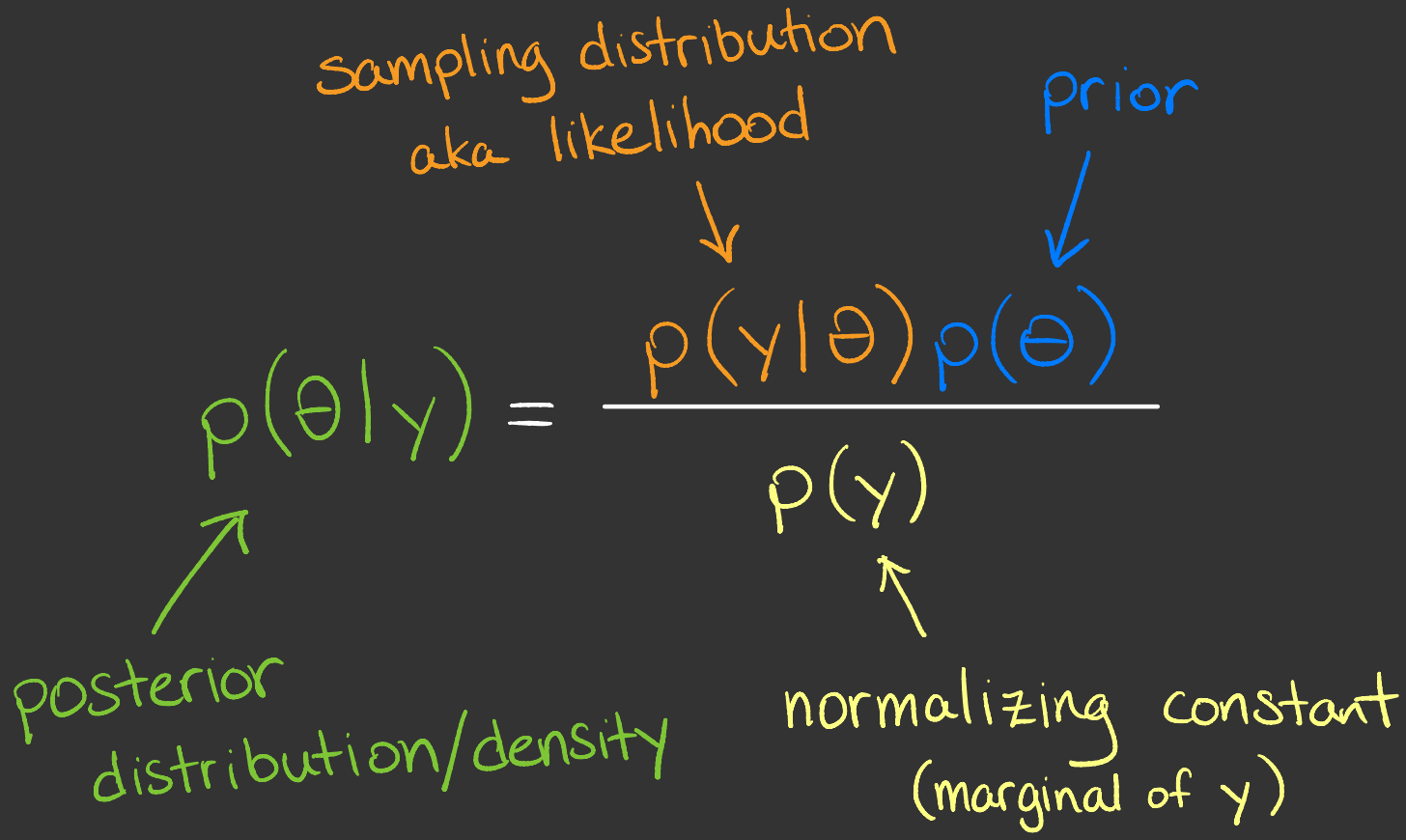
Hoff Ch. 3 : One-parameter models

Bayes rule (parameter estimation)



- This chapter: θ is a scalar (one-dimension)
- Two models, both <u>conjugate</u>:
    - Beta-binomial
    - Gamma-Poisson
- A class of priors $\mathcal{P}$ is <u>conjugate</u> for a sampling model $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}.$$

That is, the posterior belongs to the same class as the prior.

- Beta-binomial: Beta priors are conjugate for binomial sampling models
- Gamma-Poisson: Gamma priors are conjugate for Poisson sampling models

- Conjugacy greatly simplifies posterior computation

Part I : Beta-binomial

The binomial distribution

- State space $\mathcal{Y} = \{0, \ldots, n\}$.
- $\mathcal{Y}$-valued r.v. $Y \sim \text{Binomial}(n, \theta)$ if

$$p(y \mid \theta) = \mathbb{P}(Y = y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Recall, this is just the definition of $p(y \mid \theta)$

- Binom(n, θ) represents the probability

of obtaining a given number of successes (y) in n indepent trials, where each trial has the same probability of success, $\theta \in [0, 1]$.

- Examples
  - Number of heads in n flips
  - Number of defective items in a batch
  - Number of people who vote "yes"

- If $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, then

$$p(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} p(y_i | \theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$$

$$= \theta^{\sum_1^n y_i} (1-\theta)^{\sum_1^n 1-y_i} = \theta^{\Sigma y_i} (1-\theta)^{n-\Sigma y_i}$$

$$\Rightarrow \underbrace{Y = \sum_{i=1}^{n} Y_i}_{\text{sufficient statistic}} \sim \text{Binomial}(n, \theta)$$

sufficient statistic for θ and $p(y_1, \ldots, y_n | \theta)$

- If $Y \sim \text{Binomial}(n, \theta)$, then
  - $\mathbb{E}[Y|\theta] = n\theta$
  - $\text{Var}[Y|\theta] = n\theta(1-\theta)$

The beta distribution

- State space $\Theta = [0, 1]$
- Θ-valued r.v. $\theta \sim \text{Beta}(a, b)$ if

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^{a-1}(1-\theta)^{b-1}$$

↑ "proportional to"

- If $\Theta \sim \text{Beta}(a, b)$, then
  - $\mathbb{E}[\Theta] = \frac{a}{a+b}$
  - $\text{Var}[\Theta] = \frac{ab}{(a+b+1)(a+b)^2}$
- $\text{Beta}(1, 1) = \text{Uniform}(0, 1)$

Conjugacy of beta-binomial

- $p(\theta) \sim \text{Beta}(a, b)$
- $Y \sim \text{Binomial}(n, \theta)$

$$\implies p(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Recall HW 1, problem B1:

  If p and q are proportional pdfs on the same space, then $p = q$.

Proof. $p \propto q \Rightarrow \exists\, C > 0$ s.t. $p(\theta) = C q(\theta)$ $\forall\, \theta \in \Theta$. Therefore,

$$1 = \int_\Theta p(\theta)\, d\theta = C \int_\Theta q(\theta)\, d\theta = C. \quad \square$$

- Compute posterior:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

$$= \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^{y+a-1}(1-\theta)^{n+b-y-1}.$$

So $p(\theta|y)$ is a pdf on $[0,1]$ proportional

to $\theta^{y+a-1}(1-\theta)^{n+b-y-1}$

$$\Rightarrow p(\theta|y) = \text{Beta}(a+y, n+b-y)$$

- We have proved the following:

THEOREM (beta-binomial)

If $Y \sim \text{Binomial}(n,\theta)$ and $p(\theta) = \text{Beta}(a,b)$, then

$$p(\theta|y) = \text{Beta}(a+y, b+n-y).$$

- Posterior mean a combination of prior and data:

$$\mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$$

$$= \frac{a}{a+b+n} + \frac{y}{a+b+n}$$

$$= \frac{a}{a+b+n} \cdot \frac{a+b}{a+b} + \frac{y}{a+b+n} \cdot \frac{n}{n}$$

$$= \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\frac{y}{n}$$

Mean of $p(\theta)$

Mean of n iid Bernoulli r.v.s, $\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y}{n}$.

- $y = \sum_{i=1}^{n} y_i$ = number of observed successes in data
- n = total number of Bernoulli trials
- $\frac{y}{n}$ = mean probability of success (empirical)

Analogous interpretation for prior:

- a = "prior number of successes"
- b = "prior number of failures"
- $a+b$ = "prior sample size"
- $\frac{a}{a+b}$ = "prior probability of success"

Classroom exercise.

You conduct a survey asking whether or not people are happy ("yes" or "no"). 129 people respond, with 118 saying they are happy and 11 saying no. Write down a Bayesian model for these data. What are the posterior mean and variance?

Solution.

First, what is our parameter of interest?

Aswer: Proportion of people that are happy, $\theta \in [0,1]$.

Next:

Two ingredients: Likelihood and prior.

Likelihood: We have $n = 129$ samples Y_i where each Y_i is 1 (yes) or 0 (no). Assuming the responses are iid, a reasonable likelihood is

$$p(Y_1, \ldots, Y_{129} \mid \theta) = \theta^{\sum_1^{129} Y_i} (1-\theta)^{129 - \sum_1^{129} Y_i}.$$

$$= \theta^{118} (1-\theta)^{11}$$

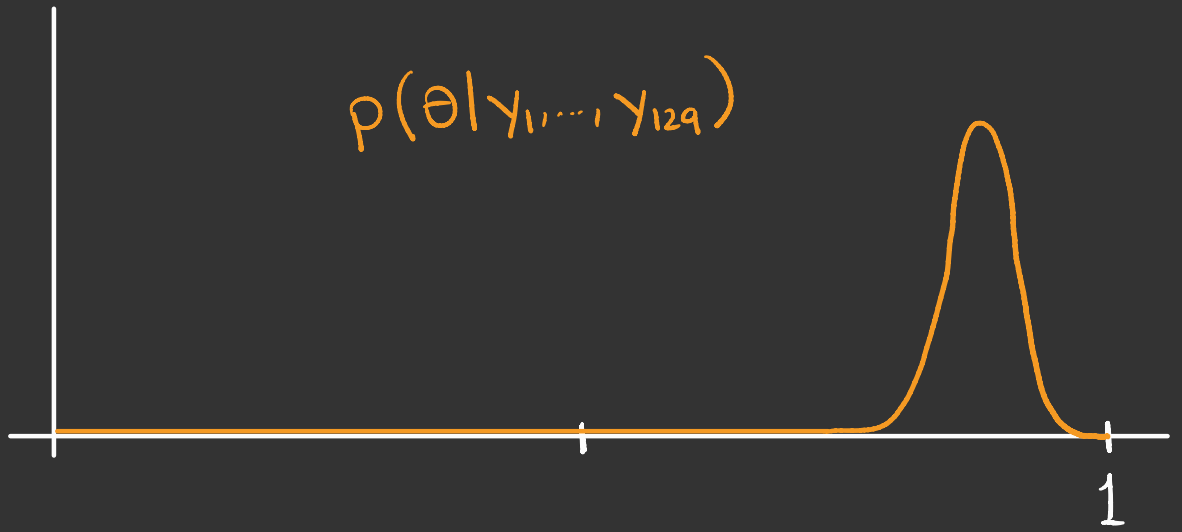
Prior: If we have no prior beliefs about θ, we can take $p(\theta) = \text{Beta}(1,1) = \text{Uniform}(0,1)$.

Compute posterior.

- $p(y_1, \dots, y_{129} | \theta) = \theta^{118}(1-\theta)^{11}$
- $p(\theta) = \text{Beta}(1,1)$

$$\Rightarrow \quad p(\theta | y_1, \dots, y_{129}) = \text{Beta}(1+118, 1+11)$$

$$= \text{Beta}(119, 12).$$



Posterior mean $= \frac{119}{131} \approx 0.908$

Sample mean $= \frac{118}{129} \approx 0.915$

Posterior stdev $= \sqrt{\frac{119(12)}{132(131)^2}} \approx 0.025$

$$\frac{ab}{(a+b+1)(a+b)^2}$$

Beta(a,b) variance

Part II : Gamma-Poisson

Poisson distribution

- State space $\mathcal{Y} = \{0\} \cup \mathbb{N} = \{0, 1, 2, \ldots\}$
- $\mathcal{Y}$-valued r.v. $Y \sim \text{Poisson}(\theta)$ if

$$p(y|\theta) = \mathbb{P}(Y = y|\theta) = \frac{\theta^y}{y!\, e^\theta}.$$

- Poisson(θ) describes the probability of observing a certain number of events

in a fixed interval (e.g. time or space) when events occur independently and the process has constant average rate θ.

- Examples
  - # of emails recieved / day
  - # of cars passing through intersection / hour
  - # of typos / page of text

- Binomial: Fixed # of trials
  Poisson: Fixed interval

- If $Y \sim \text{Poisson}(\theta)$, then
  - $\mathbb{E}[Y|\theta] = \text{Var}[Y|\theta] = \theta$.

Gamma distribution

- State space $\Theta = [0, \infty)$.
- Θ-valued r.v. $\theta \sim \text{Gamma}(a, b)$ if

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \propto \theta^{a-1} e^{-b\theta}$$

for $a, b \in (0, \infty)$.

- If $\theta \sim \text{Gamma}(a, b)$, then
  - $\mathbb{E}[\theta] = \frac{a}{b}$.
  - $\text{Var}[\theta] = \frac{a}{b^2}$.

Conjugacy of gamma-Poisson

- $p(\theta) \sim \text{Gamma}(a, b)$
- $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$

$$\Rightarrow p(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta}$$

$$= e^{-n\theta} \frac{\theta^{\sum y_i}}{\prod y_i!}$$

- Compute posterior:

$$p(\theta \mid y_1, \dots, y_n) \propto p(y_1, \dots, y_n \mid \theta)\, p(\theta)$$

$$\propto e^{-n\theta}\, \theta^{\sum y_i}\, \theta^{a-1}\, e^{-b\theta}$$

$$= e^{-(n+b)\theta}\, \theta^{a+\sum y_i - 1}$$

$$\Rightarrow p(\theta \mid y_1, \dots, y_n) = \text{Gamma}\left(a + \sum_{i=1}^{n} y_i,\ b+n\right).$$

- Posterior mean:

$$\mathbb{E}[\theta|y] = \frac{a + \sum_{i}^{n} y_i}{b + n}$$

$$= \frac{a}{b+n} \cdot \frac{b}{b} + \frac{\sum y_i}{b+n} \cdot \frac{n}{n}$$

$$= \left(\frac{b}{b+n}\right) \frac{a}{b} + \left(\frac{n}{b+n}\right) \frac{\sum y_i}{n}$$

Prior mean

Sample mean

- n = # of observations (e.g. # of 1 hour intervals)
- $\sum_{i=1}^{n} y_i$ = sum of counts from each observation
- $\frac{1}{n} \sum_{i=1}^{n} y_i$ = average # of counts

Analogous interpretation for prior:

- b = "prior # of observations"
- a = "sum of counts from prior observations"
- $\frac{a}{b}$ = "prior average # of counts"

- Beta-binomial and Gamma-Poisson are special cases of exponential family models. See Section 3.3 of Hoff for details.

Part III : Posterior inference

Prediction

- $(Y_1, \ldots, Y_n)$ observations of

$$Y_1, \ldots, Y_n \mid \theta \overset{\text{iid}}{\sim} p(\cdot \mid y).$$

- We want to make predictions about unobserved $\tilde{Y}$ based on observed data.

- Given observations $y = (y_1, \dots, y_n)$ of $Y_1, \dots, Y_n$, the <u>posterior predictive distribution</u> of an unobserved r.v. $\tilde{Y}$ is

$$p(\tilde{y} \mid y) = \int_{\Theta} p(\tilde{y}, \theta \mid y)\, d\theta$$

$$= \int_{\Theta} p(\tilde{y} \mid \theta, y)\, p(\theta \mid y)\, d\theta.$$

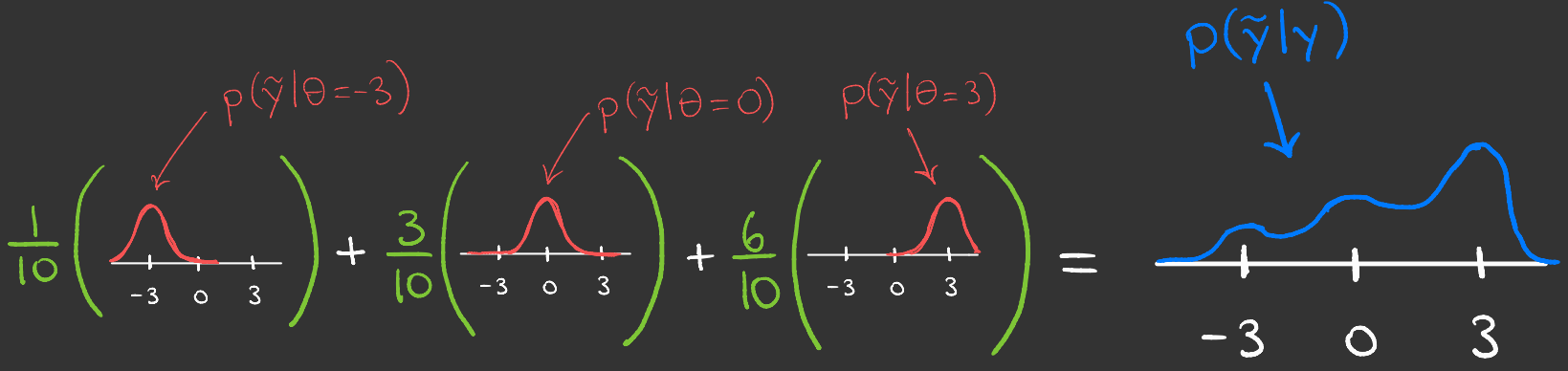
- If $\tilde{Y}$ is conditionally independent of $Y_1, \ldots, Y_n$ given θ, then

$$p(\tilde{y} \mid y) = \int_{\Theta} \underbrace{p(\tilde{y} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta \mid y)}_{\text{posterior}} d\theta.$$

- <u>Illustration.</u> Suppose $\Theta = \{-3, 0, 3\}$ and $Y_i \overset{iid}{\sim} \underbrace{N(\theta, 1)}_{\text{Normal distribution w/ mean } \theta, \text{ variance } 1}$.

Suppose we find that

- $p(-3|y) = \frac{1}{10}$
- $p(0|y) = \frac{3}{10}$
- $p(3|y) = \frac{6}{10}$



- Example. Suppose $Y_1, \ldots, Y_n$ are observations of $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. We put a $\text{Beta}(a,b)$ prior on θ, so

$$p(\theta | y_1, \ldots, y_n) = \text{Beta}\left(a + \Sigma y_i, b + n - \Sigma y_i\right).$$

The posterior predictive probability of success for an unobserved $\tilde{Y}$ that is

conditionally independent of $Y_1, \dots, Y_n$ given θ (and also Bernoulli(θ)) is then

$$\mathbb{P}(\tilde{Y}=1 \mid y_1, \dots, y_n) = p(1 \mid y_1, \dots, y_n)$$

$$= \int_0^1 p(1 \mid \theta)\, p(\theta \mid y_1, \dots, y_n)\, d\theta$$

$$= \int_0^1 \theta\, p(\theta \mid y_1, \dots, y_n)\, d\theta$$

$$= \mathbb{E}[\theta \mid Y_1, \ldots, Y_n]$$

$$= \frac{a + \Sigma Y_i}{a + b + n} .$$

Confidence regions

- Goal: Uncertainty quantification

- How confident are we in θ
  - Before we observe Y?
  - After we observe Y?

- A random interval $[\ell(Y), u(Y)]$ has <u>frequentist coverage</u> $1-\alpha \in [0,1]$ if

$$\mathbb{P}(\ell(Y) \leq \theta \leq u(Y) \mid \theta) = 1-\alpha.$$

- An interval $[\ell(y), u(y)]$ (based on observed data y) has <u>Bayesian coverage</u> $1-\alpha$ if

$$\mathbb{P}(\ell(y) \leq \theta \leq u(y) \mid Y=y) = 1-\alpha$$

- Suppose $\alpha = 0.95$.
  - Frequentist: 95% of the time, interval will contain θ. However, for observed data y, the interval either contains θ or does not — we have no way to know.
  - Bayesian: 95% chance that θ lies in the interval $[l(y), u(y)]$ constructed from observed data y.

- In many cases, confidence regions/intervals have the same Bayesian and frequentist coverage asymptotically, i.e., as $n \longrightarrow \infty$

- Bayesian confidence regions/intervals often called <u>credible regions/intervals</u>.

Computing confidence intervals

- Quantile-based intervals: To make a $100(1-\alpha)\%$ confidence interval, find $\theta_{\frac{\alpha}{2}}$ and $\theta_{1-\frac{\alpha}{2}}$ such that
  - $\theta_{\frac{\alpha}{2}} < \theta_{1-\frac{\alpha}{2}}$
  - $\mathbb{P}(\theta < \theta_{\frac{\alpha}{2}} \mid y) = \frac{\alpha}{2}$
  - $\mathbb{P}(\theta > \theta_{1-\frac{\alpha}{2}} \mid y) = \frac{\alpha}{2}$

$$\Rightarrow \mathbb{P}\left(\theta \in \left[\theta_{\frac{\alpha}{2}}, \theta_{1-\frac{\alpha}{2}}\right] \mid y\right)$$

$$= 1 - \mathbb{P}\left(\theta < \theta_{\frac{\alpha}{2}} \text{ or } \theta > \theta_{1-\frac{\alpha}{2}} \mid y\right)$$

$$= 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right)$$

$$= 1 - \alpha .$$

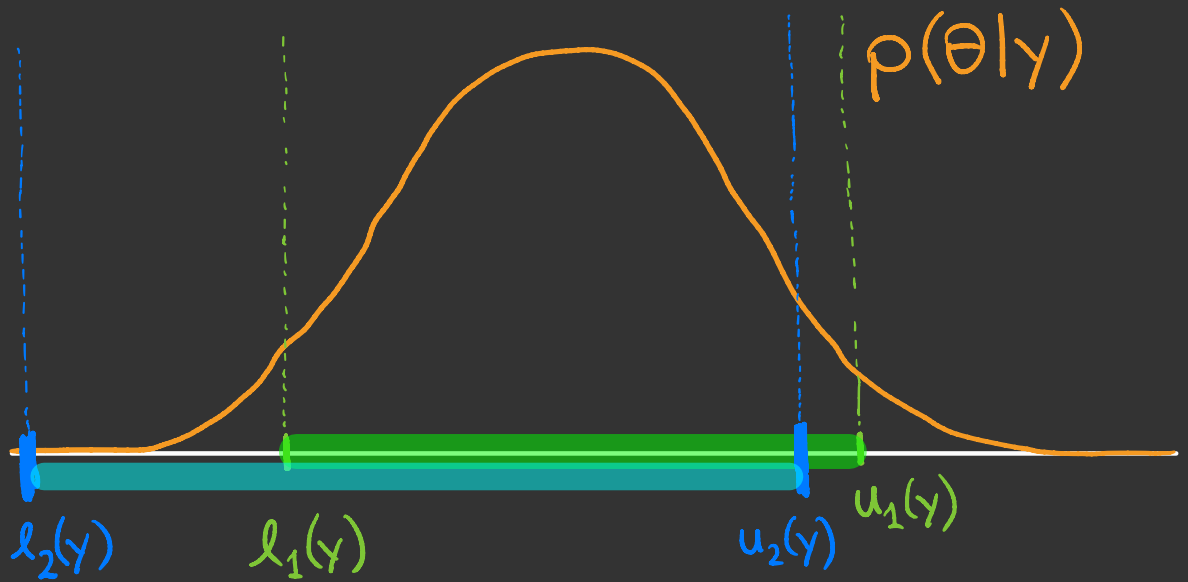
For standard distributions, quantities like

$$\mathbb{P}\left(\theta < \theta_{\frac{\alpha}{2}} \mid y\right)$$

can be computed directly. In the next chapter, we'll see how to compute this in general.

Highest posterior density (HPD) region

- In general, a Bayesian confidence interval is not unique.



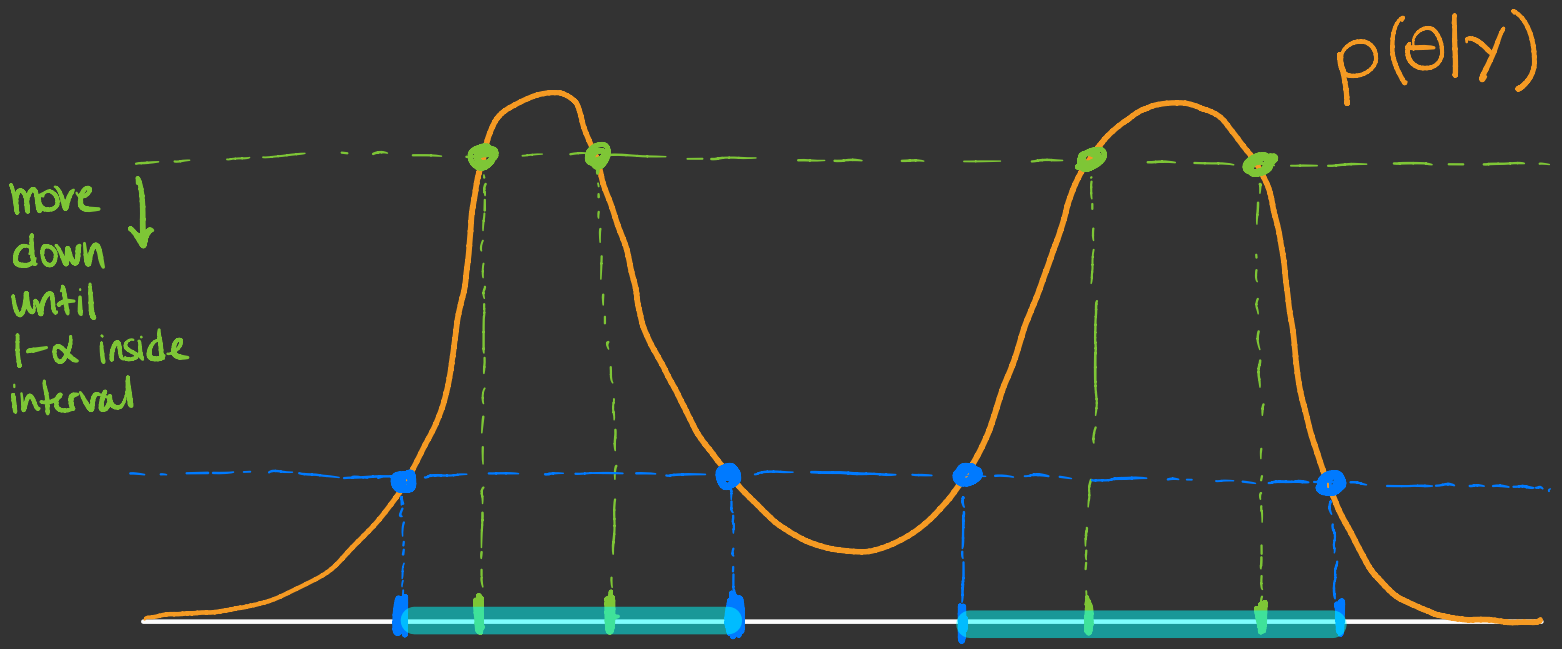
$$\mathbb{P}(l_1(y) \leq \theta \leq u_1(y) \mid Y = y) = \alpha$$

AND

$$\mathbb{P}(l_2(y) \leq \theta \leq u_2(y) \mid Y = y) = \alpha.$$

- A $100(1-\alpha)\%$ HPD region is a subset $S(y) \subseteq \Theta$ of the parameter space such that
  - (i) $\mathbb{P}(\theta \in S(y) \mid Y = y) = 1 - \alpha$
  - (ii) If $\theta_1 \in S(y)$ and $\theta_2 \notin S(y)$, then

$$p(\theta_1|y) > p(\theta_2|y).$$



End Ch. 3