0c}, \beta_1, \sigma \mid \mathcal{D}) \propto \underbrace{p(\beta_{0c}, \beta_1, \sigma)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} p(y_i \mid x_{ci}, \beta_{0c}, \beta_1, \sigma)}_{\text{likelihood}}.$$

The posterior is not a simple familiar distribution.
MCMC approximates it using many plausible parameter triples.
The draw $(3610, 83, 1300)$ describes one possible line and one possible residual spread; it is *not* a separate fitted model selected in isolation.

R: fit the model using `rstanarm`

```
library(bayesrules)
library(rstanarm)
data(bikes)
bikes <- mutate(bikes, temp_feel_c = temp_feel - 70)

bike_model <- stan_glm(
  rides ~ temp_feel_c,
  data = bikes,
  family = gaussian(),
  prior_intercept = normal(5000, 1000, autoscale = FALSE),
  prior = normal(100, 40, autoscale = FALSE),
  prior_aux = exponential(0.0008, autoscale = FALSE),
  chains = 4, iter = 10000, seed = 84735
)
```

In `rstanarm`, `prior_intercept` targets the mean at the sample-average predictor, not exactly 70°F. The reported `(Intercept)` is still the mean at 70°F.

R: visualize many plausible regression lines from Stan

One posterior draw = one plausible line. Plot 40 draws over the data:

```
# Note: run after fitting bike_model with stan_glm()

draws <- as.matrix(bike_model)
idx   <- sample(seq_len(nrow(draws)), 40)

plot(bikes$temp_feel, bikes$rides,
     pch = 16, col = "gray75",
     xlab = "Temperature (F)", ylab = "Daily rides")

for (s in idx) {
  abline(a = draws[s, "(Intercept)"] -
           70 * draws[s, "temp_feel_c"],
         b = draws[s, "temp_feel_c"],
         col = adjustcolor("maroon", alpha.f = 0.25))}
```

Interpretation: Spread among lines represents uncertainty about *mean demand*; it does *not* show day-to-day residual noise.

Check the simulation before trusting it

- **Trace plots:** chains should mix and explore similar parameter regions.
- $\widehat{R}$ **near 1:** evidence of stable between-chain behavior.
- **Effective sample size:** enough independent information from simulated draws.
- **Also inspect warnings:** divergences and other sampler problems matter.

Check the actual $\widehat{R}$ values from your refit; do not assume convergence.

Risk: bad computation can create unwarranted confidence, even when the regression equation looks correct.

What the posterior says

Refit the centered model above and obtain its *actual* posterior summaries:

```
draws <- as.matrix(bike_model)
params <- c("(Intercept)", "temp_feel_c", "sigma")
round(rbind(
  estimate = colMeans(draws[, params]),
  lower80 = apply(draws[, params], 2, quantile, 0.10),
  upper80 = apply(draws[, params], 2, quantile, 0.90)
), 1)
```

$$\widehat{\mu}(x_c) = \widehat{\beta}_{0c} + \widehat{\beta}_1 x_c, \quad x_c = x - 70.$$

- The intercept estimates expected rides at 70°F.
- The slope estimates change in expected rides per degree.
- Credible intervals describe parameters, not day-to-day rides; association is not causation.

Probability is more actionable than a yes/no claim

With posterior draws $\beta_1^{(1)}, \ldots, \beta_1^{(S)}$,

$$\mathbb{P}(\beta_1 > 0 \mid \mathcal{D}) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\{\beta_1^{(s)} > 0\}.$$

Compute this probability from the refitted model: `mean(draws[, "temp_feel_c"] > 0)`. The answer depends on its posterior draws.

But: a high posterior probability of positive association does *not* imply predictable daily demand. Residual variation still matters.

Two questions that must not be confused

At 75°F, the fitted mean is

$$\widehat{\mu}(75) = \widehat{\beta}_{0c} + 5\widehat{\beta}_1.$$

Question A: average ridership at 75°F

How uncertain is the *mean* $\mu(75)$?
Only parameter uncertainty matters.

Question B: ridership tomorrow at 75°F

How uncertain is one *future observation* Y_{new}?
Parameter uncertainty *plus* residual day-to-day variability matter.

Same forecast, but two very different questions

Suppose the temperature tomorrow is 75°F.

To make the distinction concrete,
imagine our analysis suggests about **5,000 rides** on average.

Question A: If we repeated many 75°F days,
what would their average be?

Question B: How many rides will we get on ONE 75°F day?

Question A: If we repeated many 75°F days, what would their average be?

We want the **long-run average** number of rides at this temperature.

- Example: perhaps that unknown average is between **4,900 and 5,100** rides.
- Uncertainty: we do not know the true regression line exactly.

Question B: How many rides will we get on ONE 75°F day?

Even if the true average were exactly 5,000, some days might have **4,000** rides and others **6,000**.

- Uncertainty: the unknown average *and* ordinary day-to-day fluctuations.

Knowing the average does not tell us exactly what happens tomorrow.

PS: Numbers are examples, not output from the fitted model.

Two R functions

What does each simulated number mean?
For a forecast of 75°F, each posterior draw s supplies a possible mean $\mu^{(s)}$ and residual standard deviation $\sigma^{(s)}$.

`posterior_epred()`: simulate uncertainty about the AVERAGE

Returns $\mu^{(s)} = \beta_{0c}^{(s)} + 5\beta_1^{(s)}$.
Interpret a value of 5,050: "The average for 75°F days might be 5,050."

`posterior_predict()`: simulate an ACTUAL future day

Returns $Y_{\text{new}}^{(s)} = \mu^{(s)} + \varepsilon^{(s)}$, with $\varepsilon^{(s)} \sim N(0, (\sigma^{(s)})^2)$.
Interpret a value of 5,900: "Tomorrow might have 5,900 rides."

```
new_day <- data.frame(temp_feel_c = 5) # 75 - 70
avg_draws <- posterior_epred(bike_model, newdata = new_day)
day_draws <- posterior_predict(bike_model, newdata = new_day)
```

Operational question: $P(\text{tomorrow's rides} > 5000)$? Use `mean(day_draws > 5000)`, *not* `avg_draws`.

Posterior prediction: one extra simulation step

For each saved posterior draw s:

1. Compute $\mu_{75}^{(s)} = \beta_{0c}^{(s)} + (75 - 70)\beta_1^{(s)}$.
2. Generate $Y_{\text{new}}^{(s)} \sim N(\mu_{75}^{(s)}, (\sigma^{(s)})^2)$.
3. Summarize the collection of future $Y_{\text{new}}^{(s)}$ values.

$$p(y_{\text{new}} \mid x_c = 5, \mathcal{D}) = \int p(y_{\text{new}} \mid x_c = 5, \theta)\, p(\theta \mid \mathcal{D})\, d\theta.$$

Key idea: predictions integrate over parameter uncertainty; they do not just plug in one estimated line.

To calculate the actual probability from our fitted Bayesian model

```
y_pred <- posterior_predict(
  bike_model,
  newdata = data.frame(temp_feel_c = 5)
)

mean(y_pred > 5000)
```

Average uncertainty vs tomorrow uncertainty

```
draws <- as.matrix(bike_model) # Obtain posterior parameter draws

# Expected ridership at 75 F
mu_75 <- draws[, "(Intercept)"] + (75 - 70) * draws[, "temp_feel_c"]

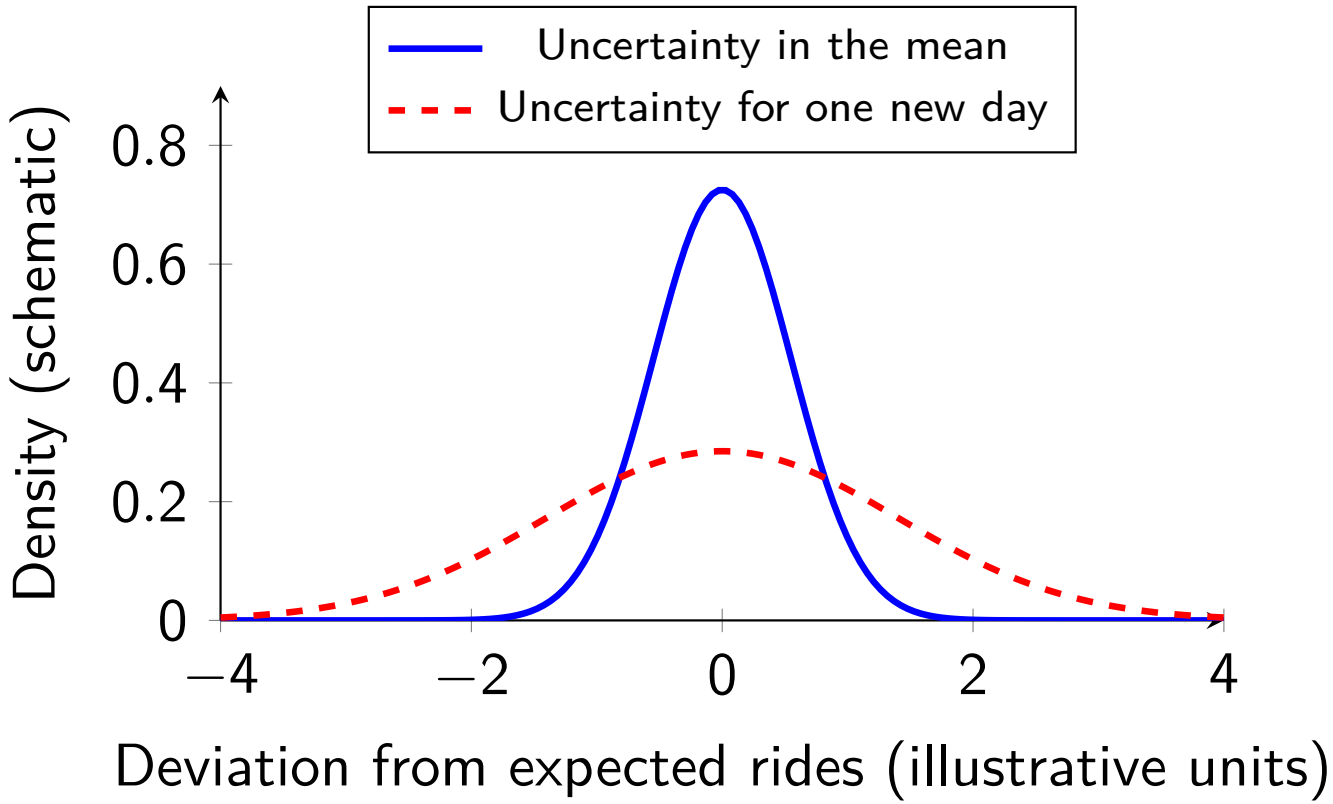
# 95% credible interval for mean ridership
quantile(mu_75, c(0.025, 0.975))

# Simulate actual ridership on a new day
set.seed(84735)
y_new <- rnorm(n = length(mu_75),
               mean = mu_75, sd = draws[, "sigma"])

# 95% posterior predictive interval
quantile(y_new, c(0.025, 0.975))

# Mean estimate and both intervals from the refit
mean(mu_75)
quantile(mu_75, c(0.025, 0.975))
quantile(y_new, c(0.025, 0.975))
```

Mean vs predictive uncertainty



Conceptual diagram, not fitted data: future-day outcomes have extra residual variation. Use R-generated quantiles from `posterior_epred()` and `posterior_predict()` for numerical intervals.

R walkthrough: centering and verifying units

```
bikes <- mutate(bikes, temp_feel_c = temp_feel - 70)
head(bikes[, c("temp_feel", "temp_feel_c")])
new_day <- data.frame(temp_feel_c = 75 - 70)
new_day
```

Exact R output for the last command:

```
  temp_feel_c
1           5
```

- `mutate()` creates a new variable without discarding the data.
- `data.frame()` builds the one-row forecast input.
- The model uses 5 degrees *above 70°F*, not 75 as the input.

R walkthrough: posterior parameter output

```
draws <- as.matrix(bike_model)
params <- c("(Intercept)", "temp_feel_c", "sigma")
head(draws[, params])  # first 6 posterior draws
colMeans(draws[, params])  # posterior means
apply(draws[, params], 2, quantile, c(.10, .90))
```

- `as.matrix()` returns a row for each saved posterior draw.
- `(Intercept)`: fitted mean at 70°F; `sigma`: residual SD.
- `colMeans()` averages parameter draws.
- `quantile()` and `apply()` find 80% credible intervals.

Output format: columns are intercept, slope, and sigma; values must come from your own run, not an old fit.

R walkthrough: expected mean versus new rides

```
new_day <- data.frame(temp_feel_c = 5)

mu_draws <- as.numeric(posterior_epred(bike_model, newdata = new_day))
y_pred <- as.numeric(posterior_predict(bike_model, newdata = new_day))

quantile(mu_draws, c(.025, .975))
quantile(y_pred, c(.025, .975))
mean(y_pred > 5000)
```

- `posterior_epred()`: uncertainty in the *expected mean*.
- `posterior_predict()`: also adds one-day residual noise.
- `quantile()`: endpoints of a 95% interval.
- `mean(y_pred > 5000)`: fraction of draws above capacity.

R output: an exact decision threshold

```
shortfall_cost <- 4000
extra_capacity_cost <- 600
threshold <- extra_capacity_cost / shortfall_cost
threshold
```

Exact output:

```
[1] 0.15
```

- Compare your fitted `mean(y_pred > 5000)` to 0.15.
- The output 0.15 follows from the stated hypothetical costs; it does *not* depend on the regression fit.

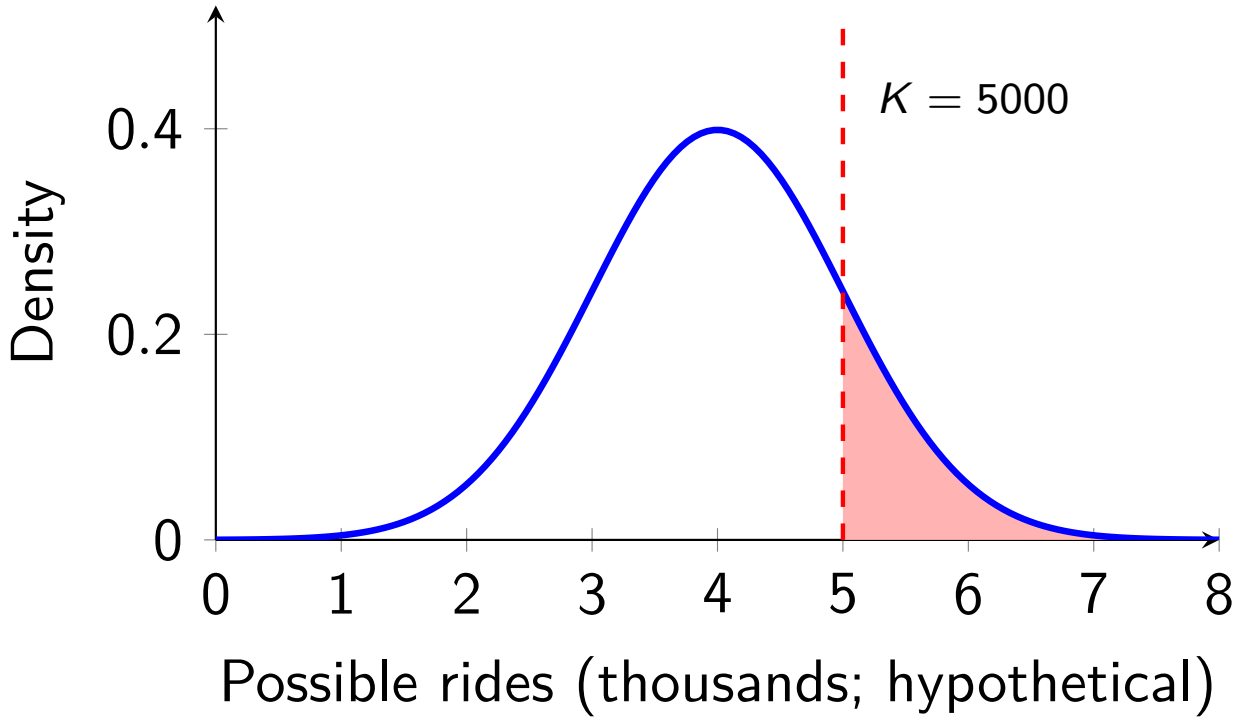
R plots: observed data and posterior predictive risk

```
plot(bikes$temp_feel, bikes$rides,
     xlab = "Temperature (F)", ylab = "Rides")
abline(lm(rides ~ temp_feel, data = bikes))

hist(y_pred, breaks = 40,
     main = "Future-day rides at 75 F", xlab = "Rides")
abline(v = 5000, col = "red", lty = 2)
```

- Scatterplot: observed temperature and rides; line summarizes association.
- Histogram: each bar counts simulated future-day outcomes.
- Red line: capacity threshold; area to its right corresponds to risk.
- Running `bikes_corrected.R` saves both graphs as PNG files.

Visualizing predictive capacity risk



Shaded right tail means $Y_{\text{new}} > K$. Diagram is hypothetical; `hist(y_pred)` visualizes the actual fitted predictive draws.

From forecasts to operational risk

A manager needs event probabilities, not just a point forecast:

$$\mathbb{P}(Y_{\text{new}} > K \mid X_c = 5, \mathcal{D})$$

where K is the capacity that can be served.

- $K = 5000$: probability that demand exceeds 5,000 rides.
- $K = 3000$: probability that demand exceeds 3,000 rides.
  The complement is the probability demand is at most 3,000.

With posterior predictive draws:

$$\widehat{\mathbb{P}}(Y_{\text{new}} > K \mid \mathcal{D}) = \frac{\#\{Y_{\text{new}}^{(s)} > K\}}{S}.$$

These probabilities directly support inventory and staffing decisions.

Example: Normal-tail calculation (hypothetical)

Suppose a planner assumes

$$Y_{\text{new}} \sim N(4000, 1000^2).$$

For capacity $K = 5000$,

$$z = \frac{5000 - 4000}{1000} = 1, \quad \mathbb{P}(Y_{\text{new}} > 5000) = 1 - \Phi(1) \approx 0.1587.$$

Interpretation: under this hypothetical distribution, about 15.9% of days exceed 5,000 rides.

For the fitted bike model: use the observed fraction of `posterior_predict()` draws exceeding 5,000 instead.

A capacity decision needs a loss model

Scenario (illustrative costs, not from the book):

- **Standard capacity** (5,000 rides): no extra upfront cost, but a high-demand shortfall event creates a simplified \$4,000 loss.
- **Expanded capacity**: costs an extra \$600 and is assumed to eliminate that shortfall loss for this example.

Define $H = \{Y_{\text{new}} > 5000\}$. The loss matrix is

	H	H^c
Standard	4000	0
Expanded	600	600

A real loss function would account for the *size* of each shortfall, capacity limits, and other costs.

Choose the action with lower expected loss

Let $p = \mathbb{P}(Y_{\text{new}} > 5000 \mid X_c = 5, \mathcal{D})$, estimated from posterior draws.

$$\mathbb{E}[L(\text{Standard}) \mid \mathcal{D}] = 4000p,$$
$$\mathbb{E}[L(\text{Expanded}) \mid \mathcal{D}] = 600.$$

Decision rule: expand if

$$4000p > 600 \quad \Longleftrightarrow \quad p > 0.15.$$

- If $p < 0.15$, choose standard capacity.
- If $p = 0.15$, both options have equal expected loss.
- If $p > 0.15$, choose expanded capacity.

Find your actual p by running the R code;
no refit output is assumed.

Sensitivity analysis: what would reverse the choice?

Assumed high-demand risk	Expected standard loss	Preferred action
10%	\$400	Standard
15%	\$600	Tie
21%	\$840	Expanded
30%	\$1,200	Expanded

- Changing the event cost also changes the threshold.
- Forecast uncertainty, data bias, and wrong model assumptions may change $\mathbb{P}(H)$.
- A robust decision considers plausible alternatives, not just one model run.

Decision risk is broader than residual noise

1. **Parameter risk:** uncertain intercept, slope, and residual SD.
2. **Outcome risk:** random variation even if parameters were known.
3. **Model risk:** nonlinearity, heteroskedasticity, seasonality, autocorrelation, count boundaries.
4. **Decision risk:** wrong costs, missing operational constraints, or forecast error in X.

Sequential learning can change the decision

The book re-fits after 30, 60, and 500 observed days.

- Early winter days cover a narrow temperature range; slope estimates may be uncertain or misleading.
- As data span more temperatures, the inferred relationship changes and tightens.
- Revisit forecast risks and capacity decisions as new evidence arrives.

Warning: having more observations is not always enough; *coverage of decision-relevant conditions* matters.

When prior information is weak

- Weak priors can allow positive, flat, or negative slopes rather than asserting a strong prior increase.
- Autoscaling adapts prior scale to data units; check the resulting prior with `prior_summary()`.
- Simulate prior predictive datasets and assess realistic magnitudes.

Decision impact: a weak prior may produce wider predictive tails, especially with limited data. Compare choices across reasonable prior specifications.

R: predictive risk and decision rule

```
new_day <- data.frame(temp_feel_c = 75 - 70)
y_sim <- posterior_predict(bike_model,
                           newdata = new_day)
y_sim <- as.numeric(y_sim)

p_short <- mean(y_sim > 5000)
loss_standard <- 4000 * p_short
loss_expanded <- 600

c(p_short = p_short,
  loss_standard = loss_standard,
  loss_expanded = loss_expanded)
```

Use the simulation output as the primary risk estimate. Use the model output; the Normal-tail classroom example is hypothetical.

Homework (part 1)

A manager can buy expanded capacity for \$600. The cost of failing to meet demand above 5,000 rides is \$4,000 (simplified event loss).

1. If $\mathbb{P}(Y_{\text{new}} > 5000 \mid \mathcal{D}) = 0.12$, what action minimizes expected loss?
2. If the shortage cost rises to \$7,500, how does the threshold change?
3. What if your Normal predictive model assigns negative rides some probability?
4. Which additional predictor or diagnostic would most improve this decision?

Answers: (1) Standard: \$480 $<$ \$600. (2) Threshold $600/7500 = 0.08$, so expand at risk 0.12. (3) Check model adequacy and consider alternatives. (4) Reasonable answers include day-of-week and seasonality.

Homework (part 2)

9.10.1 Conceptual exercises

- Exercise 9.1
- Exercise 9.2
- Exercise 9.3
- Exercise 9.4
- Exercise 9.5

9.10.2 Applied exercises

- Exercise 9.9
- Exercise 9.10

Model → Uncertainty → Decision

1. Choose a data model and mean structure that reflect the substantive problem.
2. Use defensible priors, posterior simulation, and diagnostics.
3. For actions involving one future day, use the *posterior predictive*, not a confidence interval for the mean.
4. Translate predictive draws into event probabilities and expected losses.
5. Stress-test the choice against prior sensitivity, model misspecification, and uncertain costs!