Brief review of elementary statistics: parameter estimation, confidence intervals, hypothesis testing

Roger Levy

9.S916: Statistical data analysis for scientific inference in cognitive science

11 April 2024

Running example

- I'm about to join a game of betting on the heads/tails outcome of a potentially bent coin
- I can't inspect the coin, but I can watch the coin being flipped "for a while" and record the outcomes
- The coin flips constitute a sequence of **Bernoulli random variables** conditionally independent of each other given the coin weighting $P(\text{heads}) = \pi$ with $0 \leq \pi \leq 1$
- Figuring out from observed data what the weighting is likely to be is **parameter estimation**
- In general, here we will use $\mathbf{y}$ to refer to observed-outcome **data** and θ to refer to the model parameters to be estimated

Characteristics of estimators

- **Estimator:** a procedure for guessing a quantity of interest within a population from a sample from that population
- For example, the **relative frequency estimator:** if we observe r instances of heads in n coin flips,

"this is an estimator"

$$\hat{\pi} = \frac{r}{n}$$

- Data are stochastic, so estimators give random variables!
- **Bias** of an estimator is $E[\hat{\theta}] - \theta$

$$E[\hat{\pi}] = E[\frac{r}{n}] = \frac{1}{n}E[r] = \frac{r}{n} = \pi$$ so $\hat{\pi}$ is **unbiased**

- **Variance** of an estimator is ordinary variance

$$\text{Var}(X) \equiv E[(X - E[X])^2] \qquad \text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$ (see reading materials)

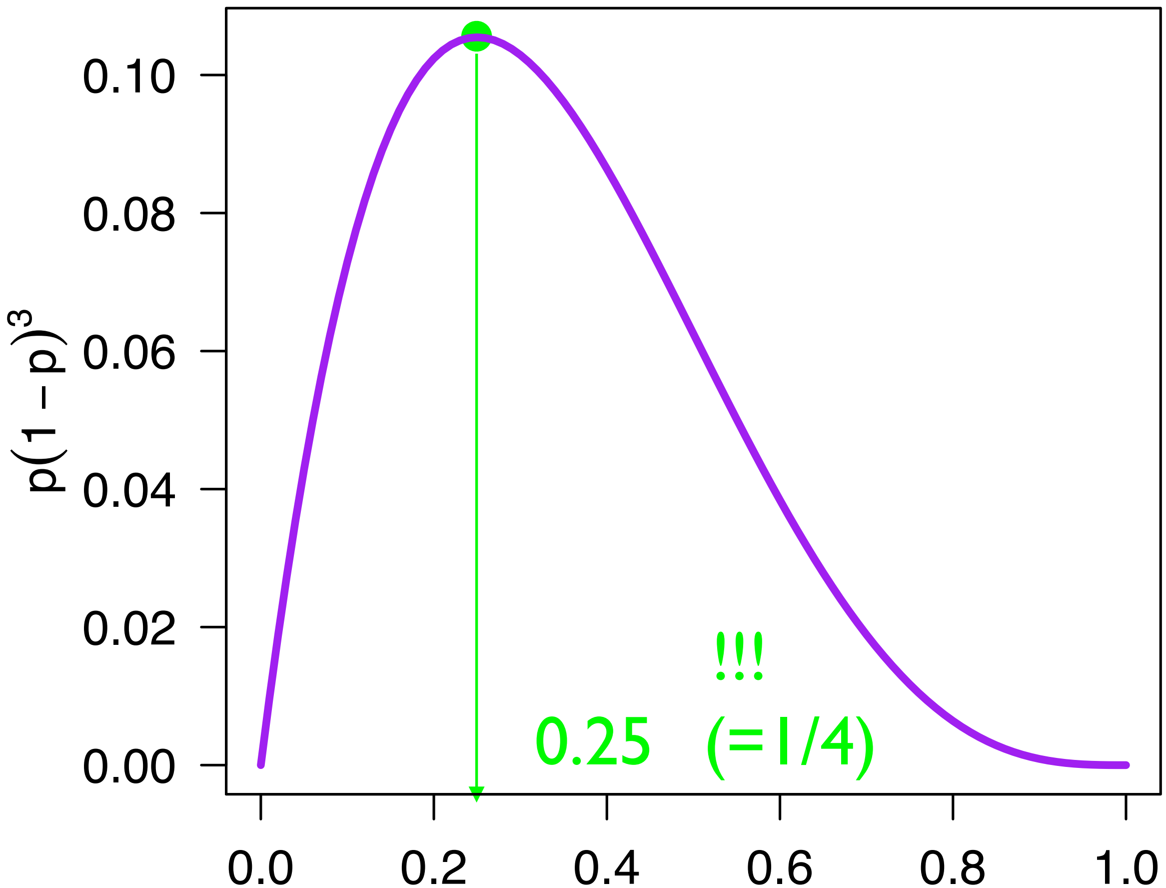
- Good estimators have favorable **bias–variance** tradeoff

Maximum likelihood estimation

$$\mathrm{Lik}(\boldsymbol{\theta};\boldsymbol{y}) \equiv P(\boldsymbol{y}|\boldsymbol{\theta}) \qquad \hat{\boldsymbol{\theta}}_{MLE} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} \mathrm{Lik}(\boldsymbol{\theta};\boldsymbol{y})$$

i	y_i
1	T
2	T
3	H
4	T

- p refers to the value of P(coin toss$_i$ = Heads)
- Likelihood for the following dataset



This is choosing the *maximum likelihood estimate* (**MLE**)

The **MLE** also turns out to be the *relative frequency estimate* (RFE)

(repeat slide from lecture 3)

The binomial distribution

- The **binomial distribution** is a two-parameter probability distribution over the number of **successes** in a number of independent, identically distributed (**iid**) **Bernoulli trials**
- Two parameters: Number of trials n & trial success parameter π
- A binomial-distributed random variable Y is simply the sum of n iid Bernoulli random variables with success parameter π

 $+$  $+ \ldots +$ 

- A binomial random variable has the following probability mass function:

$$P(Y = r) = \binom{n}{r} \pi^r (1 - \pi)^{n-r}$$

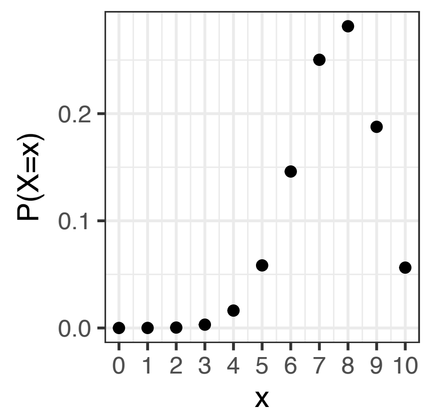


Image credit https://www.wannapik.com/vectors/75713

Bayesian parameter estimation

- Assume that the model parameters "intervene" between background knowledge I and data Y:

$$I \longrightarrow \boldsymbol{\theta} \longrightarrow Y$$

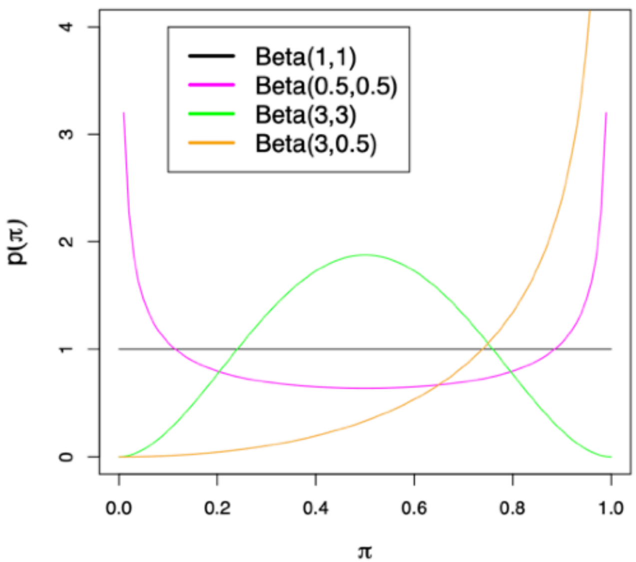
$$P(\boldsymbol{\theta}|\boldsymbol{y}, I) = \frac{P(\boldsymbol{y}|\theta, I)P(\theta|I)}{P(\boldsymbol{y}|I)}$$

$$= \frac{P(\boldsymbol{y}|\boldsymbol{\theta}) \qquad P(\boldsymbol{\theta}|I)}{P(\boldsymbol{y}|I)} \qquad (\text{because } \boldsymbol{y} \perp I \mid \boldsymbol{\theta})$$

- Then, if we assume a parametric form for $P(\mathbf{y} \,|\, \theta)$, we just need the prior $P(\theta \,|\, I)$

Example for coin flips: the beta distribution

- Express background knowledge I as two "pseudo-count" parameters α_1, α_2



- The beta distribution has form

$$P(\pi|\alpha_1,\alpha_2) = \boxed{\frac{1}{B(\alpha_1,\alpha_2)}}\boxed{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}$$

Normalizing constant, not of great interest for present purposes

Where the action is!

$$B(\alpha_1,\alpha_2) = \int_0^1 \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}\,d\pi$$

- Cool thing about the beta distribution: the posterior is also beta distributed! For $\mathbf{y} = m$ successes in n trials:

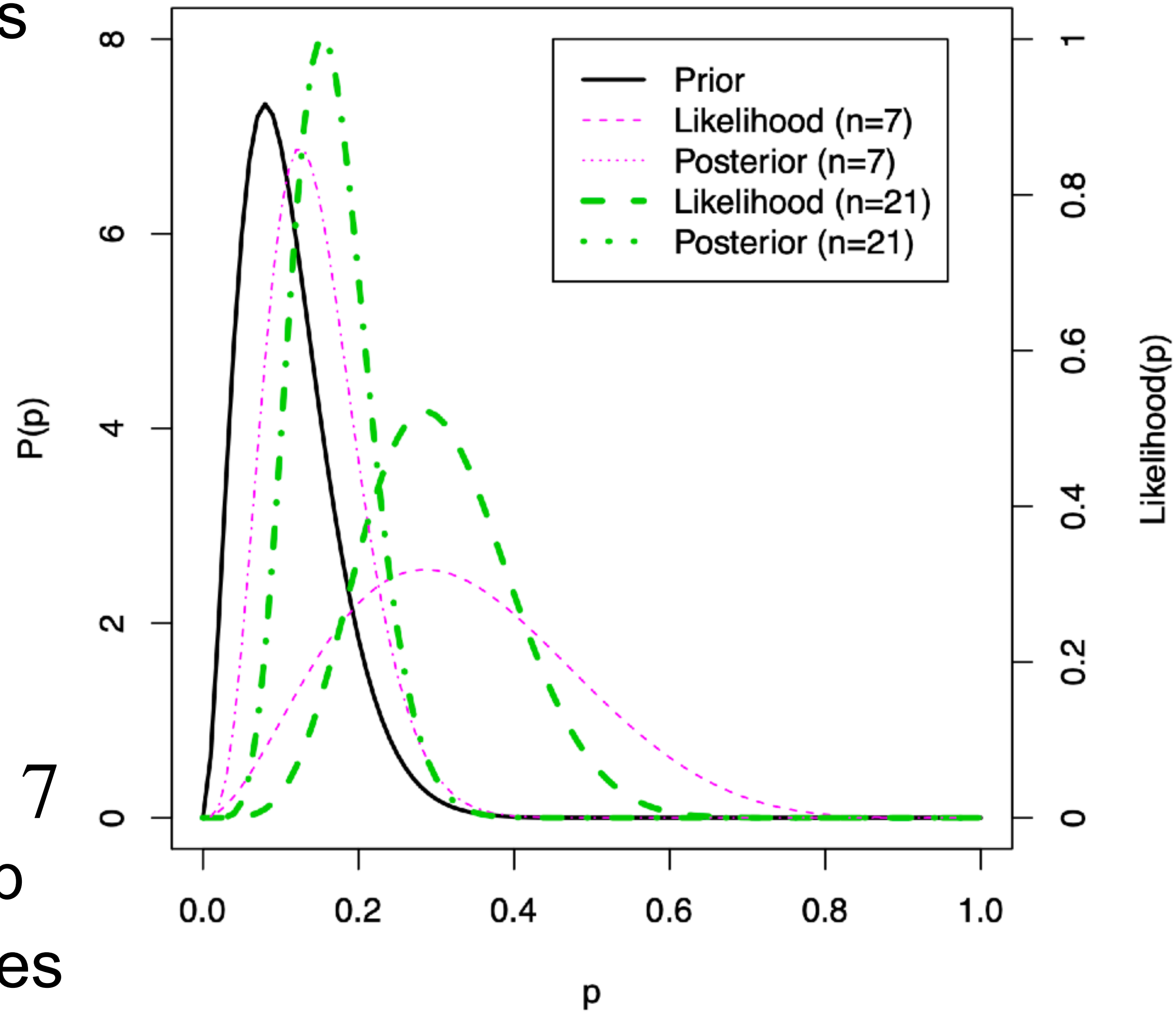
$$P(\pi|\boldsymbol{y},\alpha_1,\alpha_2) \propto \overbrace{\pi^m(1-\pi)^{n-m}}^{\text{Likelihood}}\overbrace{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}^{\text{Prior}}$$

- This property is called **conjugacy** and is convenient where available!

Example of Bayesian parameter estimation

- I inspect my coin and notice serious irregularities!

- My prior for $P(\text{heads})$: a $\alpha_1 = 3, \alpha_2 = 24$ Beta prior
- I flip the coin $n = 7$ times, it comes up heads $m = 2$ times

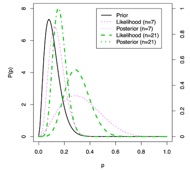


https://commons.wikimedia.org/wiki/
File:Post_medieval_coin,_Sixpence_of_William_III_(FindID_605939).jpg

Posterior prediction

$P(\text{heads}) = \pi$
$P(\pi) = \text{Beta}(\alpha_1, \alpha_2)$
Observe m heads out of n flips

Beta distribution



Our example

Posterior mean

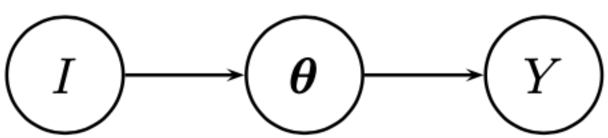
$$E[\pi \mid I] = \frac{\alpha_1}{\alpha_1 + \alpha_2} \qquad E[\pi \mid y, I] = \frac{\alpha_1 + m}{\alpha_1 + \alpha_2 + n}$$

Posterior mode
(when it exists)

$$\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2} \qquad \frac{\alpha_1 + m - 1}{\alpha_1 + \alpha_2 + n - 2}$$

Posterior predictive distribution

If I flip the same coin k more times, what is the distribution on the resulting # heads r?



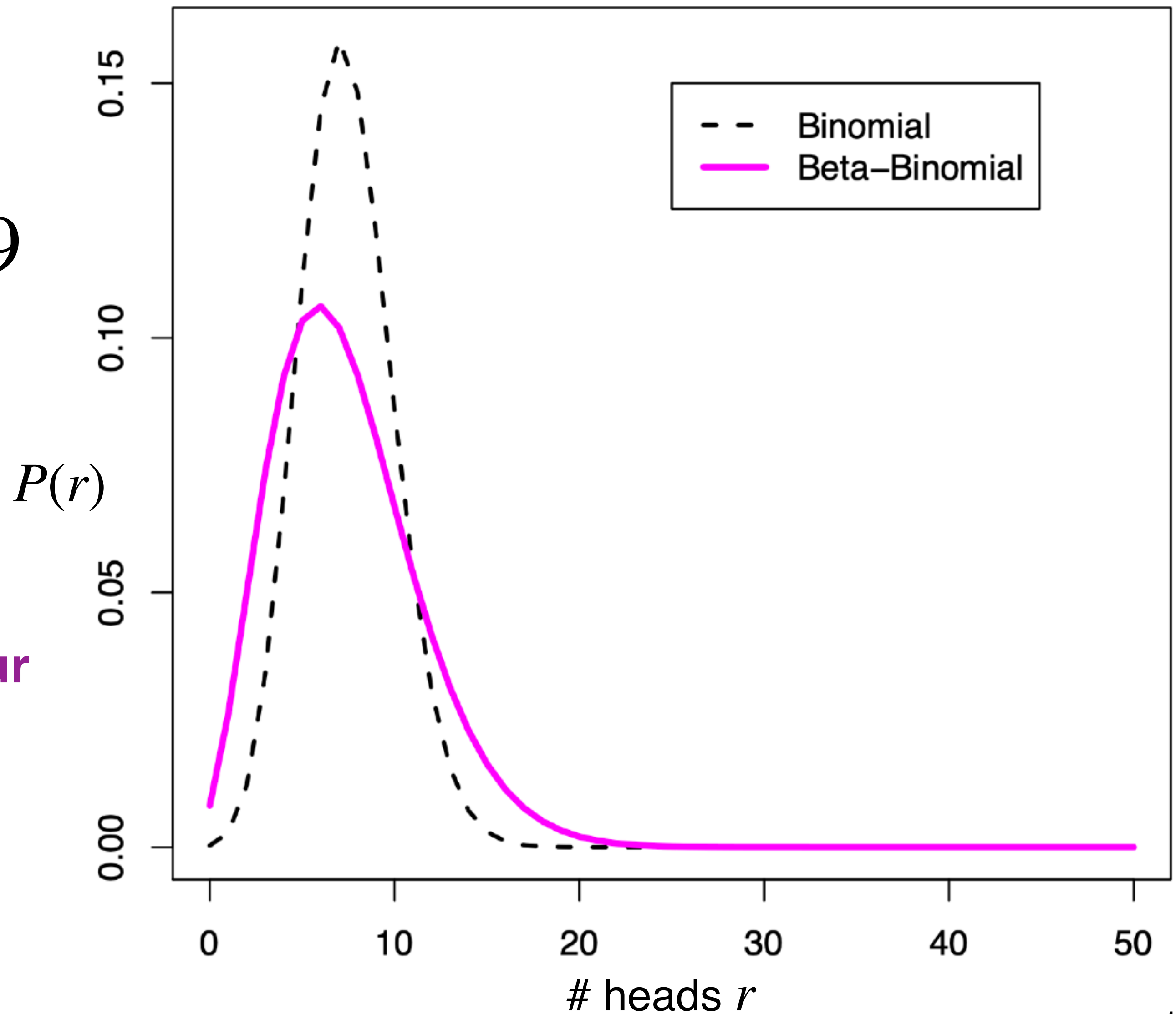
$$P(\boldsymbol{y}_{new} | \boldsymbol{y}, I)$$

→The **Beta-Binomial** model: $P(r|k, I, \boldsymbol{y}) = \binom{k}{r} \dfrac{B(\alpha_1 + m + r, \alpha_2 + n - m + k - r)}{B(\alpha_1 + m, \alpha_2 + n - m)}$

Point estimation vs Bayesian prediction

- Say we'll flip the coin $k = 50$ more times

$$\alpha_1 + m = 5$$
$$\alpha_2 + n - m = 29$$



Bayesian inference takes into account our uncertainty about model parameters θ, leading to more hedged predictions

A note on Bayesian priors

- The Bayesian prior is a double-edged sword
  - We get to specify it
  - We have to specify it
- In the above example, we used an **informative prior**
  - When we have strong domain knowledge, this can potentially be useful in various ways
- In scientific data analysis, however, our general goal is to *allow the data to speak to us about what we care about*
- For this reason, I generally advocate **vague priors for any part of the model whose posterior we care about**
- If your qualitative conclusions depend on choice of prior, it is a reason to be wary of the robustness of your analysis!
- As data become plentiful*, choice of prior *often but not always* recedes in importance

*What counts as "plentiful" depends on size of the model and structure of the data