# Recap: Probabilistic Modeling

DS701 Session 6 — Wed Sep 23, 2026

# Today’s plan

# Knowledge-Check Review

# KC 1: The recipe, twice

*The lecture fits a probability model to data in the same four steps twice — for Boston July temperatures and for the Prussian horse-kick counts. List the four steps, and for each dataset name the distribution family chosen and give one phrase saying why that family fits the shape of the data.*

# KC 2: What “maximum likelihood” means

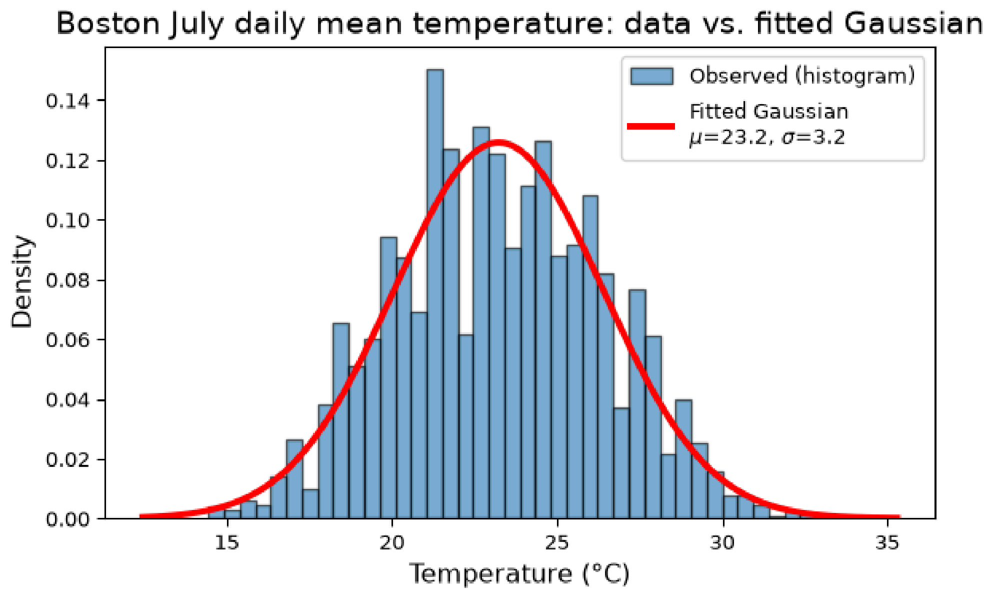
*In one or two sentences, what does it mean for* $\hat{\mu}$ *and* $\hat{\sigma}^2$ *to be the maximum likelihood estimates of a Gaussian’s parameters, and what do those estimates turn out to be? Why does maximizing the log-likelihood give the same answer as maximizing the likelihood?*

# KC 3: Reading a confidence interval

*From 2,821 July days the lecture reports a 95% confidence interval for the **mean** July temperature of $[23.12, 23.35]$ °C, with $\hat{\mu} = 23.2$ °C and $\hat{\sigma} = 3.2$ °C. A friend reads this as "so 95% of July days in Boston are between 23.12 and 23.35 °C." Explain what is wrong with that reading, and use the two fitted numbers to say roughly what range does contain about 95% of individual July days.*

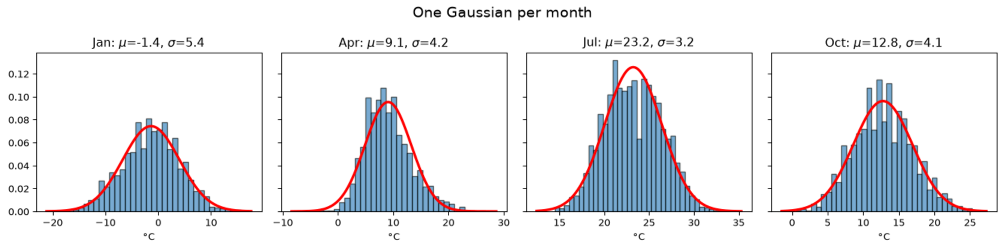
# Highlights

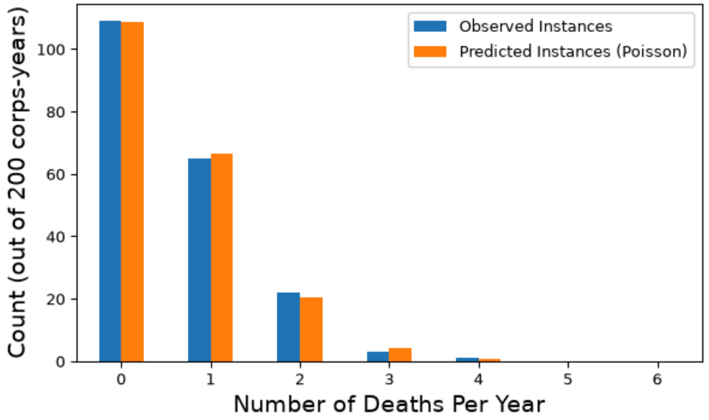
# Idea 1 — fit, then check



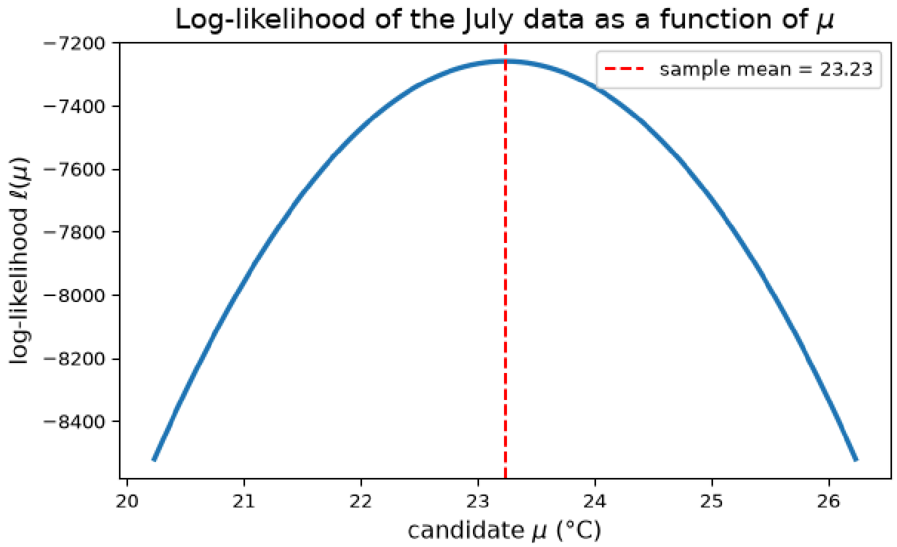

Fitting is cheap. **Checking is what makes it a model rather than an assumption.**

# Idea 1, continued — when the check fails, believe the check



# Idea 2 — MLE: maximize the likelihood of what you saw

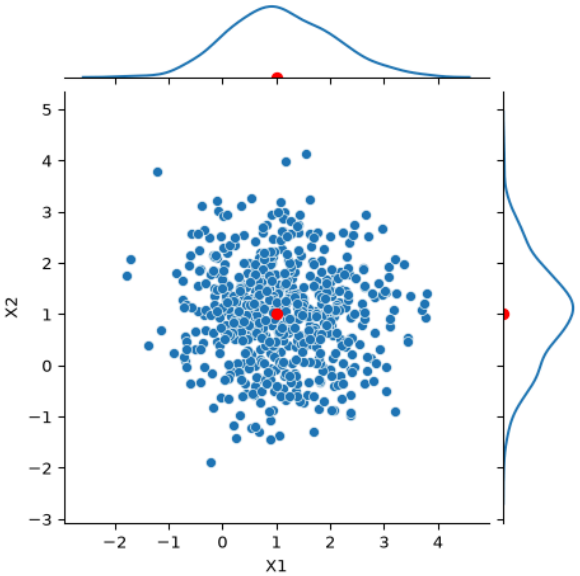


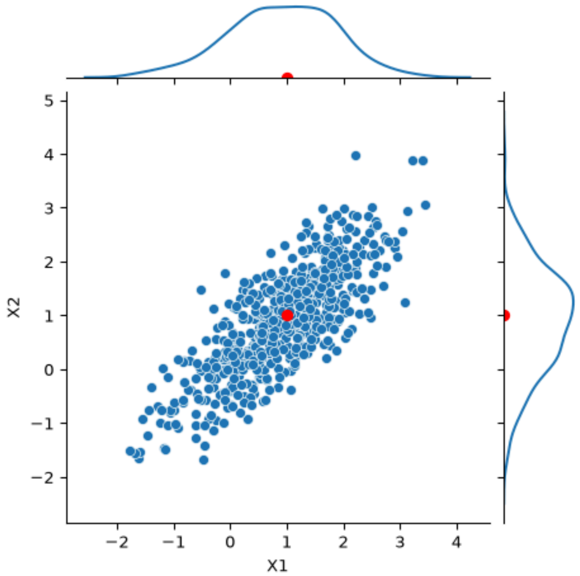

The Gaussian log-likelihood of the July data as $\mu$ varies ($\sigma$ fixed) — it peaks **exactly** at the sample mean.

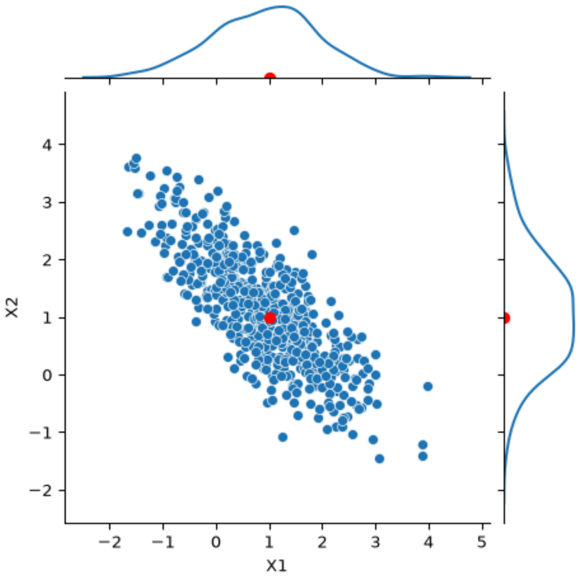
The recipe, for *any* model:

# Idea 2, continued — the same recipe, and where you will meet it again

# Idea 3 — covariance shapes the multivariate Gaussian



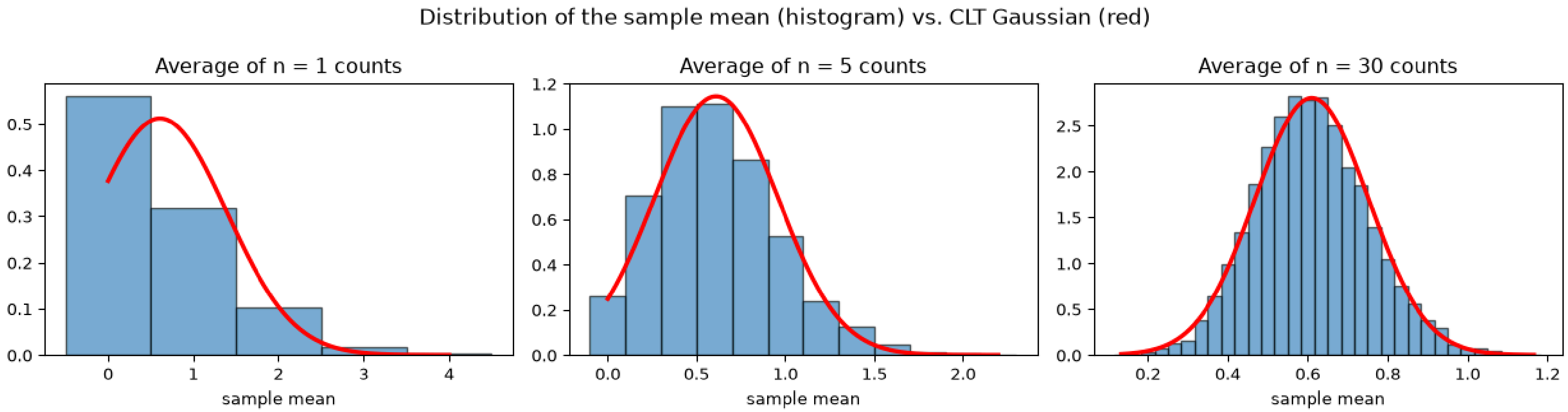




$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

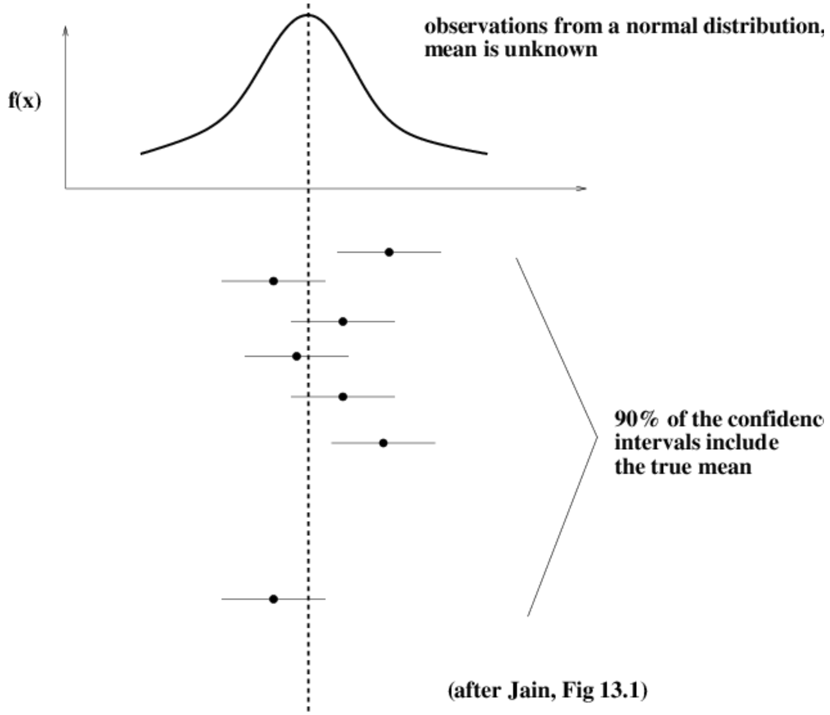
$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

# Idea 4 — the CLT: sample means become Gaussian, and narrow



# Idea 4, continued — a confidence interval, and how to read it

$$\bar{x} \;\pm\; z_{1-\alpha/2}\,\frac{s}{\sqrt{n}}, \qquad z_{0.975} \approx 1.96$$




Each bar is the interval from one dataset; most cover the true mean, a f...

# In-Class Activity

# Activity: fit, check — then simulate what the theorems promise

Goal: run the recipe yourself on a **month of Boston temperatures** (Gaussian, MLE by hand and with `scipy`) and on the **horse-kick counts** (Poisson), check each fit and use it — then, as the **stretch**, treat your fitted Poisson as a known population and *watch* the Law of Large Numbers, the Central Limit Theorem, and the coverage of your own 95% confidence intervals.

- Work in groups of 2–3. Sections A1 and B1 get different months, corps, and seeds — open **your** section's notebook.
- Parts marked **(autograded)** are submitted to Gradescope; the rest is participation.
- Staff will circulate — be ready to explain any part of your work.
- Dependencies: `numpy`, `pandas`, `matplotlib`, `scipy` only.

### Section A1

Open in Colab | Open on GitHub

# Going deeper

Full lecture notes: Probabilistic Modeling: Fitting Distributions to Data

- Checking the Fit and Checking the Fit Quantitatively — the overlay and the $\pm k\sigma$ fractions
- The Model: The Poisson Distribution — counts, and the observed-vs-predicted table
- Likelihood Function → Gaussian MLEs → The Same Recipe for the Poisson; full derivation in Appendix 06A
- Covariance, Correlation, The Multivariate Gaussian
- The Central Limit Theorem, The Standard Error, Reading a Confidence Interval Correctly
- Background: Probability and Statistics Refresher
- Next up: Gaussian Mixture Models and EM (06) — the same likelihood, several Gaussians at once.