

Bayes' Rule and Learning from Data

The Basic Formula

When we know $P(A | B)$ but want $P(B | A)$:

$$P(B | A) = \frac{P(A | B) * P(B)}{P(A)}$$

Alternative Form:

$$P(B | A) = \frac{P(A | B) * P(B)}{P(A | B) * P(B) + P(A | \neg B) * P(\neg B)}$$

Components:

- Prior: $P(B)$
- Likelihood: $P(A | B)$
- Marginal likelihood: $P(A)$
- Posterior: $P(B | A)$

In many cases, we know $P(A|B)$ but we really want to know $P(B|A)$. This commonly occurs in medical screening, where we know $P(\text{positive test result} | \text{disease})$ but what we want to know is $P(\text{disease} | \text{positive test result})$.

If we have only two outcomes, we can express Bayes' rule in a somewhat clearer way, using the sum rule to redefine $P(A)$.

Putting Bayes into Practice

Construction company drug testing

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. Consider the following scenario:

- In the UK construction industry during 2023, the prevalence of substance use affecting work-place safety was estimated at **2.5%** of the workforce
- The rapid saliva test used has a sensitivity (true positive rate) of **85%** when conducted according to protocol
- The specificity (true negative rate) of these tests is **99.2%**

Let's consider a specific example. Suppose that a worker is selected for a random drug screening. The test result is positive. What is the probability that this worker is actually positive for substances?

Putting Bayes into Practice

Construction company drug testing

Let's consider a specific example. Suppose that a worker is selected for a random drug screening. The test result is positive. What is the probability that this worker is actually positive for substances?

Context: The company's current policy is immediate suspension without pay following a positive test result, pending a more accurate laboratory confirmation test that takes 48 hours.

- Mandatory screening
- Rapid saliva test
- Safety-critical roles
- Immediate consequences

Putting Bayes into Practice

Construction company drug testing

Construction Site Testing:

- Sensitivity: $P(\text{positive}|\text{substance}) = 0.85$
- Specificity: $P(\text{negative}|\text{no substance}) = 0.992$
- Base rate: $P(\text{substance}) = 0.025$

Key Values:

- $P(S) = 0.025$ (prevalence)
- $P(P|S) = 0.85$ (sensitivity)
- $P(P|\text{not } S) = 0.008$ (1 - specificity)

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. In the UK construction industry during 2023, the prevalence of substance use affecting workplace safety was estimated at 2.5% of the workforce. The rapid saliva test used has a sensitivity of 85% when conducted according to protocol, and a specificity of 99.2%.

Let's Work Through It

Using Bayes' Theorem, calculate the probability that this worker is actually positive for substances given their positive test result.

- $P(S) = 0.025$ (prevalence)
- $P(P|S) = 0.85$ (sensitivity)
- $P(P|\text{not } S) = 0.008$ (1 - specificity)

A construction worker is randomly selected for testing at the start of their shift. Their saliva test comes back positive. Using Bayes' Theorem, calculate the probability that this worker is actually positive for substances given their positive test result.

Solution

Calculate $P(\text{substance}|\text{positive})$:

Interpretation:

- ~73.1% chance true positive
- ~26.9% chance false positive

- Much higher than 2.5% base rate
- Still significant uncertainty

Using Bayes' Theorem, we find that given a positive test result, there is a 73.1% probability that the worker actually has substances present. This is much higher than the base rate of 2.5%, but still leaves significant uncertainty with a 26.9% false positive rate.

Discussion: The Real-world Implications

The company's current policy is immediate suspension without pay following a positive test result.

What do these results mean for this business policy? Is it fair to immediately suspend workers without pay for a positive test result?

The company's current policy is immediate suspension without pay following a positive test result, pending a more accurate laboratory confirmation test that takes 48 hours.

Given that approximately 26.9% of positive test results may be false positives, an immediate suspension without pay could unfairly penalize innocent workers; however, the high stakes of construction safety and the 73.1% probability of a true positive suggest that temporary removal from safety-critical roles is prudent while awaiting confirmation.

Learning from Data

Bayes' Rule as Learning:

$$P(B | A) = \frac{P(A | B) * P(B)}{P(A)}$$

Components:

- Prior belief: $P(B)$
- Evidence strength: $\frac{P(A | B)}{P(A)}$
- Updated belief: $P(B | A)$

Key Insights:

- Updates prior knowledge
- Evidence can strengthen/weaken
- Systematic way to learn
- Combines knowledge & data

Another way to think of Bayes' rule is as a way to update our beliefs on the basis of data. The different parts of Bayes' rule have specific names, that relate to their role in using Bayes' rule to update our beliefs.

The part on the left tells us how much more or less likely the data A are given B, relative to the overall likelihood of the data, while the part on the right side tells us how likely we thought B was before we knew anything about the data.

Odds and Odds Ratios

Converting to Odds:

$$\text{odds of A} = \frac{P(A)}{P(\neg A)}$$

Example:

Drug test odds:

- Prior: $\frac{0.025}{0.975} = 0.026$
- Posterior: $\frac{0.7314974}{0.2685026} = 2.724$

Odds Ratio:

$$\frac{\text{posterior odds}}{\text{prior odds}} = \frac{2.724}{0.026} = 106.25$$

Interpretation:

- Odds increased 105×
- Much stronger evidence
- Shows test's power
- Despite false positives

We can convert probabilities into odds which express the relative likelihood of something happening or not. An odds ratio is an example of what we will later call an effect size, which is a way of quantifying how relatively large any particular statistical effect is.

First, remember the rule for computing a conditional probability. We can rearrange this to get the formula to compute the joint probability using the conditional. Using this we can compute the inverse probability.

Bibliography