Bayesian inference

A: Frequentist versus Bayesian estimation

Inferring an unknown parameter

We get data $x_1, \ldots, x_n \sim p_\theta$ and would like to estimate θ.

- **Frequentist: treat θ as an unknown, non-random quantity.**

  For instance, pick the maximum-likelihood value

  $$\arg\max_\theta \Pr(x_1, \ldots, x_n | \theta) \;=\; \arg\max_\theta \prod_{i=1}^n p_\theta(x_i).$$

- **Bayesian: treat θ as a random variable with a prior distribution q_o.**

  Given the data, the posterior distribution of θ is

  $$q_n(\theta) \;=\; \Pr(\theta | x_1, \ldots, x_n) \;\propto\; q_o(\theta) \Pr(x_1, \ldots, x_n | \theta).$$

Some good references

- Bradley Efron. A 250-year argument: Belief, behavior, and the bootstrap.
- Andrew Gelman, John Carlin, Hal Stern, Donald Rubin. Bayesian Data Analysis.
- Kevin Murphy. Machine Learning: A Probabilistic Perspective.

Inferring a binomial parameter

(From Gelman.) What fraction of human births are female?

- Laplace looked at children born in Paris, 1745–1770.

$$\# \text{ girls} = 241{,}945, \ \# \text{ boys} = 251{,}527, \ \text{total} = 493{,}472.$$

  Female fraction = 0.490.

- Mathematical setup:
  - Let θ be the probability that a child is female.
  - Let n be the number of children observed.

  Suppose we see F females and M males. Then $F \sim \text{binomial}(n, \theta)$.

- Bayesian inference: put a prior on θ.
  Simple choice: uniform prior, $q_o(\theta) = 1$ for all $\theta \in [0, 1]$.

- Laplace asked: What is the probability that $\theta < 0.5$?

- Uniform prior: $q_o \equiv 1$.
- What is the posterior q_n after seeing $F = f, M = m$?

B: Binomial and beta

The beta distribution

For $\alpha, \beta > 0$, the beta(α, β) distribution on $[0, 1]$ has functional form

$$p(x) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

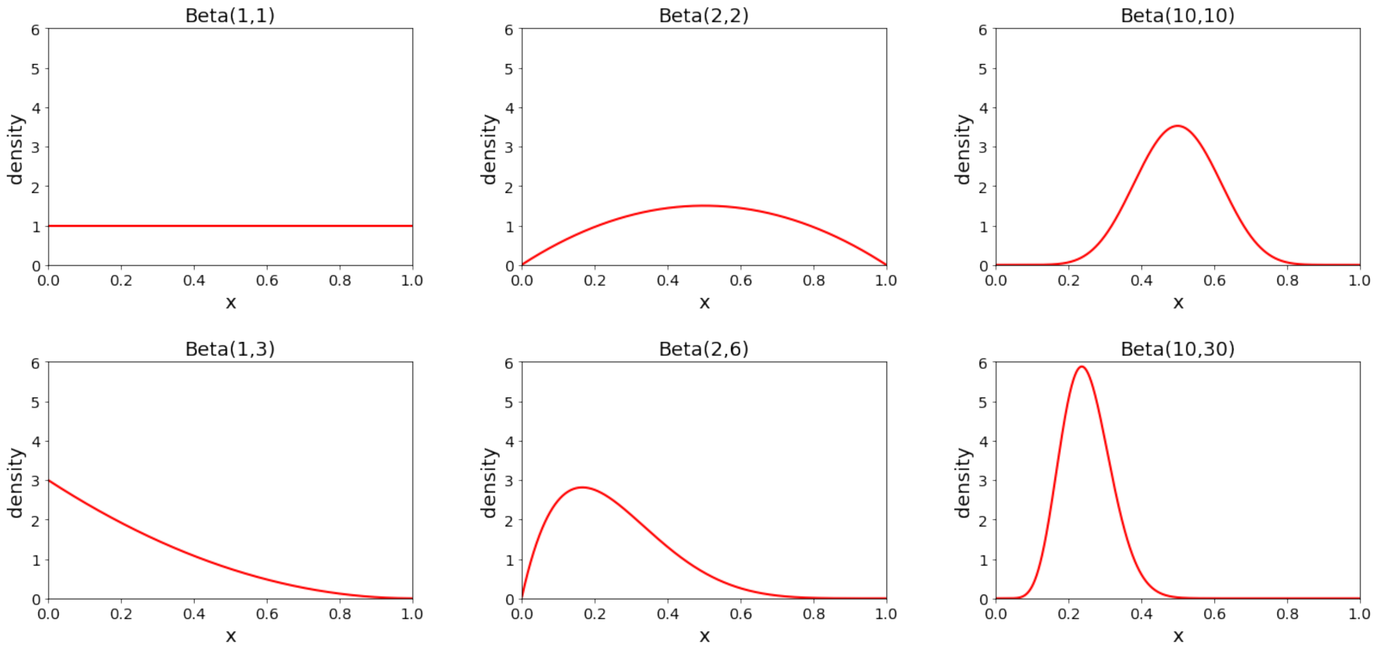
where $\Gamma(\cdot)$ is the gamma function.

Recall $\Gamma(z) \;=\; \int_0^\infty t^{z-1}e^{-t}dt$. Useful identity: $\Gamma(z+1) = z\Gamma(z)$.

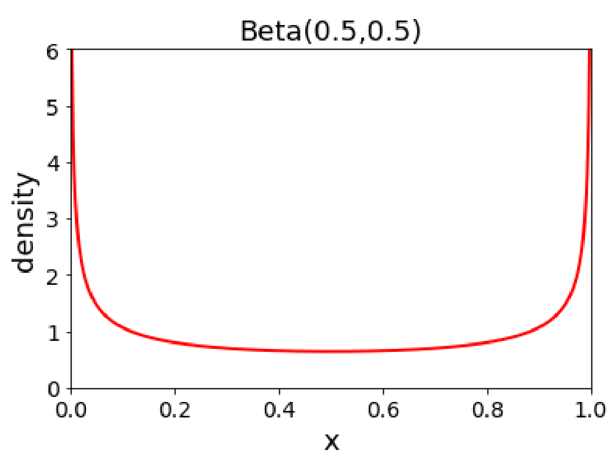
Basic properties of the beta(α, β) distribution:

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}$$
$$\text{Mode} = \frac{\alpha-1}{\alpha+\beta-2}$$
$$\text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta pictures



An interesting case



Laplace's law of succession

Another question that Laplace considered: What is the probability that the sun will rise tomorrow, given that it has risen every day for the past 5000 years?

- Let θ be the probability that the sun rises on any given day.
- Place a uniform prior on θ.
- We have $n = 5000 \times 365 = 1{,}826{,}213$ observations.
- Recall: **uniform prior** + **binomial likelihood** $\implies$ **beta posterior**.
- Posterior q_n is beta$(n+1, 1)$.

Answer:

$$\Pr(\text{sun rises tomorrow}) = \int_0^1 \theta \, q_n(\theta) \, d\theta = \mathbb{E}_{q_n}\theta = \frac{n+1}{n+2}.$$

Beta priors

Unknown probability $\theta \in [0, 1]$ (of female child, sun rising, etc.)

- Prior $q_o = \text{beta}(\alpha, \beta)$
- See n observations, of which s are successes
- What is the posterior q_n?

The beta is a conjugate prior for the binomial

Conjugate prior: posterior is from the same family as the prior.

$$\boxed{\textbf{beta prior} + \textbf{binomial likelihood} \implies \textbf{beta posterior}}$$

- Prior beta(α, β)
- Observe s successes and f failures
- Posterior is beta$(\alpha + s, \beta + f)$

Equivalent sample size: prior beta(α, β) is the same as

- Uniform prior
- Seeing $\alpha - 1$ successes and $\beta - 1$ failures

Prior is eventually overwhelmed by observations.

C: Handling non-conjugate priors

Binomial with an arbitrary prior

Coin of unknown bias θ. Use any prior q_o on $[0, 1]$:

See n observations, with h heads and t tails. Posterior:

$$q_n(\theta) \;\propto\; q_o(\theta) \cdot \theta^h \cdot (1-\theta)^t$$

How to answer questions like "What is $\Pr(\theta < 1/2)$?"

1. Fine gridding of $[0, 1]$, approximate q_o, q_n by discrete distribution over grid points.
2. Sample from q_n. Easy if q_n is beta. Otherwise, methods like rejection sampling.

Rejection sampling

Wish to sample from a distribution f, but we don't know how.

- There is a distribution g from which we know how to sample.
- $f(x)/g(x)$ is bounded, say $\leq M$.

Repeat:
- Draw $X \sim g$ and $U \sim \text{Unif}[0, 1]$
- If $U < f(x)/Mg(x)$: output X and halt

- What is the probability this procedure outputs a given x?
- What is the expected number of trials before a sample is generated?

D: Conjugate priors for exponential families

Conjugate priors for exponential families

Take any exponential family

$$p_\theta(x) \;=\; e^{\theta \cdot T(x) - G(\theta)} \pi(x)$$

with features $T(x) = (T_1(x), \ldots, T_k(x))$ and $\theta \in \Theta$.

Conjugate prior over Θ:

- Define features $U(\theta) = (\theta_1, \ldots, \theta_k, -G(\theta)) \in \mathbb{R}^{k+1}$
- Gives a $(k+1)$-parameter family indexed by $\eta = (\eta_1, \ldots, \eta_k), \lambda$:

  $$p_{\eta,\lambda}(\theta) \;=\; e^{(\eta,\lambda)\cdot U(\theta) - F(\eta,\lambda)} \nu(\theta)$$

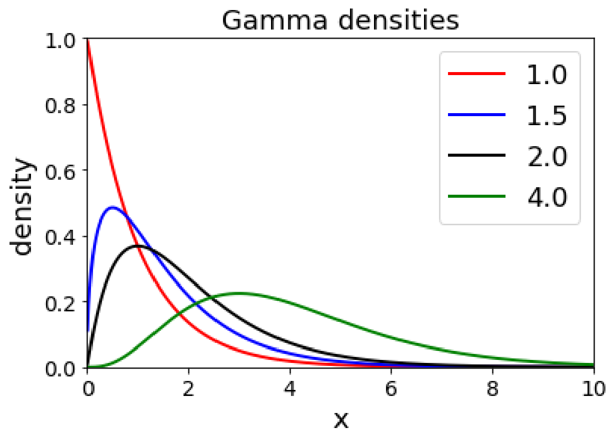
  with base measure ν on Θ.

Poisson with a gamma prior

Recall Poisson(θ) distribution over $\mathbb{N}$: $p_\theta(x) = e^{-\theta}\,\theta^x/x!$.

Conjugate prior: gamma(α, β) distribution over $\mathbb{R}^+$:

$$\Pr(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}e^{-\beta\theta}.$$



- Mean: α/β
- Mode: $(\alpha-1)/\beta$
- Variance: α/β^2

After seeing $x_1, \ldots, x_n$, what is the posterior?

- Poisson distribution $p_\theta(x) = e^{-\theta}\,\theta^x/x!$
- Gamma (α, β) prior: $\Pr(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}e^{-\beta\theta}$

E: Multinomial and Dirichlet

Inference with the multinomial distribution

Collection of proteins from the same family (similar sequence, structure, function). Each is a sequence of amino acids over a 20-letter alphabet $S = \{a_1, \ldots, a_{20}\}$. You align them and you want to model the distribution over S at each position.

$$\begin{array}{cccccc} \cdots & a_1 & a_6 & a_7 & a_{20} & \cdots \\ \cdots & a_3 & a_8 & a_7 & a_8 & \cdots \\ \cdots & a_1 & a_9 & a_7 & a_{20} & \cdots \end{array}$$

Infer the distribution $\theta \in \Delta_{20}$ for the first position.

- Vector of counts at this position: $(x_1, \ldots, x_{20}) = (2, 0, 1, 0, \ldots, 0)$
- Here $(x_1, \ldots, x_{20}) \sim \text{multinomial}(n = 3, \theta)$

What is the maximum-likelihood estimate θ_{ML}?

Dirichlet distribution

For $\alpha \in \mathbb{R}_+^k$, Dirichlet($\alpha$) is the distribution

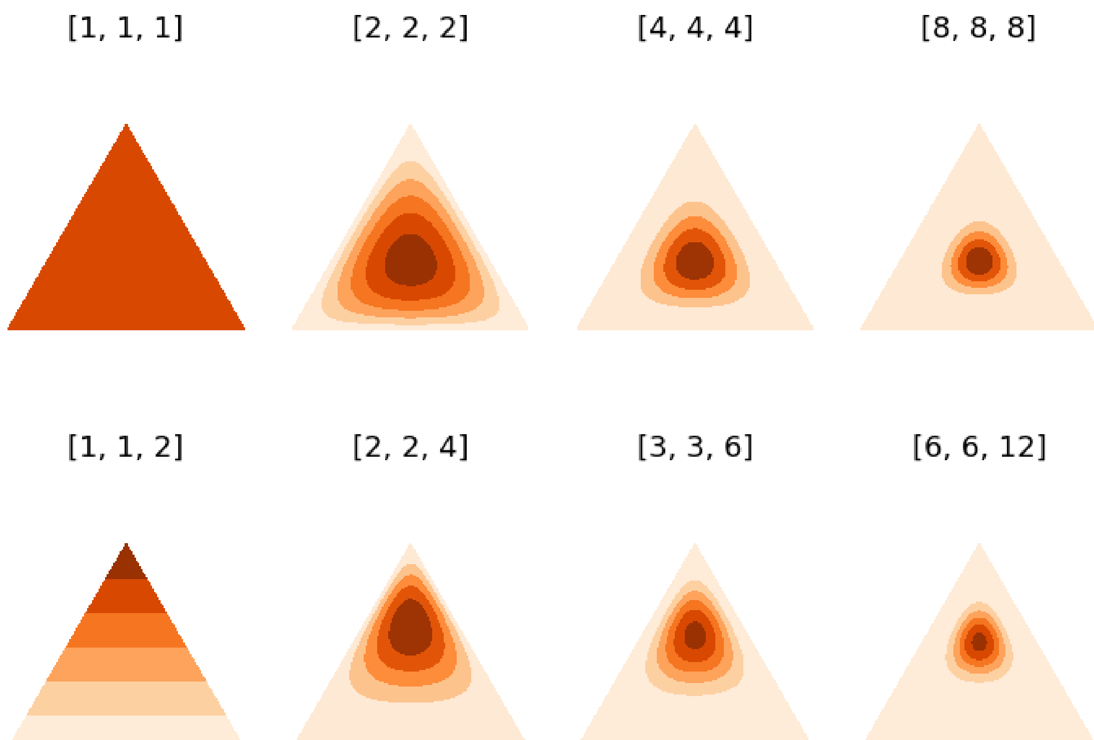
$$\Pr(\theta_1, \ldots, \theta_k) \;=\; \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \, \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}$$

over the probability simplex $\Delta_k = \{(\theta_1, \ldots, \theta_k) : \theta_i \geq 0, \sum_i \theta_i = 1\}$.

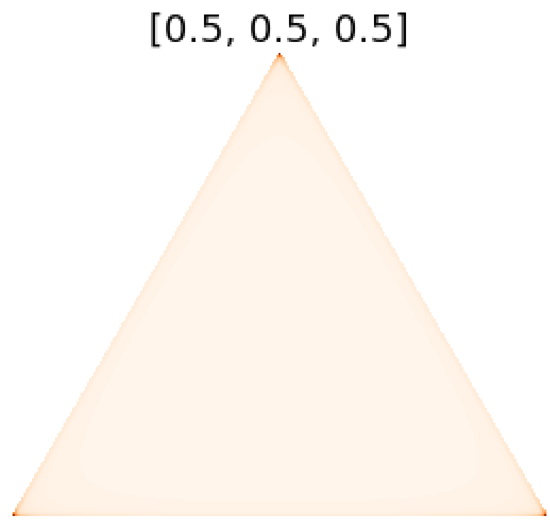
Properties:

$$\begin{aligned} \mathbb{E}\theta_i &= \frac{\alpha_i}{\alpha_1 + \cdots + \alpha_k} \\ \text{Mode: } \; \theta_i &= \frac{\alpha_i - 1}{\alpha_1 + \cdots + \alpha_k - k} \\ \mathrm{var}(\theta_i) &= \frac{\alpha_i(\alpha_1 + \cdots + \alpha_k - \alpha_i)}{(\alpha_1 + \cdots + \alpha_k)^2(\alpha_1 + \cdots + \alpha_k + 1)} \end{aligned}$$

Dirichlet pictures



One more



Dirichlet as a conjugate prior for the multinomial

Bayesian inference for distribution $\theta \in \Delta_k$:

- Prior: Dirichlet$(\alpha_1, \ldots, \alpha_k)$
- Draw n samples, get counts $(x_1, \ldots, x_k)$

What is the posterior distribution?

Choosing the prior parameters

By analyzing databases of proteins, we find (hypothetically):

- 25% of positions are *highly conserved*: concentrated on a single amino acid.

  a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11} a_{12} a_{13} a_{14} a_{15} a_{16} a_{17} a_{18} a_{19} a_{20}

- 12% of positions combine a particular set of amino acids with similar properties.

  a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11} a_{12} a_{13} a_{14} a_{15} a_{16} a_{17} a_{18} a_{19} a_{20}

- 8% of positions combine a different set of amino acids.

  a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11} a_{12} a_{13} a_{14} a_{15} a_{16} a_{17} a_{18} a_{19} a_{20}

But, how to combine these clusters?

F: Mixtures of conjugate priors

A mixture of conjugate priors is conjugate

Example: Beta-binomial. Coin of unknown bias $\theta \in [0, 1]$.

- Use prior $w_1 \, \text{beta}(\alpha_1, \beta_1) + w_2 \, \text{beta}(\alpha_2, \beta_2)$.
- See n coin tosses, of which h are heads and t are tails.

What is the posterior?

Mixtures of conjugate priors, cont'd

Unknown parameter θ. Prior is a mixture:

$$\sum_{j=1}^{k} \Pr(J = j) \Pr(\theta | J = j).$$

After seeing data, posterior is a mixture:

$$\sum_{j=1}^{k} \Pr(J = j | \text{data}) \Pr(\theta | J = j, \text{data}).$$