Foundations of Data Science

Lecture 9: Model criticism and comparison

Prof. Gilles Louppe
g.louppe@uliege.be

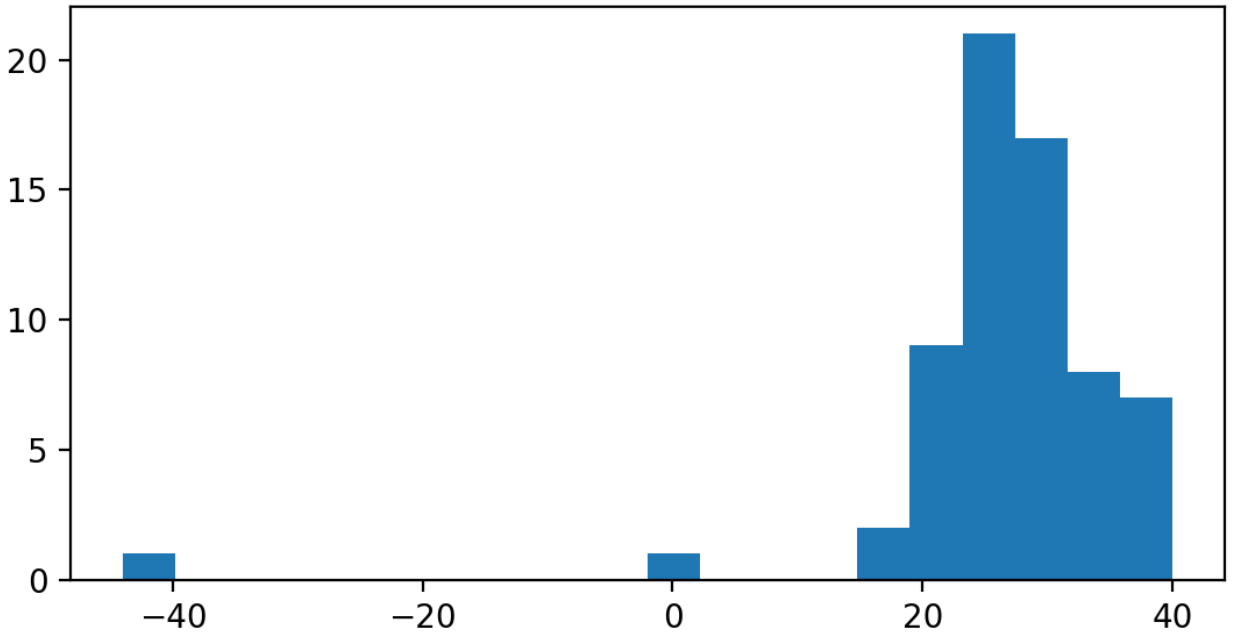
LIÈGE université

If it (a model) disagrees with experiment, it is wrong.
In that simple statement is the key to science. -- Richard Feynman

Model checking

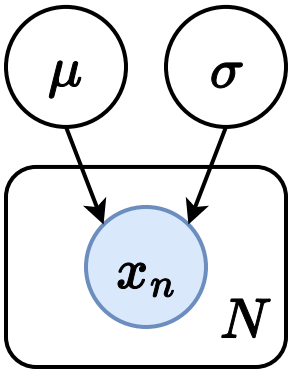
Newcomb's experiment (1882)

In a famous experiment, Simon Newcomb measured the speed of light using a rotating mirror. He collected 66 measurements of the time taken by light to travel a known distance.



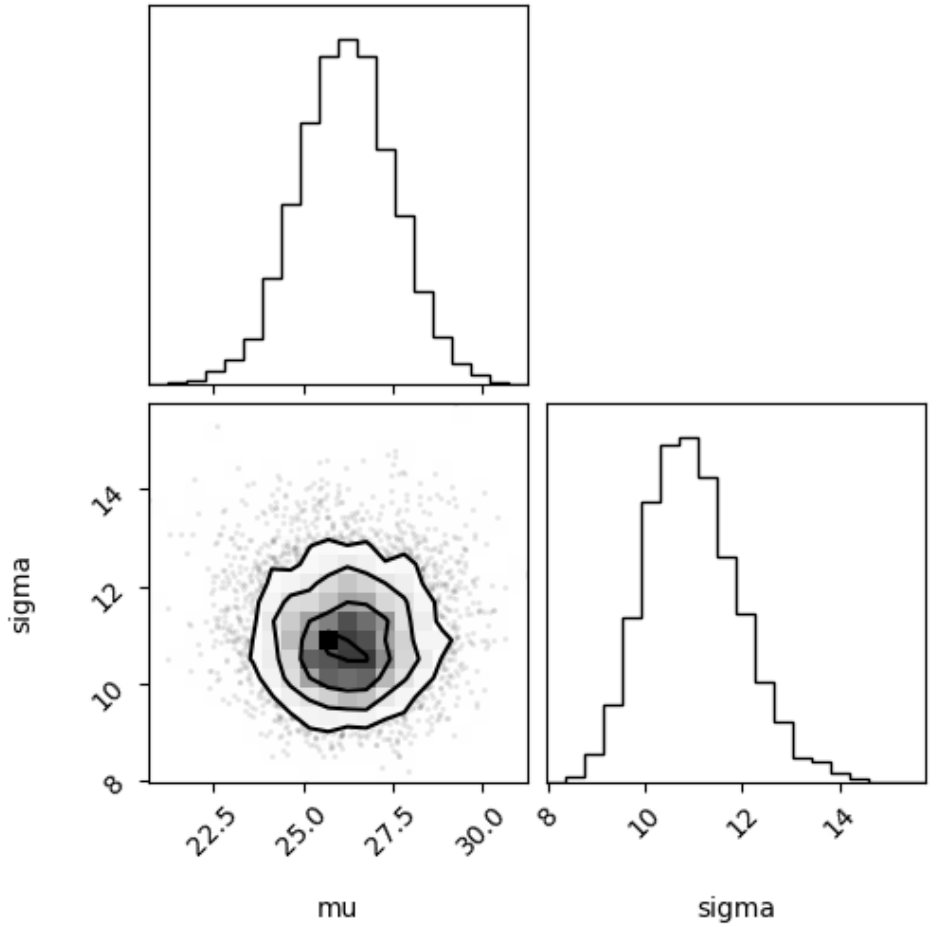
Histogram of Newcomb's measurements of the speed of light (in deviations from 24800 nanoseconds).

We consider a simple Gaussian model for these measurements:



$$\mu \sim \mathcal{U}(-1000, 1000)$$
$$\sigma \sim \mathcal{U}(0.1, 1000)$$
$$x_n \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for } n = 1, \ldots, 66.$$

Using MCMC, we obtain samples from the posterior $p(\mu, \sigma \mid x_{1:N})$:



From the samples, we can estimate the speed of light (in deviations from 24800 nanoseconds) as the empirical mean of the posterior samples of μ,

$$\hat{\mu} = \frac{1}{M} \sum_{m=1}^{M} \mu^{(m)} = 26.20949771717204 \text{ nanoseconds.}$$

Reporting this many digits is misleading, as it suggests a precision that is not supported by the data or model. Do not report more digits than justified by the uncertainty in the estimate $\hat{\mu}$!

Credible intervals

In the Bayesian framework, **credible intervals** provide a way to quantify uncertainty in parameter estimates.

Assuming a joint model $p(\theta, x)$ over parameters θ and data x, a credible interval at level $1 - \alpha$ is an interval $[a, b]$ such that

$$P(a \leq \theta \leq b \mid x) = 1 - \alpha.$$

The highest posterior density (HPD) interval is a common choice, defined as the narrowest interval containing $1 - \alpha$ of the posterior probability.

For our speed of light estimate, a 95% credible interval would be computed from the posterior samples of μ as the interval between the 2.5th and 97.5th percentiles of the samples, yielding

$$[\mu_{2.5\%}, \mu_{97.5\%}] = (23.71, 28.89).$$

Note that credible intervals capture our Bayesian uncertainty about parameter estimates given the data x, unlike **Frequentist confidence intervals** which would capture the variability of estimates across hypothetical repeated samples, assuming the parameters are fixed but unknown.

Posterior predictive checks

The posterior predictive distribution is the distribution

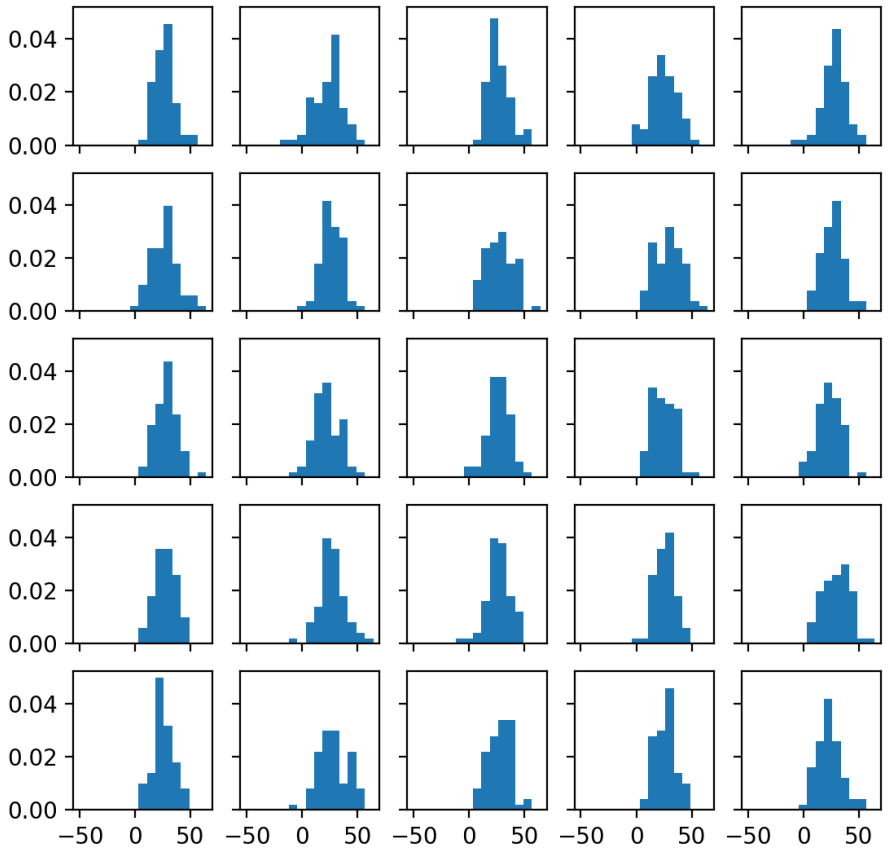
$$p(x^{\text{rep}} \mid x) = \int p(x^{\text{rep}} \mid \theta) p(\theta \mid x) d\theta$$

of replicated data x^{rep} given observed data x.

If our model is a good fit to the data, then replicated data x^{rep} should resemble the observed data x: **posterior predictive checks** aim to assess this resemblance or lack thereof.

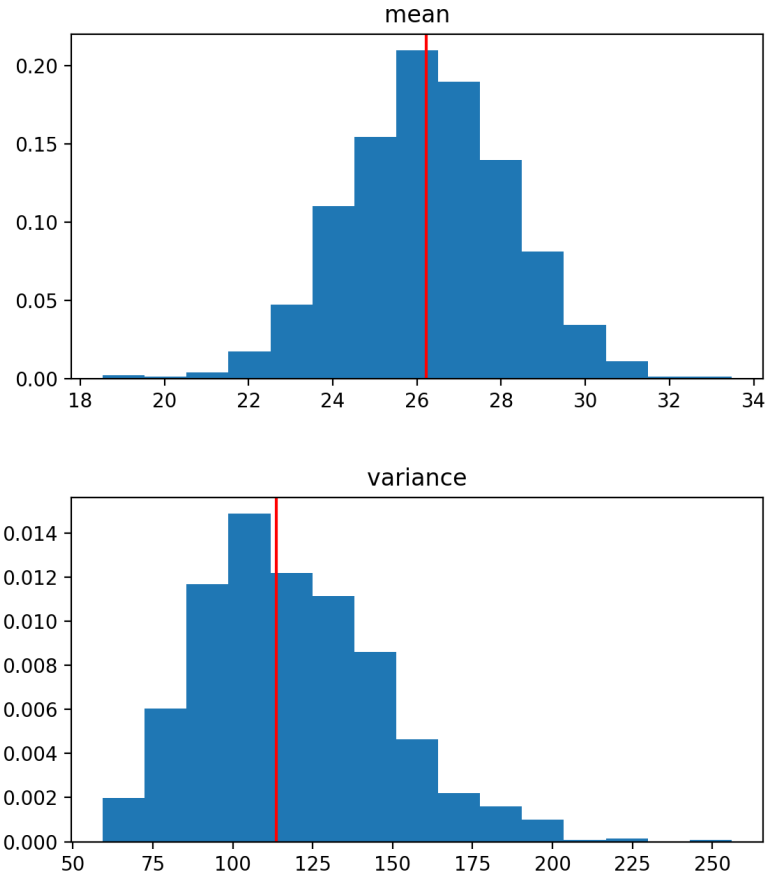
In our example, we generate replicated datasets by

- sampling parameters $(\mu^{(m)}, \sigma^{(m)}) \sim p(\mu, \sigma \mid x_{1:N})$ from the posterior,
- and simulating new data points $x_n^{\mathrm{rep}(m)} \sim \mathcal{N}(\mu^{(m)}, \sigma^{2(m)})$ for $n = 1, \ldots, N$ and $m = 1, \ldots, M$.

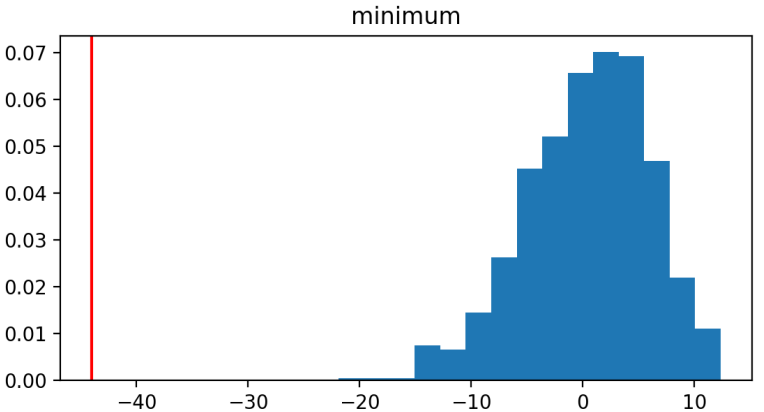


Histograms of 25 replicated datasets drawn from the posterior predictive distribution.

To quantify specific aspects of the model fit, we can also compare **summary statistics** $T(x)$ of the observed data to the distribution $p(T(x^{\text{rep}})|x)$ of those statistics computed on replicated datasets.



For our Gaussian model, posterior predictive distributions of the mean and variance of replicated datasets are consistent with the observed statistics, indicating a good fit in terms of location and spread.



For the minimum statistic, however, the situation is different. The observed statistic is poorly captured, indicating a potential model misfit.

Beyond visual checks, we can also quantify whether the observed statistics are extreme under the posterior predictive distribution, using **Bayesian p-values** defined as

$$P(T(x^{\text{rep}}) \geq T(x) \mid x) = \int P(T(x^{\text{rep}}) \geq T(x) \mid \theta) p(\theta \mid x) d\theta.$$

A Bayesian p-value close to 0 or 1 indicates a poor model fit for the statistic T.

Note that Bayesian p-values account for uncertainty in parameters via the posterior distribution, whereas Frequentist p-values

$$P(T(x^{\text{rep}}) \geq T(x) \mid \theta)$$

condition on a fixed parameter value θ.

For our example, we find the following Bayesian p-values:

- Mean: 0.513
- Variance: 0.528
- Minimum: 1.0

The extreme p-value for the minimum statistic confirms the poor fit of our Gaussian model to the lower tail of the data.

Residual analysis

When we have multiple observations $x_{1:N}$, another way to assess model fit is through **residual analysis**. Assuming the forward model is defined as a deterministic function plus additive noise, i.e.,

$$x_n = f(\theta) + \sigma\epsilon_n,$$

where σ is a scale parameter and $\epsilon_n \sim p(\epsilon)$ is noise, (standardized) **residuals** are computed as

$$r_n = \frac{x_n - f(\theta)}{\sigma}$$

for $n = 1, \ldots, N$ and $\theta \sim p(\theta \mid x_{1:N})$.

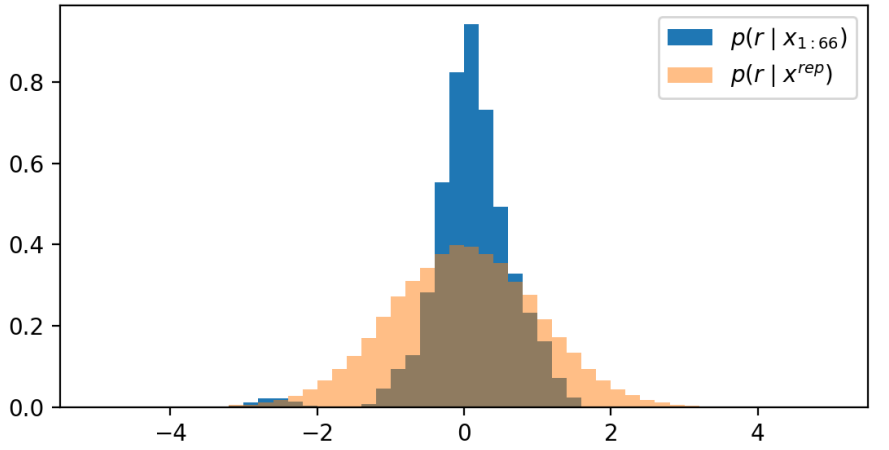
If the model is appropriate, the distribution of residuals $p(r \mid x_{1:N})$ for the observed data should match the distribution $p(r \mid x^{\text{rep}})$ of residuals for replicated data.

If the posterior is concentrated, then residuals should approximately follow the noise distribution $p(\epsilon)$.

For our Gaussian model, standardized residuals are computed as

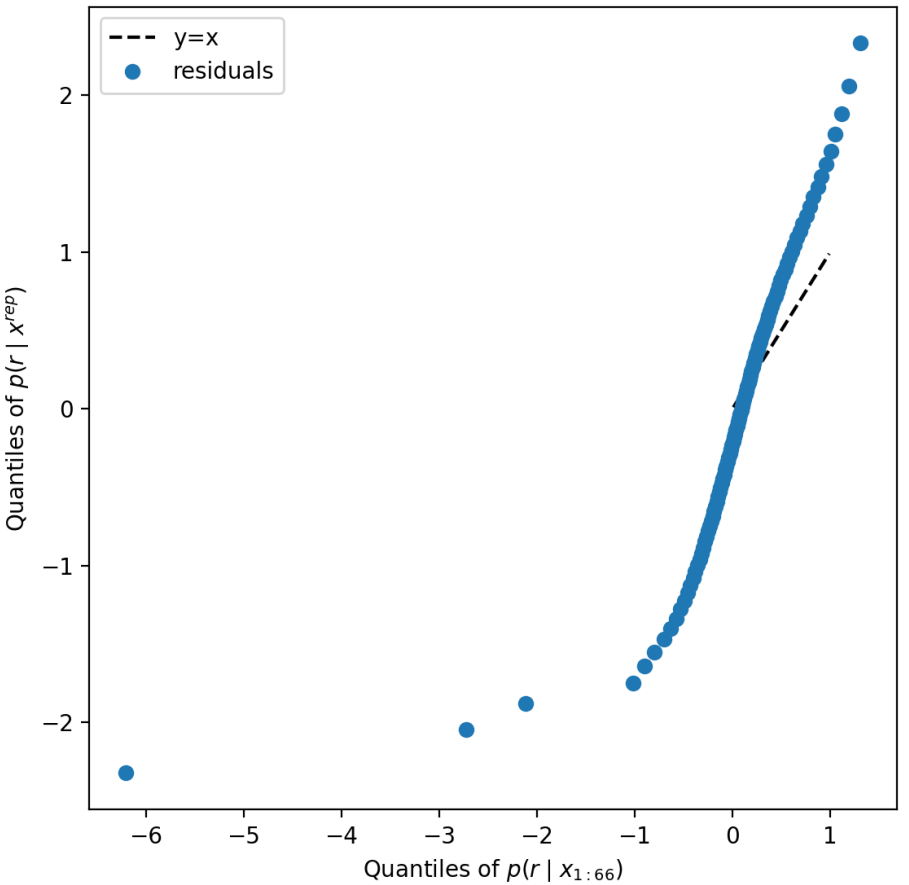
$$r_n = \frac{x_n - \mu}{\sigma},$$

for $n = 1, \ldots, 66$ and $\mu, \sigma \sim p(\mu, \sigma \mid x_{1:66})$.



A **quantile-quantile (Q-Q) plot** compares the quantiles of two distributions.

In our context, a Q-Q plot can be used to compare the distribution of residuals for the observed data to the distribution of residuals for replicated data. Systematic deviations from the diagonal line indicate model misfit.



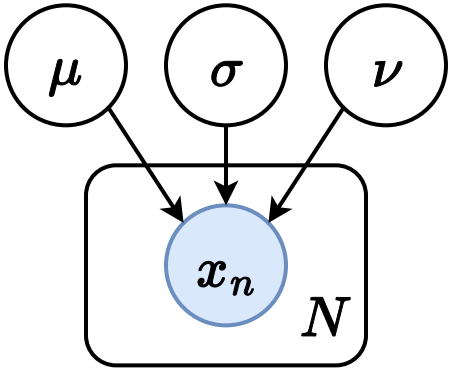
Deviations from the diagonal line indicate model misfit, particularly in the lower tail.

In summary, model checking provides a principled way to assess model fit by comparing observed data to data simulated from the model, using both visualizations and summary statistics.

It is meant to reveal **model misspecifications** and failures, guiding us towards better models.

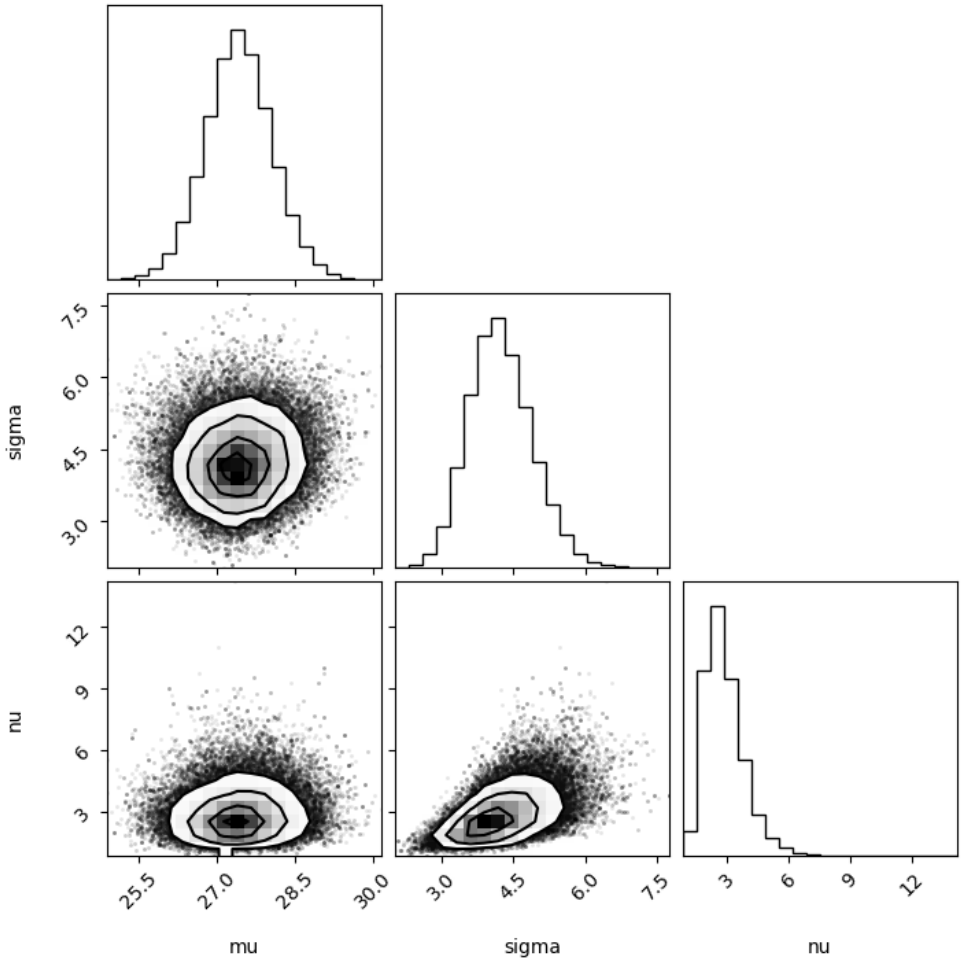
Model comparison

We now assume an alternative *t*-location-scale model for Newcomb's data to account for the heavy tails observed in the measurements:



$$
\begin{aligned}
\mu &\sim \mathcal{U}(-1000, 1000) \\
\sigma &\sim \mathcal{U}(0.1, 1000) \\
\nu &\sim \text{Gamma}(2, 0.1) \\
\epsilon_n &\sim t_\nu \\
x_n &= \mu + \sigma\epsilon_n \quad \text{for } n = 1, \ldots, 66.
\end{aligned}
$$

Again, we use MCMC to obtain samples from the posterior $p(\mu, \sigma, \nu \mid x_{1:N})$ under this new model.



Bayes factors

Assume we have two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$. As good Bayesian citizens, we also assign prior probabilities to each model, $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$.

The Bayesian approach to model comparison then consists in comparing the posterior probabilities of each model given the data x,

$$\frac{p(\mathcal{M}_1 \mid x)}{p(\mathcal{M}_2 \mid x)} = \frac{p(x \mid \mathcal{M}_1)}{p(x \mid \mathcal{M}_2)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}$$

where the first term on the right-hand side is called the **Bayes factor** $\mathrm{BF}_{1,2}$.

The Bayes factor quantifies how much more likely the observed data is under one model compared to another. If $\mathrm{BF}_{1,2} > 1$, the data favors model $\mathcal{M}_1$ over $\mathcal{M}_2$, and vice versa.

A common scale for interpreting Bayes factors is:

- 1 to 3: Weak evidence
- 3 to 10: Moderate evidence
- 10+: Strong evidence

For Newcomb's data, we find that $\log \mathrm{BF}_{t,\mathrm{Gaussian}} \approx 30$, indicating strong evidence in favor of the *t*-location-scale model over the Gaussian model.

In practice, evaluating Bayes factors requires computing the **marginal likelihoods**

$$p(x \mid \mathcal{M}_i) = \int p(x \mid \theta, \mathcal{M}_i) p(\theta \mid \mathcal{M}_i) d\theta,$$

which is typically challenging, even for simple models.

One practical approach to approximate the marginal likelihood is the **Laplace approximation**.

Let $\hat{\theta} = \arg\max_\theta p(\theta \mid x, \mathcal{M})$ be the MAP estimate. The Laplace approximation approximates the posterior as

$$p(\theta \mid x, \mathcal{M}) \approx \mathcal{N}(\theta \mid \hat{\theta}, \Sigma)$$

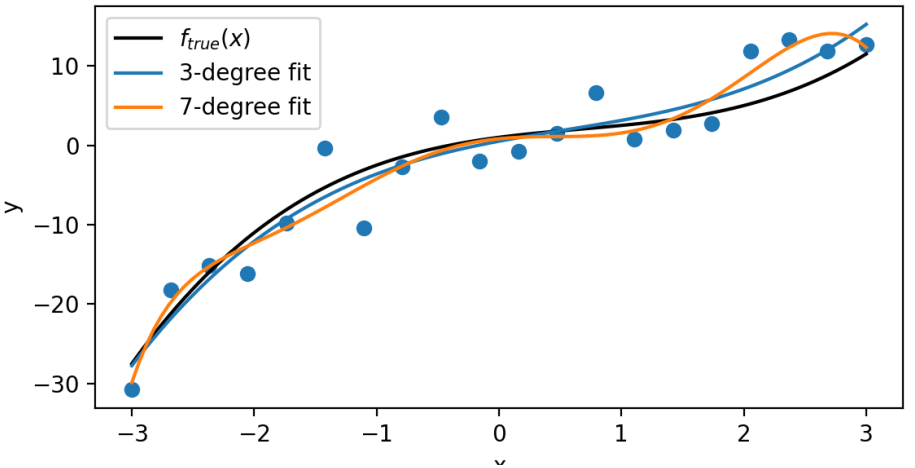
where Σ is the inverse Hessian of $-\log p(\theta \mid x, \mathcal{M})$ evaluated at $\hat{\theta}$.

Using this approximation, the marginal likelihood can be approximated as

$$\begin{aligned} p(x \mid \mathcal{M}) &= \frac{p(x \mid \hat{\theta}, \mathcal{M}) p(\hat{\theta} \mid \mathcal{M})}{p(\hat{\theta} \mid x, \mathcal{M})} \\ &\approx p(x \mid \hat{\theta}, \mathcal{M}) p(\hat{\theta} \mid \mathcal{M}) (2\pi)^{d/2} |\Sigma|^{1/2} \end{aligned}$$

where $(2\pi)^{d/2}|\Sigma|^{1/2}$ is the inverse posterior density at the MAP estimate under the Laplace approximation.

Since the marginal likelihood integrates over all parameter values, using it for model comparison automatically implements a form of Bayesian **Occam's razor**: simpler models with less capacity are favored unless the data strongly supports the need for a more complex model.



Polynomial regression fits of varying degrees to noisy data. The marginal likelihood favors the 3-degree model ($\log p(x \mid \mathcal{M}) \approx -26$) over the 7-degree model ($\log p(x \mid \mathcal{M}) \approx -45$), balancing fit and complexity, even if the 7-degree model includes the 3-degree model as a special case.

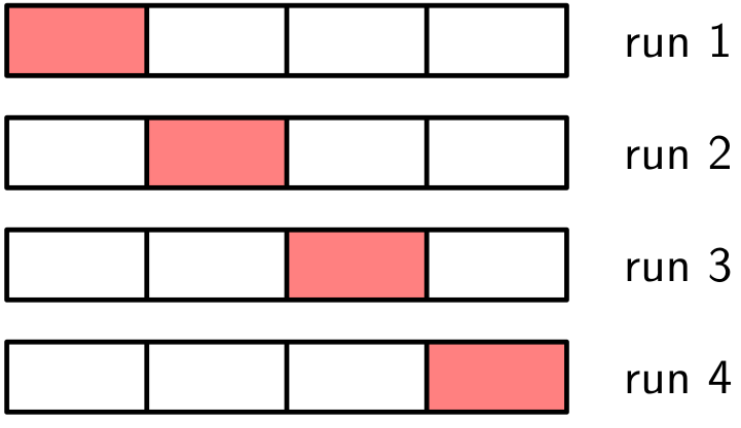
Cross-validation

An alternative approach to model comparison is **cross-validation**, which estimates the predictive performance of models on held-out data.

Assume we have a dataset $x_{1:N}$ and we want to evaluate how well a model $\mathcal{M}$ predicts unseen data.

The most common form of cross-validation is **k-fold cross-validation**, where the data $x_{1:N}$ is partitioned into k equally sized folds of N/k observations each.

Each fold is used once as a validation set while the remaining $k-1$ folds form the training set. The predictive performance is averaged over the k folds.



A built-in performance metric for Bayesian models is the **expected log predictive density (ELPD)**, defined as

$$\mathbb{E}_{p_{\text{true}}(x')}[\log p(x'|x)]$$

where $p_{\text{true}}(x')$ is the true data-generating distribution and $p(x'|x)$ is the posterior predictive distribution given observed data x (those in the training set).

ELPD ideally combines with cross-validation, as held-out data can be used to estimate the expectation.

For Newcomb's data, using 5-fold cross-validation, we find that the ELPD for the *t*-location-scale model (EPLD=-48) is higher than that of the Gaussian model (EPLD=-68), confirming its superior predictive performance.

Note that approximating the posterior distribution by a point estimate (e.g., MAP) reduces ELPD to the familiar **log-likelihood** evaluated on held-out data.

The end.