Section 2

Probability

Review of probability

- The crux of Bayesian statistics is to compute the posterior distribution, i.e., the uncertainty distribution of the parameters (θ) after observing the data (**Y**)

- This is the conditional distribution of θ given **Y**

- Therefore, we need to review the probability concepts that lead to the conditional distribution of one variable conditioned on another

- We will cover Chapters 1.1-1.2:
  1. Probability mass (PMF) and density (PDF) functions
  2. Joint distributions
  3. Marginal and conditional distributions
  4. Bayes Rule

Random variables

- X (capital) is a random variable
- We want to compute the probability that X takes on a specific value x (lowercase)
- This is denoted $\text{Prob}(X = x)$
- We also might want to compute the probability of X being in a set $\mathcal{A}$
- This is denoted $\text{Prob}(X \in \mathcal{A})$
- The set of possible value that X can take on is called its support, $\mathcal{S}$

Random variables - example

Example 1: X is the roll of a die

- The support is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
- $\text{Prob}(X = 1) = 1/6$

Example 2: X is a newborn baby's weight

- The support is $\mathcal{S} = (0, \infty)$
- $\text{Prob}(X \in [0, \infty]) = 1$

What is probability?

Objective (associated with frequentist)

- $\text{Prob}(X = x)$ as a purely mathematical statement
- If we repeatedly sampled X, the value that the proportion of draws equal to x converges to defined as $\text{Prob}(X = x)$

Subjective (associated with Bayesian)

- $\text{Prob}(X = x)$ represents an individual's degree of belief
- Often quantified as the amount an individual would be willing to wager that X will be x

A Bayesian analysis makes use of both of these concepts

What is uncertainty?

Aleatoric uncertainty (likelihood)

- Uncontrollable randomness in the experiment
- For example, the results of a fair coin flip can never be predicted with certainty

Epistemic uncertainty (prior/posterior)

- Uncertainty about a quantity that could theoretically be known
- For example, if we flipped a coin infinitely-many times we could know the true probability of a head

A Bayesian analysis makes use of both of these concepts

Probability versus statistics

Probability is the forward problem

- We assume we know how the data are being generated and compute the probability of events
- For example, what is the probability of flipping 5 straight heads if the coin is fair?

Statistics is the inverse problem

- We use data to learn about the data-generating mechanism
- For example, if we flipped five straight head, can we conclude the coin is biased?

Any statistical analysis obviously relies on probability

Univariate distributions

- We often distinguish between **discrete** and **continuous** random variables
- The random variable X is discrete if its support $\mathcal{S}$ is countable
- Examples:

  $X \in \{0, 1, 2, 3\}$ is the number of successes in 3 trials

  $X \in \{0, 1, 2, ...\}$ is the number users that visit a website

Univariate distributions

- We often distinguish between **discrete** and **continuous** random variables
- The random variable X is continuous if its support $\mathcal{S}$ is uncountable
- Examples with $\mathcal{S} = (0, \infty)$:

  $X > 0$ is weight of a baby

  $X > 0$ is the wind speed

Discrete univariate distributions

- If X is discrete we describe its distribution with its **probability mass function** (PMF)
- The PMF is $f(x) = \text{Prob}(X = x)$
- The domain of X is the set of x with $f(x) > 0$
- We must have $f(x) \geq 0$ and $\sum_x f(x) = 1$
- The mean is $\text{E}(X) = \sum_x xf(x)$
- The variance is $\text{V}(X) = \sum_x [x - E(X)]^2 f(x)$
- The last three sums are over X's domain

Parametric families of distributions

- A statistical analysis typically proceeds by selecting a PMF that seems to match the distribution of a sample
- We rarely know the PMF exactly, but we assume it is from a parametric family of distributions
- For example, Binomial(10,0.5) and Binomial(4,0.1) are different but both from the normal family
- A family of distributions have the same equation for the PMF but differ by some unknown parameters θ
- We must estimate these parameters

Example: $X \sim \text{Bernoulli}(\theta)$

- Example: X is a success (1) or failure (0)
- Domain: $X \in \{0, 1\}$ (i.e., X is binary)
- PMF: $P(X = 0) = 1 - \theta$ and $P(X = 1) = \theta$
- Parameter: $\theta \in [0, 1]$ is the success probability
- Mean: $\text{E}(X) = \sum_x xf(x) = 0(1 - \theta) + 1\theta = \theta$
- Variance:

$$\text{V}(X) = \sum_x (x-\theta)^2 f(x) = (0-\theta)^2(1-\theta)+(1-\theta)^2\theta = \theta(1-\theta)$$

Example: $X \sim \text{Binomial}(N, \theta)$

- Example: X is a number of successes in N trials
- Domain: $X \in \{0, 1, ..., N\}$
- PMF: $P(X = x) = \binom{N}{x}\theta^x(1 - \theta)^{N-x}$
- Parameter: $\theta \in [0, 1]$ is the success probability of each trial
- Mean: $\text{E}(X) = \sum_{x=0}^{N} xf(x) = n\theta$
- Variance: $\text{V}(X) = n\theta(1 - \theta)$

Example: $X \sim \text{Poisson}(N\theta)$

- Example: X is the number events that occur in N units of time
- Often the distribution is presented with $N = 1$
- Domain: $X \in \{0, 1, 2, ...\}$
- PMF: $P(X = x) = \frac{\exp(-N\theta)(N\theta)^x}{x!}$
- Parameter: θ is the expected number of events per unit of time
- Mean: $\text{E}(X) = N\theta$
- Variance: $\text{V}(X) = N\theta$

Continuous univariate distributions

- If X is continuous we describe its distribution with the probability density function (PDF) $f(x) \geq 0$
- Since there are uncountably many possible values, the probability of any one value must be zero
- For example,

$P(X=x) = c > 0$
for all $x \in (7.01, 7.02)$

$P(X \in (7.01, 7.02))$
$= c \cdot$ # elements in interval
$= \infty$

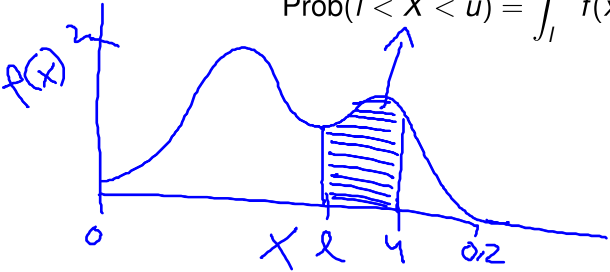
$P(\geq 0.5) + P(\leq 0.5) = 1$
$P = 0.5 = 0$

- Therefore, $P(X = x) = 0$ for all x, and so the PMF is meaningless

Continuous univariate distributions

- Probabilities are computed as areas under the PDF curve

$$\text{Prob}(l < X < u) = \int_l^u f(x)dx$$



- Therefore, $f(x)$ must satisfy $f(x) \geq 0$ and

$$\text{Prob}(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

Continuous univariate distributions

- The domain is the set of x values with $f(x) > 0$
- The mean and the variance are defined similarly to the discrete case but with the sums replaced by integrals
- The mean is
$$\mathrm{E}(X) = \int xf(x)dx$$
- The variance is
$$\mathrm{V}(X) = \int [x - E(X)]^2 f(x)dx$$

Example: $X \sim \text{Normal}(\mu, \sigma^2)$

- Example: X is an IQ score
- Domain: $X \in (-\infty, \infty)$
- PDF: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
- Parameters: μ is the mean, $\sigma^2 > 0$ is the variance
- Mean: $\text{E}(X) = \mu$
- Variance: $\text{V}(X) = \sigma^2$

Example: $X \sim \text{Gamma}(a, b)$

- Example: X is a height
- Domain: $X \in (0, \infty)$
- PDF: $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$
- Parameters: $a > 0$ is the shape, $b > 0$ is the rate
- Mean: $\text{E}(X) = \frac{a}{b}$
- Variance: $\text{V}(X) = \frac{a}{b^2}$
- Be careful: Sometimes the PDF is given as

$$f(x) = \frac{1}{\Gamma(a) b^a} x^{a-1} \exp(-x/b)$$

Example: $X \sim \text{InverseGamma}(a, b)$

- If $Y \sim \text{Gamma}(a, b)$ and $X = 1/Y$, then $X \sim \text{InverseGamma}(a, b)$
- Domain: $X \in (0, \infty)$
- PDF: $f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-b/x)$
- Parameters: $a > 0$ is the shape, $b > 0$ is the rate
- Mean: $\text{E}(X) = \frac{b}{a-1}$ if $a > 1$
- Variance: $\text{V}(X) = \frac{b^2}{(a-1)^2(a-2)}$ if $a > 2$
- Be careful: Sometimes the PDF is given as

$$f(x) =\propto x^{-a-1} \exp[-1/(bx)]$$

Example: $X \sim \text{Beta}(a, b)$

- Example: X is a probability
- Domain: $X \in [0, 1]$
- PDF: $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$
- Parameters: $a > 0$ and $b > 0$
- Mean: $\text{E}(X) = \frac{a}{a+b}$
- Variance: $\text{V}(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Joint distributions

- $\mathbf{X} = (X_1, ..., X_p)$ is a random vector (vectors and matrices should be in bold).

- For notational convenience, let's consider only $p = 2$ random variables X and Y.

- (X, Y) is discrete if it can take on a countable number of values, such as

  X = number of hearts and Y = number of face cards.

- (X, Y) is continuous if it can take on an uncountable number of values, such as

  X = birthweight and Y = gestational age.

Discrete random variables

- The joint PMF is

$$f(x, y) = \text{Prob}(X = x, Y = y)$$

- Example: patients are randomly assigned a dose and followed to determine whether they develop a tumor.
- $X \in \{5, 10, 20\}$ is the dose; $Y \in \{0, 1\}$ is 1 if a tumor develops and 0 otherwise

$X=5$ $\quad P(Y=1 \mid X=5) = \frac{0.231}{0.700} = \frac{1}{3}$ $\quad f_Y(0) = 0.642$

- The joint PMF is

$\text{Prob}(X=5) = 0.7 = f_X(5)$

	X		
Y	5	10	20
0	0.469	0.124	0.049
1	0.231	0.076	0.051

$0.469 + 0.124 + 0.049 = 0.642$

$0.358 = f_Y(1)$

$0.469 + 0.231 = 0.700$ $\quad 0.200$ $\quad 0.100$

Discrete random variables

$P(Y=0 \mid X=5) = \frac{0.469}{0.700} = \frac{2}{3}$

- The **marginal PMF** for X is

$$f_X(x) = \text{Prob}(X = x) = \sum_y f(x, y)$$

- The **marginal PMF** for Y is

$$f_Y(y) = \text{Prob}(Y = y) = \sum_x f(x, y)$$

- The marginal distribution is the same as univariate distribution as if we ignored the other variable

Discrete random variables

- Example: X = dose and Y = tumor status
- Find the marginal PMFs of X and Y
- See "dose marginal" in the online derivations

Discrete random variables

values in table

- The **Conditional PMF** of Y given X is

$$f(y|x) = \text{Prob}(Y = y|X = x) = \frac{\text{Prob}(X = x, Y = y)}{\text{Prob}(X = x)} = \frac{f(x,y)}{f_X(x)}.$$

row sums

- Here x is treated as a fixed number, and so $f(y,x)$ is only a function of y.
- However, we can't use $f(x,y)$ as the PMF for Y because

$$\sum_y f(x,y) = f_X(x) \neq 1$$

- Dividing by $f_X(x)$ makes $f(y|x)$ valid

$$\sum_y f(y|x) = \sum_y \frac{f(y,x)}{f_X(x)} = \frac{\sum_y f(y,x)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$$

Discrete random variables

- Example: X = dose and Y = tumor status

- Find the $f(x|y)$ and $f(y|x)$

$$f(x=5|y=1) = \frac{f(5,1)}{f_Y(1)} = \frac{0.231}{0.358}$$

$$= 0.645$$

- See "dose conditional" in the online derivations

Monte Hall problem

- `http://en.wikipedia.org/wiki/Monty_Hall_problem`
- Suppose you're on a game show, and you're given the choice of three doors.
- Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- **Is it to your advantage to switch your choice?**

Discrete random variables

- X and Y are **independent** if

$$f(x, y) = f_X(x) f_Y(y)$$

  for all x and y

- Variables are dependent if they are not independent
- Equivalently, X and Y are **independent** if

$$f(x|y) = f_X(x)$$

  for all x and y

- Prove these two definitions are equivalent

Discrete random variables

- Notation: $X_1, ..., X_n \overset{iid}{\sim} f(x)$ means that $X_1, ..., X_n$ are independent and identically distributed

- This implies the joint PMF is

$$\text{Prob}(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} f(x_i)$$

- The same notation and definitions of independence apply to continuous random variables

Hurricane proportions by landfall (X) and category (Y)

$Y =$

$X =$	Category 1	2	3	4	5	Total
US	0.0972	0.0903	0.0694	0.0069	0.0069	0.2708
Not US	0.3194	0.1319	0.1389	0.1181	0.0208	0.7292
Total	0.4167	0.2222	0.2083	0.1250	0.0278	1.0000

Problem: Prove X and Y are dependent

Show $f(y|x) \neq f_Y(y)$ for some x & y

$f_Y(5) = 0.0278$ $\quad P(5|US) = \frac{0.0069}{0.2708} \neq 0.0278$

Continuous random variables

- Manipulating joint PDFs is similar to joint PMFs but sums are replaced by integrals
- The joint PDF is denoted $f(x, y)$
- Probabilities are computed as volume under the PDF:

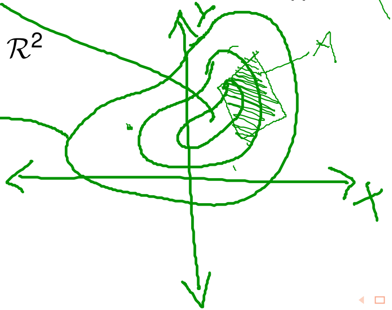
high $f(x,y)$

$$\text{Prob}[(X, Y) \in A] = \int_A f(x, y)dxdy$$

volume under the surface

where $A \subset \mathcal{R}^2$

$f(x,y)$ = small

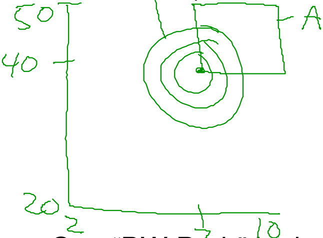


Continuous random variables

- Example: X=birthweight, Y=gestational age
- Domain: $X \in (2, 10)$ lbs and $Y \in (20, 50)$ weeks
- PDF: $f(x, y) = 0.26 \exp(-|x - 7| - |y - 40|)$
- Find: Prob$(X > 7, Y > 40)$ $= \int\int_A f(x,y)\,dx\,dy$

$$= \int_{40}^{50} \int_{7}^{10} 0.26\, e^{-|x-7|-|y-40|}\, dx\, dy$$

$$= \dots = 0.25$$



- See “BW Prob” in the online derivations

Continuous random variables

If (X,Y,Z) are 3 variables,
$f_X(x)=\iint f(x,y,z)\,dy\,dz$

- The **Marginal PDF** of X is

$$f_X(x) = \int f(x,y)dy$$

- f_X is the univariate PDF for X as if we never considered Y
- Find: $f_X(x)$ for the birthweight example

$f_X(x_0)=\int f(x_0,y)dy$

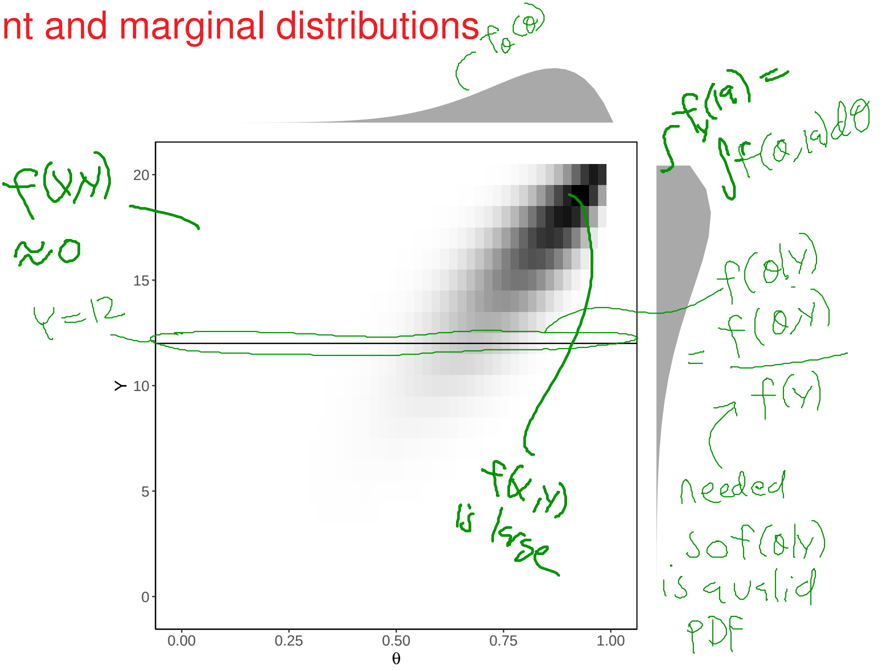
$f_X(x)=\int f(x,y)dy$

$= \int_{20}^{50} 0.26 e^{-|x-7|-|y-40|} dy$

$= \ldots 0.52 e^{-|x-7|}$

- See "BW marginal" in the online derivations

Joint and marginal distributions



Continuous random variables

- The **Conditional PDF** of Y given X is

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

- Proper: $\int f(y|x)dy = \int \frac{f(x,y)}{f_X(x)}dy = \frac{\int f(x,y)dy}{f_X(x)} = 1$

- Find: $f(y|x)$ for the birthweight example

- See “BW conditional” in the online derivations

Bivariate normal distribution

- The **bivariate normal distribution** is the most common multivariate family
- There are 5 parameters:
  - The marginal means of X and Y are μ_X and μ_Y
  - The marginal variances of X and Y are $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$
  - The correlation between X and Y is $\rho \in (-1, 1)$
- The joint PDF is $f(x, y) =$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)}{2(1-\rho^2)}\right\}$$

-

Bivariate normal distribution

- Assume $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$, find the marginal distribution of X.

In general, if $\mu_x \neq 0$ & $\sigma_x \neq 1$

$$X \sim N(\mu_x, \sigma_x^2)$$

&

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

- See "MVN marginal" in the online derivations

Bivariate normal distribution

- Assume $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$, find the conditional distribution of Y given X.

So $Y|X \sim N(\underbrace{\qquad\qquad}_{\text{function of } \rho \,\&\, x}, \underbrace{\qquad\qquad}_{\text{function of } \rho})$

- See "MVN conditional" in the online derivations

Defining joint distributions conditionally

$$f(y|x) = \frac{p(x,y)}{f(x)}$$

- Specifying joint distributions is hard
- Every joint distribution can be written

$$f(x, y) = f(y|x)f(x)$$

- Therefore, any joint distribution can be defined by
  1. X's marginal distribution
  2. The conditional distribution of $Y|X$

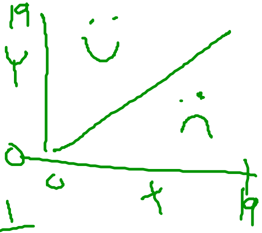
ex

$Y|X \sim \text{Beta}(X, 3+X^2)$

- The joint problem reduces to two univariate problems
- This idea forms the basis of hierarchical modeling

Defining joint distributions conditionally

- Let Y be the number of robins in the forest
- Let X be the number of robins we observe $\in \{0, 1, \dots, Y\}$
- Model $\text{Prob}(Y = y) = 1/20$ for $y \in \{0, \dots, 19\}$ and $X|Y \sim \text{Binomial}(Y, 0.2)$
- What is the support of (X, Y)?



- What is $\text{Prob}(X = 1, Y = 10)$?

$= P(X|Y)P(Y) = \left[\binom{10}{1} .2^1 .8^9\right]\frac{1}{20}$

- What is $\text{Prob}(X = 0)$?

$P(X) = \sum P(X,Y) = \sum P(X|Y)P(Y)$

$P(0) = \sum_Y P(0|Y)\frac{1}{20} = \sum\left[\binom{Y}{0} .2^0 .8^Y\right]\frac{1}{20} = \left[\sum_{Y=0}^{20} .8^Y\right]\frac{1}{20}$

Bayes' theorem

- In Bayesian statistics, we select the prior, $p(\theta)$, and the likelihood, $p(y|\theta)$
- Based on these two pieces of information, we must compute the posterior $p(\theta|y)$
- Bayes' theorem is the mathematical formula to convert the likelihood and prior to the posterior
- Bayes theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- This holds for discrete (PMF) and continuous (PDF) cases

Bayes' theorem

- Bayes theorem in math:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Bayes theorem in words:

$$p(\theta|y) = \frac{\text{Likelihood} * \text{Prior}}{\text{marginal distribution of Y}}$$

- As in the formula for a conditional distribution, $p(y)$ is just the normalizing constant required so that $\int p(\theta|y)d\theta = 1$

- Most of the time $p(y)$ can be ignored because it doesn't depend on θ and the objective is to study the posterior of θ

Derivation of Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \boxed{\frac{p(y|\theta)\,p(\theta)}{p(y)}}$$

$$p(\theta, y) = p(\theta|y)\,p(y)$$

or

$$= p(y|\theta)\,p(\theta)$$

Football example

$p(\tilde{w} \mid H=0) = 0.4$ $p(H=0) = p(H=1) = \frac{1}{2}$ $p(\tilde{w} \mid H=1) = 0.7$

- A team plays half its games at home, wins 70% of its home games, and 40% of its road games. Given that the team wins a game, what's the probability it was a home game?

$$W = \begin{cases} 1 & \text{win} \\ 0 & \text{lose} \end{cases} \qquad H = \begin{cases} 1 & \text{home} \\ 0 & \text{road} \end{cases}$$

$$p(H=1 \mid W=1) = p(W=1 \mid H=1)\,p(H=1) / p(W=1)$$

$$= 0.7 \cdot 0.5 / p(W=1)$$

$$= 0.7 \cdot .5 / 0.55 = \frac{7}{11} > 0.5$$

$$p(W=1) = p(\text{winning}) = \sum_H p(W=1, H) = \sum p(W=1 \mid H)\,p(H)$$

$$= p(W=1 \mid H=0)\,p(H=0) + p(W=1 \mid H=1)\,p(H=1)$$

$$= .4 \cdot .5 + .7 \cdot .5 = \boxed{0.55}$$

- See “Football” in the online derivations

HIV example

- Let θ be the parameter of interest with

$$\theta = \begin{cases} 0 & \text{patient does not have HIV} \\ 1 & \text{patient has HIV} \end{cases}$$

- The data is Y, defined as

$$Y = \begin{cases} 0 & \text{test is negative} \\ 1 & \text{test is positive} \end{cases}$$

- Objective: Derive the probability that the patient has HIV given the test results

- That is, we want $p(\theta|y)$

HIV example - Likelihood

- The likelihood describes the distribution of the data as if we knew the parameters
- This is a statistical model for the data
- Since Y is binary, we use a Bernoulli PMF for the likelihood
- We must specify the likelihood for both $\theta = 0$ and $\theta = 1$
- $\text{Prob}(Y = 1|\theta = 0) = q_0$ is the false positive rate
- $\text{Prob}(Y = 1|\theta = 1) = q_1$ is the true positive rate
- How might we select q_0 and q_1?

HIV example - Prior

- The prior represents our uncertainty about the parameters before we observe the data
- Since θ is binary, we use a Bernoulli PMF for the prior
- $\text{Prob}(\theta = 1) = p$ is the population prevalence of HIV
- How might we select p?

HIV example - Posterior

- Derive the posterior probability that the patient has HIV given a positive test
- That is Prob($\theta = 1 | Y = 1$)
- See "HIV" in the online derivations

HIV example - Posterior

- Derive the posterior probability that the patient has HIV given a negative test
- That is Prob($\theta = 1 | Y = 0$)

- See "HIV" in the online derivations

Robins example

- Let Y be the number of robins in the forest
- Let X be the number of robins we observe
- Model $\text{Prob}(Y = y) = 1/20$ for $y \in \{0, ..., 19\}$ and $X|Y \sim \text{Binomial}(Y, 0.2)$
- Given that we do not observe any birds, what is the probability that no birds are in the forest?
- Intuitively, how would this change if Y could be as large as 100?
- Intuitively, how would this change if the detection probability increased from 0.2 to 0.9?