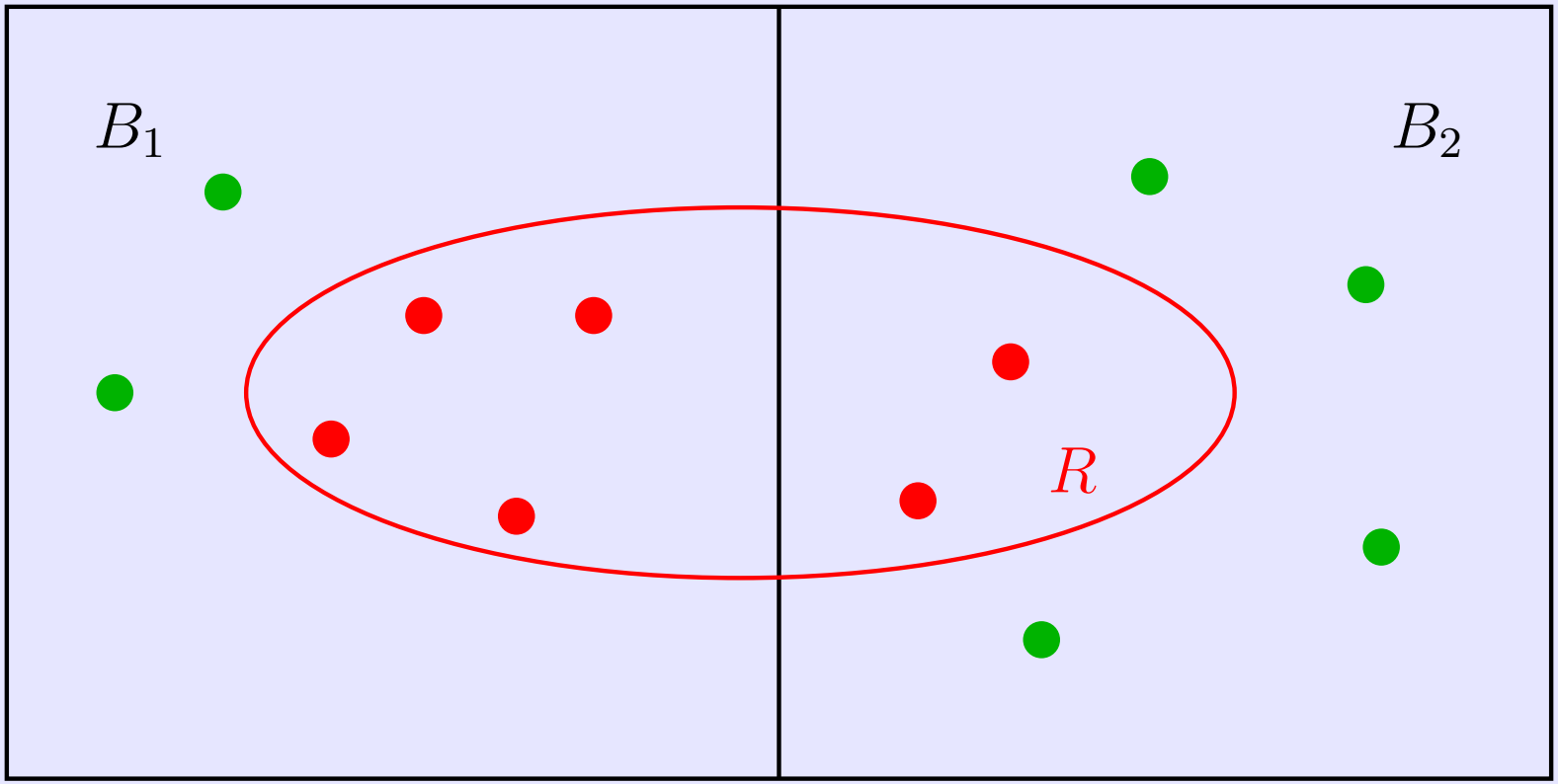
## Quiz

Two boxes are available:

- Box 1 contains 4 red and 2 green balls.
- Box 2 contains 2 red and 4 green balls.
- The probability of choosing either box is equal:

$$P(E_1) = P(E_2) = \frac{1}{2}.$$

- A box is selected at random, and one ball is drawn.
- Study the **likelihood** of observing a red ball under each box.


$B_1$
$B_2$
$R$

**Answer:** A red ball is more consistent with Box 1 than with Box 2.

The likelihoods are: $P(R \mid B_1) = \frac{4}{6} = \frac{2}{3}$ and $P(R \mid B_2) = \frac{2}{6} = \frac{1}{3}$.

So, if Box 1 had been selected, observing a red ball would be more likely.

$$P(R \mid B_1) > P(R \mid B_2)$$

Therefore, the red ball provides stronger evidence for Box 1 than for Box 2.

$$\frac{P(R \mid B_1)}{P(R \mid B_2)} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

The observed red ball is **twice** as likely under Box 1 as under Box 2. Since the boxes were equally likely before the draw, the red ball favors Box 1.

# Homework

- Questions 1–5 next; questions 6-9 at the end of the topic.
- Submit all answers by Fri Sep 11.
- You may discuss with peers but submissions must be individual.
- Responses may be shared anonymously during class discussion!

# Question 1

For each scenario below, you're given a pair of events, $A$ and $B$. Explain what you believe to be the relationship between the posterior and prior probabilities of $B$: $P(B|A) > P(B)$ or $P(B|A) < P(B)$.:

a) $A =$ you just finished reading Lambda Literary Award-winning author Nicole Dennis-Benn's first novel, and you enjoyed it!
$B =$ you will also enjoy Benn's newest novel.

b) $A =$ it's 0 degrees Fahrenheit in Minnesota on a January day.
$B =$ it will be 60 degrees tomorrow.

c) $A =$ the authors only got 3 hours of sleep last night.
$B =$ the authors make several typos in their writing today.

d) $A =$ your friend includes three hashtags in their tweet.
$B =$ the tweet gets retweeted.

## Question 2

The probability that Gabby will like a restaurant is 0.7. Among the restaurants that she likes, 20% have five stars on Yelp, 50% have four stars, and 30% have fewer than four stars. What other information do we need if we want to find the posterior probability that Gabby likes a restaurant given that it has fewer than four stars on Yelp?

## Question 3

Matt is on a dating app looking for love. Matt swipes right on 8% of the profiles he views. Of the people that Matt swipes right on, 40% are men, 30% are women, 20% are non-binary, and 10% identify in another way. Of the people that Matt does not swipe right on, 45% are men, 40% are women, 10% are non-binary, and 5% identify in some other way.

a) What's the probability that a randomly chosen person on this dating app is non-binary?

b) Given that Matt is looking at the profile of someone who is non-binary, what's the posterior probability that he swipes right?

## Question 4

For a certain airline, 30% of the flights depart in the morning, 30% depart in the afternoon, and 40% depart in the evening.
Frustratingly, 15% of all flights are delayed.
Of the delayed flights, 40% are morning flights, 50% are afternoon flights, and 10% are evening flights.
Alicia and Mine are taking separate flights to attend a conference.

a) Mine is on a morning flight. What's the probability that her flight will be delayed?

b) Alicia's flight is not delayed. What's the probability that she's on a morning flight?

## Question 5 (answer revealed on the next slide!)

A medical test correctly detects a disease 95% of the time when the disease is present. This value represents:

a. Prior probability
b. Posterior probability
c. False positive rate
d. Likelihood

**Answer:** d. Likelihood

The statement describes: $P(\text{Test positive} \mid \text{Disease present}) = 0.95$

This is the probability of observing the test result given that the disease is present, which represents the likelihood (also called **sensitivity** or **true positive rate**).

- **Prior**: Probability of disease before observing the test result.
- **Posterior**: Probability of disease after observing the test result.
- **False positive rate**: Probability that the test is positive when the disease is absent.

# In-class Mini Exercise

A student applies to 6 internships. Let $\pi$ be their chance of getting any one offer.

| $\pi$ | 0.3 | 0.4 | 0.5 | Total |
|---|---|---|---|---|
| $f(\pi)$ | 0.25 | 0.60 | 0.15 | 1.00 |

They receive $y = 4$ offers.

1. Write the data model for $Y \mid \pi$.
2. Compute $L(\pi \mid y = 4)$ for each $\pi$.
3. Multiply by the prior and normalize.
4. Which $\pi$ has the highest posterior probability?

## 1. Data Model

Assuming the internship outcomes are conditionally independent,

$$Y \mid \pi \sim \text{Binomial}(6, \pi).$$

Thus,

$$P(Y = y \mid \pi) = \binom{6}{y} \pi^y (1 - \pi)^{6-y}.$$

For the observed value $y = 4$,

$$L(\pi \mid y = 4) = \binom{6}{4} \pi^4 (1 - \pi)^2.$$

## 2. Compute the Likelihood in R

Evaluate the likelihood at each possible value of $\pi$.

```
pi <- c(0.3, 0.4, 0.5)

likelihood <- dbinom(
x = 4,
size = 6,
prob = pi
)

likelihood
```

| $\pi$ | $0.3$ | $0.4$ | $0.5$ |
|---|---|---|---|
| $L(\pi \mid y = 4)$ | $0.059535$ | $0.138240$ | $0.234375$ |

Notice that the likelihood alone is largest at $\pi = 0.5$.

# 3. Multiply by the Prior and Normalize

Bayes' rule for a discrete prior is

$$f(\pi \mid y) = \frac{L(\pi \mid y) f(\pi)}{\sum_{\pi} L(\pi \mid y) f(\pi)}.$$

```
prior <- c(0.3, 0.4, 0.5)
weight <- likelihood * prior
posterior <- weight / sum(weight)
data.frame(pi, prior, likelihood, weight, posterior)
```

| $\pi$ | 0.3 | 0.4 | 0.5 |
|---|---|---|---|
| $L(\pi \mid y) f(\pi)$ | 0.014884 | 0.082944 | 0.035156 |
| $f(\pi \mid y)$ | 0.112 | 0.624 | 0.264 |

## 4. Posterior Conclusion

The posterior distribution is

| $\pi$ | $0.3$ | $0.4$ | $0.5$ |
|---|---|---|---|
| $f(\pi \mid y = 4)$ | $0.112$ | $0.624$ | $0.264$ |

In R, the most probable value is

```
pi[which.max(posterior)]
```

which gives $\pi = 0.4$.

Although $\pi = 0.5$ has the largest likelihood, the prior places much more probability on $\pi = 0.4$.

Therefore,

$$P(\pi = 0.4 \mid Y = 4) \approx 0.624$$

is the largest posterior probability.

## Question 6

Muhammad applies for six equally competitive data science internships. His prior model for his probability $\pi$ of receiving an offer from any given internship is

| $\pi$ | $0.3$ | $0.4$ | $0.5$ |
|---|---|---|---|
| $f(\pi)$ | $0.25$ | $0.60$ | $0.15$ |

a) Let $Y$ be the number of internship offers that Muhammad gets. Specify the model for the dependence of $Y$ on $\pi$ and the corresponding pmf, $f(y|\pi)$.

a) Muhammad got some pretty amazing news. He was offered four of the six internships! How likely would this be if $\pi = 0.3$?

c) Construct the posterior model of $\pi$ in light of Muhammad's internship news.

## Question 7

Whether you like it or not, cats have taken over the internet. Joining the craze, Michael has written an algorithm to detect cat images. It correctly identifies 80% of cat images as cats, but falsely identifies 50% of non-cat images as cats. Michael tests her algorithm with a new set of images, 8% of which are cats. What's the probability that an image is actually a cat if the algorithm identifies it as a cat? Answer this question by simulating data for 10,000 images.

## Question 8 (no spoilers, next slide has the answer!)

A university uses an AI system to detect plagiarism.

- 5% of assignments are actually plagiarized.
- The AI flags plagiarized assignments 92% of the time.
- The AI incorrectly flags honest assignments 8% of the time.

If an assignment is flagged, what is the probability it was actually plagiarized? Briefly discuss whether the AI system seems reliable.

**Answer:**

$$P(P \mid F) = \frac{0.92 \cdot 0.05}{0.92 \cdot 0.05 + 0.08 \cdot 0.95} \approx 0.377$$

Approximately 37.7% chance that it is actually plagiarism.
The system catches most plagiarism cases, but because plagiarism is rare, many flagged assignments are false positives.

# Question 9



Let's suppose that you are an air traveler along with 99 other passengers from your flight, each of whom, like you, checked one item of luggage.

A recording piped through the speakers reminds you that: *"Many bags look alike. Please check your bag carefully before exiting the terminal."*

Indeed, your bag is one of the most popular models on the market, a medium-sized black rectangular case **used by 5% of all travelers**.

Given your distance from the luggage chute, you can discern only the shape, size, and color of the bags that emerge, not any detailed markings on the bags.

## Question 9, cont.

a. Let's suppose that the **first bag** from your flight to enter the luggage carousel indeed has the same shape, size, and color as yours. Is it yours? Explain (i.e., calculate the posterior probability that the 1st bag is yours).

b. Now suppose that we continue to wait for our bag to appear, failing to see it among the first 98 bags to enter the carousel (i.e., only 2 bags are left to come off the airplane). Calculate the posterior probability that the **99th** bag, which also matches ours in shape, size, and color, is our own.

(maybe the answers will follow ...)

**Answer: a.** Let $Y$ = first bag is yours, and $M$ = first bag matches.

$$P(Y) = \frac{1}{100}$$
$$\mathcal{L}(Y^c \mid M) = P(M \mid Y^c) = 0.05$$
$$\mathcal{L}(Y \mid M) = P(M \mid Y) = 1$$

$$P(Y \mid M) = \frac{P(M \mid Y)P(Y)}{P(M)} = \frac{P(M \mid Y)P(Y)}{P(M \mid Y)P(Y) + P(M \mid Y^c)P(Y^c)} =$$
$$= \frac{1 \cdot 0.01}{1 \cdot 0.01 + 0.05 \cdot 0.99} = \frac{0.01}{0.0595} \approx 0.168$$

So even though the first bag matches, it is probably not yours.

**Answer: b.**

$$P(Y \mid M) = \frac{1 \cdot \frac{1}{\text{total of bags remaining}}}{1 \cdot \frac{1}{\text{total of bags remaining}} + 0.05 \cdot \frac{\text{total of bags remaining} - 1}{\text{total of bags remaining}}}$$

(if 2 bags remain)

$$= \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + 0.05 \cdot \frac{1}{2}} = \frac{1}{1 + 0.05} \approx 0.9524$$

There is a 95% chance that the 99th bag is yours, given that it looks like yours.

Or, it is **19 times more likely** to be yours than not yours $^{(0.95/0.05=19)}$.