# Frequentist and Bayesian Inference: Practice

Section 10 handout (answer key)

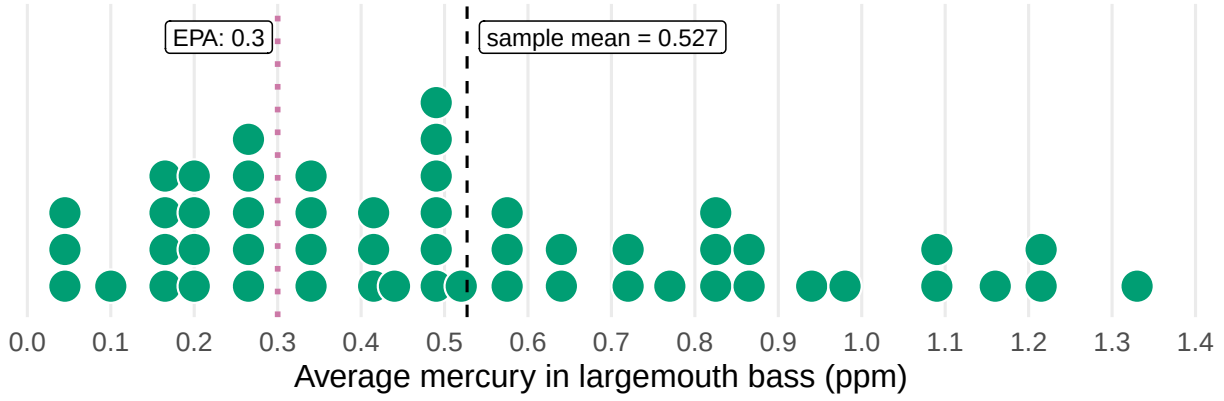
Name: SOLUTIONS

## Is the bass safe to eat?

Mercury builds up in fish, and people who eat a lot of fish with high mercury levels can be harmed. The U.S. Environmental Protection Agency (EPA) set **0.3 ppm** (parts per million, or mg of methylmercury per kg of fish) as the level of mercury in fish tissue that protects the health of people who eat fish.

Researchers caught **largemouth bass** from 53 lakes in Florida and measured the mercury in each fish. For each lake, we have the **average mercury level** (in ppm) of the bass from that lake.

Data: `FloridaLakes` in the `Lock5Data` package. Lange, T. R., Royals, H. E., & Connor, L. L. (1993). Influence of water chemistry on mercury concentration in largemouth bass from Florida lakes. *Transactions of the American Fisheries Society*, 122(1), 74–84. U.S. EPA (2001). *Water Quality Criterion for the Protection of Human Health: Methylmercury* (EPA-823-R-01-001).




Sample mean: $\bar{x} = 0.527$ ppm. Standard error: 0.047 (given to you).

Let $\mu$ be the mean of the average mercury levels in largemouth bass across **all** Florida lakes (like the ones in the study).

Is $\bar{x}$ a statistic or a parameter? **a statistic: it describes the sample of 53 lakes**

Is $\mu$ a statistic or a parameter? **a parameter: it describes all lakes**

Today we will analyze the same data in two ways: first the **frequentist** way (Week 9), then the **Bayesian** way (Week 10). Pay attention to **what is random** and **what each probability is about**.

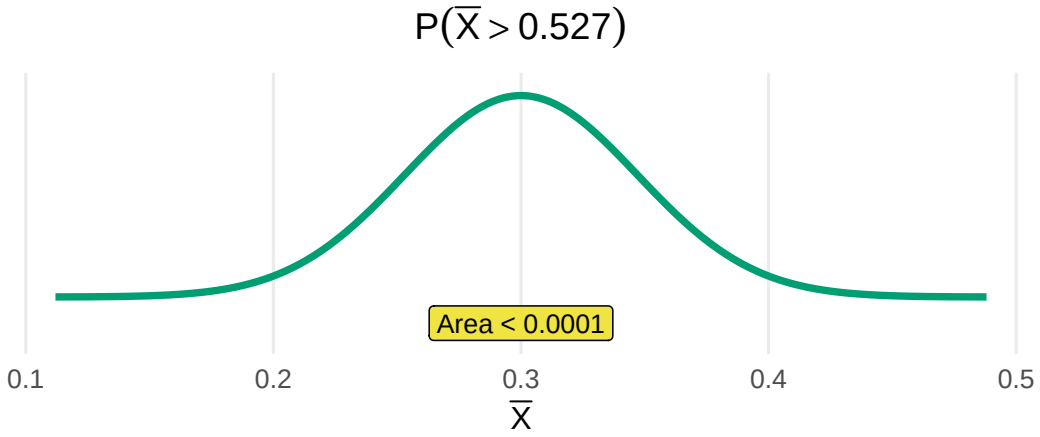
## Part 1: The frequentist analysis

**Hypotheses.**

- $H_0$: the mean mercury level is at the EPA level. $\mu = \mathbf{0.3}$
- $H_A$: the mean mercury level is not at the EPA level. $\mu \neq 0.3$

In frequentist inference, $\mu$ is a **fixed, unknown number**, and we start by **assuming $H_0$ is true**.

**Null sampling distribution.** If $\mu$ were 0.3, the sample mean of 53 lakes would be approximately $\bar{X} \mid \mu = 0.3 \sim N(0.3, 0.047^2)$. The plot shows this distribution with the area above 0.527 shaded.




a. Where is $\bar{x} = 0.527$ on this plot? About how many standard errors is it above 0.3?

It is off the right side of the plot: $\frac{0.527-0.3}{0.047} \approx 4.8$ standard errors above 0.3.

b. Write the p-value in conditional probability notation. Remember to count both tails.

$P(\bar{X} \geq 0.527 \text{ or } \bar{X} \leq 0.073 \mid \mu = 0.3)$

c. Use the plot to find the p-value. $2\times$ **(less than 0.0001), so less than 0.001**

d. What is your conclusion at $\alpha = 0.05$? Write it in context.

The p-value is less than 0.05, so we reject $H_0$. If the mean mercury level of bass across Florida lakes were 0.3 ppm, a sample mean as far from 0.3 as 0.527 would be extremely unlikely. The data provide convincing evidence that the mean mercury level is not 0.3 ppm.

e. Which type of error could you have made? Describe it in context.

A Type I error: the mean mercury level actually **is** 0.3 ppm, but we concluded that it is not.

**Confidence interval.** The margin of error for a 95% confidence interval is 0.092.

f. Write the 95% confidence interval. **from 0.435 to 0.619 ppm**

g. Interpret the interval in context.

We are 95% confident that the mean of the average mercury levels in largemouth bass across all Florida lakes (like the ones in the study) is between 0.435 and 0.619 ppm.

h. Is 0.3 inside the interval? How does this agree with your conclusion in (d)?

No. 0.3 is not a plausible value of $\mu$, which agrees with rejecting $H_0 : \mu = 0.3$.

i. What does "95%" in "95% confidence" describe: this one interval, or the method? Explain.

The method. If we took many samples of 53 lakes and built an interval from each one, about 95% of the intervals would capture $\mu$. This one interval either captures $\mu$ or it does not.

j. Is the result practically significant? Does it mean that the bass in **every** Florida lake is unsafe to eat? Explain.

Yes, it is practically significant: the sample mean is almost twice the EPA level. But $\mu$ is a mean across lakes. The dot plot shows that some lakes have bass well below 0.3 ppm and some well above, so the result does not tell us about any single lake.

## Part 2: The Bayesian analysis

Jordan, a state official, analyzes the same data the Bayesian way. Before seeing the data, Jordan believes that the bass are probably close to the EPA level. Jordan's **prior** is

$$\mu \sim N(0.3, 0.1^2)$$

In Bayesian inference, $\mu$ is a **random variable**: its distribution describes what we believe about it.

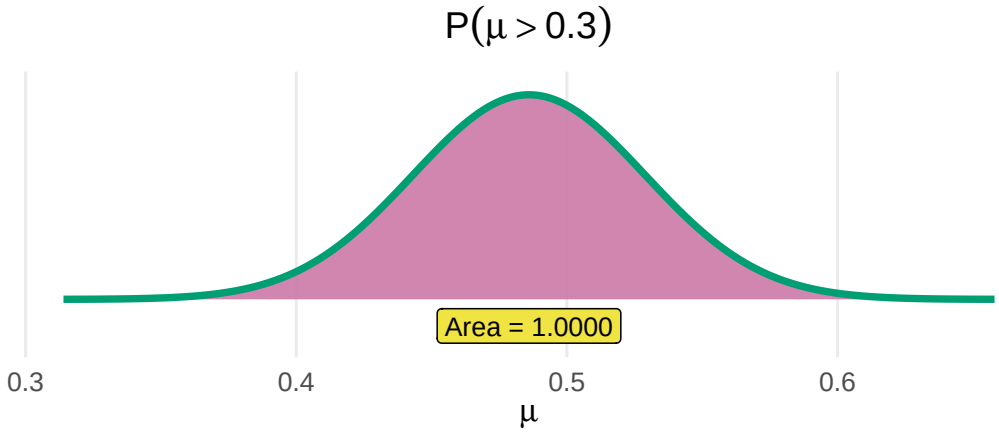
**Hypotheses.** The Bayesian question: is the mean mercury level **above** the EPA level?

- $H_0$: $\mu \leq 0.3$. The mean mercury level is at or below the EPA level.
- $H_A$: $\mu > 0.3$. The mean mercury level is above the EPA level.

a. What is Jordan's **prior** probability of $H_A$, $P(\mu > 0.3)$? How do you know without a calculation?

0.5. Jordan's prior is a normal curve centered at 0.3, so half of the area is above 0.3. $H_0$ and $H_A$ are equally probable before the data.

After seeing the 53 lakes, Jordan's **posterior** is $\mu \mid \text{data} \sim N(0.486, 0.043^2)$. The plot shows this distribution with the area above 0.3 shaded.

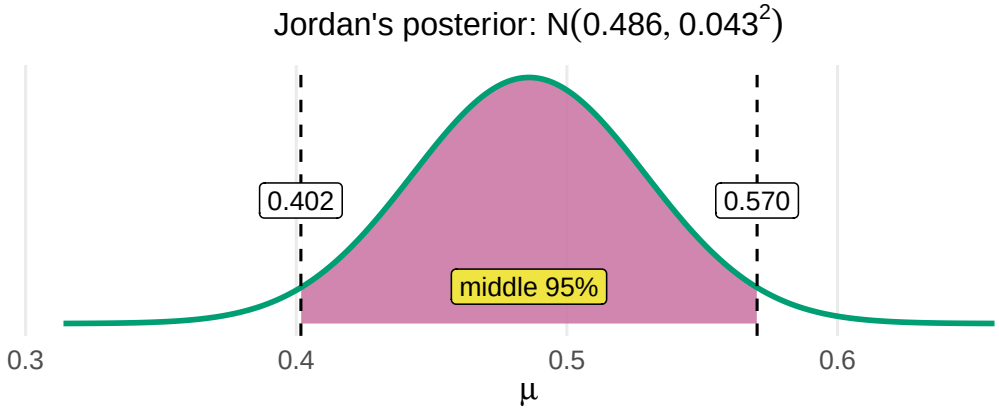



b. Fill in: $P(H_A \mid \text{data}) = P(\mu > 0.3 \mid \text{data})$ **is more than 0.9999** and $P(H_0 \mid \text{data})$ **is less than 0.0001**

c. Write a conclusion in context.

Given the data and Jordan's prior, the probability that the mean mercury level of bass across Florida lakes is above the EPA level of 0.3 ppm is more than 0.9999. The data moved Jordan from "equally probable" to very strong evidence for $H_A$.

**Credible interval.** The plot shows Jordan's posterior with the middle 95% shaded.




d. Jordan's 95% credible interval is from **0.402** to **0.570** ppm.

e. Interpret the interval in context, and write it in probability notation.

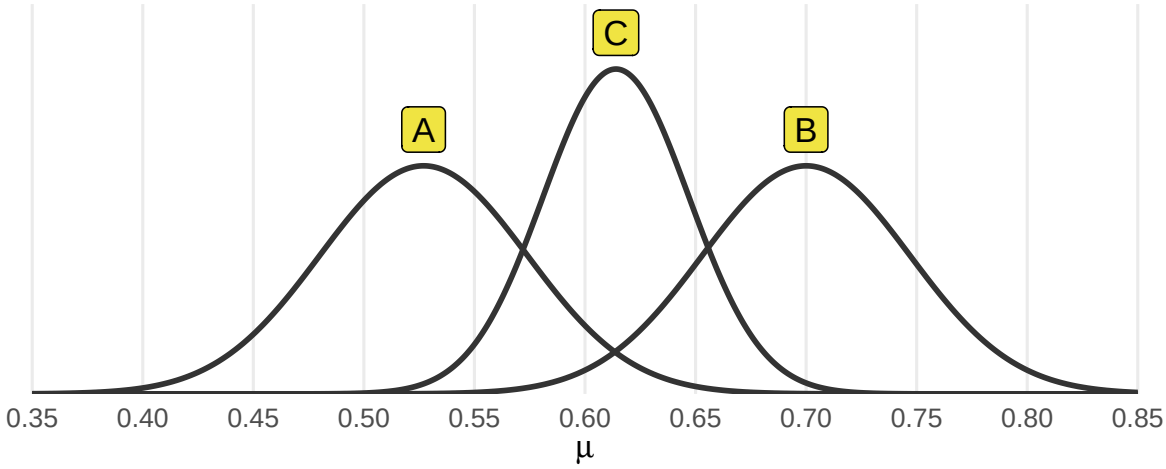
Given the data (and Jordan's prior), there is a 95% probability that the mean of the average mercury levels in bass across Florida lakes is between 0.402 and 0.570 ppm. $P(0.402 < \mu < 0.570 \mid \text{data}) = 0.95$.

f. Jordan's posterior mean, 0.486, is between the prior mean, 0.3, and the sample mean, 0.527. Which one is it closer to? Why?

The sample mean. The data (standard error 0.047) are more certain than Jordan's prior (standard deviation 0.1), so the data get more weight: about 0.82 for the sample mean and 0.18 for the prior mean.

**A different prior.** Riley, an environmental scientist, expects mercury in Florida bass to be high. Riley's prior is $\mu \sim N(0.7, 0.047^2)$. Riley sees the same 53 lakes.

The plot shows Riley's prior, the likelihood, and Riley's posterior, labeled A, B, and C (not in that order).




g. Which curve is Riley's prior? **B** The likelihood? **A** Riley's posterior? **C**

h. Explain how you decided.

The prior is centered at Riley's prior mean, 0.7 (B). The likelihood is centered at the sample mean, 0.527 (A). The posterior is a compromise between the two, so it is in between, and it is narrower than both because it combines two sources of information (C).

i. Riley's prior (standard deviation 0.047) and the likelihood (standard error 0.047) are equally certain. Riley's posterior mean is the weighted average $0.5 \times 0.7 + 0.5 \times 0.527 \approx 0.614$.

j. Jordan's posterior mean is 0.486 and Riley's is 0.614. They saw the same data. Why are their posteriors different? What would bring them closer together?

They started with different priors: Jordan believed $\mu$ was around 0.3 and Riley believed it was around 0.7. With 53 lakes, the prior still matters. With data from many more lakes, the likelihood would get narrower and dominate, and their posteriors would become almost the same.

## Part 3: Same data, two kinds of answers

**Which inference?** For each statement about the mercury data, write **F** (frequentist) or **B** (Bayesian), and write the probability in notation.

Table 1: Frequentist or Bayesian? Statements about the mercury data

| Statement | F or B? | In notation |
|---|---|---|
| If $\mu$ were 0.3, there would be less than a 0.1% chance of a sample mean as far from 0.3 as 0.527. | **F** | $P(\bar{X} \geq 0.527$ **or** $\bar{X} \leq 0.073 \mid \mu = 0.3)$ |
| Given the data, it is almost certain (probability more than 0.9999) that the mean mercury level is above 0.3 ppm. | **B** | $P(\mu > 0.3 \mid$ **data**$)$ |
| Before seeing the data, Jordan thought there was a 50% chance that the mean mercury level is above 0.3 ppm. | **B** | $P(\mu > 0.3)$**, the prior** |
| If we repeated the study many times, about 95% of the intervals built this way would capture $\mu$. | **F** | $P($**interval captures** $\mu \mid \mu) = 0.95$ |
| There is a 95% probability that $\mu$ is between 0.402 and 0.570 ppm. | **B** | $P(0.402 < \mu < 0.570 \mid$ **data**$) = 0.95$ |
| If $\mu$ were 0.3, there would be a 5% chance that we wrongly reject $H_0$. | **F** | $P($**reject** $H_0 \mid H_0$ **is true**$) = 0.05$ |

Look at your notation column. What do all of the frequentist probabilities have in common? What about the Bayesian ones?

Frequentist probabilities are about the **data** (sample statistics, intervals, decisions), assuming a value of $\mu$: $P(\text{data} \mid \text{parameter})$. Bayesian probabilities are about the **parameter** $\mu$, given the data: $P(\text{parameter} \mid \text{data})$ (or $P(\text{parameter})$ for the prior).

**Same curve, different meaning.** Both of these are normal curves that we used today. Complete the table.

Table 2: Comparing the null sampling distribution and the posterior distribution

| | $\bar{X} \mid \mu = 0.3 \sim N(0.3, 0.047^2)$ | $\mu \mid \text{data} \sim N(0.486, 0.043^2)$ |
|---|---|---|
| Name of this distribution | **null sampling distribution** | **posterior distribution** |
| What is on the x-axis (what is random)? | **the sample mean $\bar{X}$** | **the parameter $\mu$** |
| What is fixed (after the \|)? | $\mu = 0.3$ | **the observed data** |
| Frequentist or Bayesian? | **frequentist** | **Bayesian** |
| An area under it is a probability about | **the data (if $\mu = 0.3$)** | **$\mu$ (given the data)** |

**Two tiny numbers.** The p-value is less than 0.001, and Jordan's $P(H_0 \mid \text{data})$ is less than 0.0001. A classmate says: "Both are tiny, so they mean the same thing: there is almost no chance that $H_0$ is true." Do you agree? Explain.

No. Only the Bayesian $P(H_0 \mid \text{data}) = P(\mu \leq 0.3 \mid \text{data})$ is a probability that $H_0$ is true, and it depends on Jordan's prior. The p-value is $P(\text{data this extreme} \mid \mu = 0.3)$: a probability about the **data**, assuming $H_0$ is true. $P(A \mid B) \neq P(B \mid A)$! (The hypotheses are also different: the frequentist $H_0$ was $\mu = 0.3$, the Bayesian $H_0$ was $\mu \leq 0.3$.)

## Part 4: Spot the error

Each statement about the mercury data is **not correct**. Explain what is wrong, and fix it.

a. "The p-value is less than 0.001, so there is less than a 0.1% chance that the mean mercury level is 0.3 ppm."

The p-value is $P(\text{data this extreme} \mid \mu = 0.3)$, not $P(\mu = 0.3 \mid \text{data})$. Fix: "If the mean mercury level were 0.3 ppm, there would be less than a 0.1% chance of a sample mean as far from 0.3 as 0.527." In frequentist inference, we cannot give a probability that $H_0$ is true.

b. "We rejected $H_0$, so we have proven that the mean mercury level is not 0.3 ppm."

A hypothesis test does not prove anything. If $H_0$ were true, results like ours would be extremely rare, but not impossible: we could have made a Type I error. Fix: "The data provide convincing evidence that the mean mercury level is not 0.3 ppm."

c. "There is a 95% probability that $\mu$ is between 0.435 and 0.619 ppm." (This is the frequentist confidence interval.)

In frequentist inference, $\mu$ is fixed, so this interval either contains $\mu$ or it does not. Fix: "We are 95% confident that $\mu$ is between 0.435 and 0.619 ppm": 95% of intervals built this way would capture $\mu$. The "95% probability" interpretation belongs to a Bayesian **credible** interval.

d. "95% of Florida lakes have bass with an average mercury level between 0.435 and 0.619 ppm."

The interval is about $\mu$, the **mean** across all lakes, not about individual lakes. The dot plot shows that many lakes are far outside this interval, for example 24 of the 53 lakes are below 0.435.

e. "In Bayesian hypothesis testing, Jordan started by assuming $H_0 : \mu \leq 0.3$ was true."

Bayesian hypothesis testing does not assume either hypothesis is true. Each hypothesis gets a probability: before the data from the prior, after the data from the posterior. Assuming $H_0$ is true is the frequentist approach.

f. "The likelihood is highest at $\mu = 0.527$, so there is a 100% probability that $\mu = 0.527$."

The likelihood $L(\mu \mid \text{data})$ is not a probability distribution of $\mu$. The highest likelihood means 0.527 explains the data best (the maximum likelihood estimate), not that it is certain. To get probabilities of $\mu$, we need a prior and the posterior.

## Part 5: Multiple choice

Write the letter of the best answer. Then explain your choice in one sentence.

**1.** Which of these is the p-value for the mercury test in Part 1?

(A) $P(\mu = 0.3 \mid \text{data})$

(B) $P(\mu > 0.3 \mid \text{data})$

(C) $P(\bar{X} \geq 0.527 \text{ or } \bar{X} \leq 0.073 \mid \mu = 0.3)$

(D) $P(\bar{X} = 0.527)$

Answer: **C**

The p-value is a probability about the data (the sample statistic), assuming $H_0$ is true.

**2.** In frequentist inference, what is random?

(A) The parameter $\mu$

(B) The sample mean $\bar{X}$, and intervals built from samples

(C) Both $\mu$ and $\bar{X}$

(D) Neither

Answer: **B**

In frequentist inference, $\mu$ is a fixed number; what changes from sample to sample is $\bar{X}$ (and the interval).

**3.** Jordan's 95% **credible** interval for $\mu$ is (0.402, 0.570). Which interpretation is correct?

(A) 95% of all samples of lakes would give an interval that captures $\mu$.

(B) 95% of Florida lakes have bass with an average mercury level between 0.402 and 0.570 ppm.

(C) There is a 95% probability that the sample mean is between 0.402 and 0.570.

(D) Given the data and Jordan's prior, there is a 95% probability that $\mu$ is between 0.402 and 0.570.

Answer: **D**

A credible interval comes from the posterior, so it is a probability statement about $\mu$ given the data. (A) describes a confidence interval, and (B) is about individual lakes, not the mean.

**4.** Jordan also asks whether $\mu$ is above 0.5 ppm. From Jordan's posterior, $P(\mu > 0.5 \mid \text{data}) \approx 0.37$ and $P(\mu \leq 0.5 \mid \text{data}) \approx 0.63$. Which is correct?

(A) Given the data and Jordan's prior, $\mu \leq 0.5$ is about 1.7 times as probable as $\mu > 0.5$: weak evidence either way.

(B) Jordan fails to reject $\mu \leq 0.5$ because $0.37 > 0.05$.

(C) If $\mu$ were 0.5, there would be a 37% chance of data this extreme.

(D) The data prove that $\mu \leq 0.5$, because $0.63 > 0.5$.

Answer: **A**

These are posterior probabilities of the two hypotheses: $\frac{0.63}{0.37} \approx 1.7$. (B) and (C) treat a posterior probability like a p-value.

**5.** Two researchers with very different priors analyze mercury data from a very large number of lakes. Their posteriors will most likely be

(A) still very different, because their priors were very different.

(B) almost the same.

(C) the same as their priors.

(D) impossible to find, because Bayesians must agree on a prior.

Answer: **B**

With a lot of data, the likelihood is very narrow and gets almost all of the weight, so the prior matters very little.

**6.** A significance level of $\alpha = 0.05$ means

(A) $P(H_0 \text{ is true} \mid \text{reject } H_0) = 0.05$

(B) $P(H_A \text{ is true}) = 0.95$

(C) $P(\text{reject } H_0 \mid H_0 \text{ is true}) = 0.05$

(D) $P(\text{Type II error}) = 0.05$

Answer: **C**

$\alpha$ is the probability of a Type I error: rejecting $H_0$ when it is true. $P(A \mid B) \neq P(B \mid A)$, so (A) is a different probability.

**7.** With the same 53 lakes ($\bar{x} = 0.527$, standard error 0.047), a frequentist test of $H_0 : \mu = 0.5$ vs. $H_A : \mu \neq 0.5$ gives a p-value of about 0.57. Which is correct?

(A) The data prove that the mean mercury level is 0.5 ppm.

(B) If $\mu$ were 0.5, results like these would not be surprising, so there is not convincing evidence that $\mu$ is different from 0.5.

(C) There is a 57% probability that $\mu = 0.5$.

(D) There is a 43% probability that $\mu \neq 0.5$.

Answer: **B**

Failing to reject $H_0$ is not evidence for $H_0$, and a p-value is not a probability that a hypothesis is true. Values other than 0.5 are also plausible: the confidence interval is from 0.435 to 0.619.

**8.** The first 5 lakes in the data have average mercury levels 1.23, 1.33, 0.04, 0.44, 1.20 ppm. Mark each one A if it is above 0.3 and N if not: **A, A, N, A, A**. Let $\pi$ be the proportion of **all** Florida lakes whose bass average above 0.3 ppm. Which value of $\pi$ has the highest likelihood?

(A) 0.2
(B) 0.5
(C) 0.6
(D) 0.8

Answer: **D**

$L(\pi \mid \text{data}) = \pi^4(1-\pi)^1$. $L(0.5 \mid \text{data}) \approx$ **0.031** and $L(0.8 \mid \text{data}) \approx$ **0.082**. (Hint: $0.8^4 \approx 0.41$.)

The maximum likelihood estimate is the sample proportion, $\frac{4}{5} = 0.8$. But it is only about 2.6 times as likely as $\pi = 0.5$: with 5 lakes, the likelihood is very spread out. (In all 53 lakes, the proportion is 0.68.) The likelihoods do not add up to 1.

**9.** In our course notation, $\mu \sim N(0.3, 0.1^2)$, with nothing after a |, most likely describes

(A) a prior distribution.
(B) a null sampling distribution.
(C) a posterior distribution.
(D) the sampling distribution of a sample mean.

Answer: **A**

The random variable is the parameter $\mu$ (so it is Bayesian), and nothing is given yet: no data. A posterior would be $\mu \mid$ data; (B) and (D) would be about $\bar{X}$.

## Exit ticket

Explain to a friend who missed class the difference between the frequentist 95% **confidence** interval (0.435, 0.619) and Jordan's 95% **credible** interval (0.402, 0.570) for the mean mercury level. Use the words **random** and **fixed**. Why is Jordan's interval shifted toward 0.3?

For the confidence interval, $\mu$ is **fixed** and the interval is **random**: it changes from sample to sample, and 95% of intervals built this way would capture $\mu$. Our one interval either contains $\mu$ or not. For the credible interval, the data are **fixed** (we condition on them) and $\mu$ is **random**: given the data and Jordan's prior, there is a 95% probability that $\mu$ is between 0.402 and 0.570 ppm. Jordan's interval is shifted toward 0.3 because Jordan's prior was centered at 0.3, and the posterior is a compromise between the prior and the likelihood.