Unit 2: Probability and distributions
Lecture 2: Bayesian Inference

Statistics 104

Mine Çetinkaya-Rundel

May 22, 2013

What's the chance of winning?

Poll

What is the probability of getting an outcome ≥ 4 when rolling a 6-sided die? What is the probability when rolling a 12-sided die?

(a) 6-sided: $\frac{3}{4}$, 12-sided: $\frac{1}{2}$
(b) 6-sided: $\frac{1}{3}$, 12-sided: $\frac{2}{3}$
(c) 6-sided: $\frac{1}{2}$, 12-sided: $\frac{3}{4}$
(d) 6-sided: $\frac{2}{3}$, 12-sided: $\frac{1}{3}$

Which die is the good die?

Poll

You're playing a game where you win if the die roll is ≥ 4. If you could get your pick, which die would you prefer to play this game with?

(a) 6-sided
(b) 12-sided

Set up

- I have two dice: one 6-sided, the other 12-sided.
- We're going to play a game where I keep one die on the left side (die L) and one die on the right (die R), and you won't know which is the 6-sided die and which is the 12-sided.
- You pick die (L or R), I roll it, and I tell you if you win or not, where winning is getting a number ≥ 4.
- We'll play this multiple times with different contestants.
- I will not swap the sides the dice are on at any point.
- We'll record which die each contestant picks and whether they won or lost.
- The ultimate goal is to come to a class consensus about whether the die on the left or the die on the right is the "good die".
- If you pick right, you all get one extra point on your next problem set. If you pick wrong, you all lose one point.

Hypotheses and decisions

	Truth	
Decision	L good, R bad	L bad, R good
Pick L	*You get an extra point!*	*You lose a point :(*
Pick R	*You lose a point :(*	*You get an extra point !*

Sampling isn't free!
You get to pick how long you want play, but remember, there is a cost associated with too many tries – you're losing precious class time.

Initial guess

Poll

You have no idea if I have chosen the die on the left (L) to be the good die (12-sided) or bad die (6-sided). Then, before we collect any data, what are the probabilities associated with the following hypotheses?

H_1: L good, R bad
H_2: L bad, R good

(a) P(L good, R bad) = 0.33; P(L bad, R good) = 0.67
(b) P(L good, R bad) = 0.50; P(L bad, R good) = 0.50
(c) P(L good, R bad) = 0; P(L bad, R good) = 1
(d) P(L good, R bad) = 0.25; P(L bad, R good) = 0.75

Prior probabilities

- These are your *prior probabilities* for the two competing claims (hypotheses):
  H_1: L good, R bad
  H_2: L bad, R good
- That is, these probabilities represent what you believe before seeing any data.
- You could have conceivably made up these probabilities, but instead you have chosen to make an educated guess.

Results

	Choice (L or R)	Result (win or loss)
Roll 1	L	
Roll 2		
Roll 3		
Roll 4		
Roll 5		
Roll 6		
Roll 7		
Roll 8		
Roll 9		
Roll 10		

Decision making

What is your decision?

How did you make this decision?

Probability tree - roll 1

What is the probability, based on the outcome of the first roll, that L is the good die (and R is the bad die)?

- We want to find P(L good | outcome of 1^{st} roll), a conditional probability.
- We know that
  - P(win | L good) = $\frac{3}{4} = 0.75$
  - P(lose | L good) = $\frac{1}{4} = 0.25$
  - P(L good) = 0.5 (our prior probability)
- We can summarize what we know in a *probability tree* in order to help us calculate the probability we're interested in.
- And we'll implicitly make use of the Bayes' Theorem.

$$\text{P(L good | outcome of } 1^{st} \text{ roll)} = \frac{\text{P(L good AND outcome of } 1^{st} \text{ roll)}}{\text{P(outcome of } 1^{st} \text{ roll)}}$$

Probability tree - roll 1 (cont.)

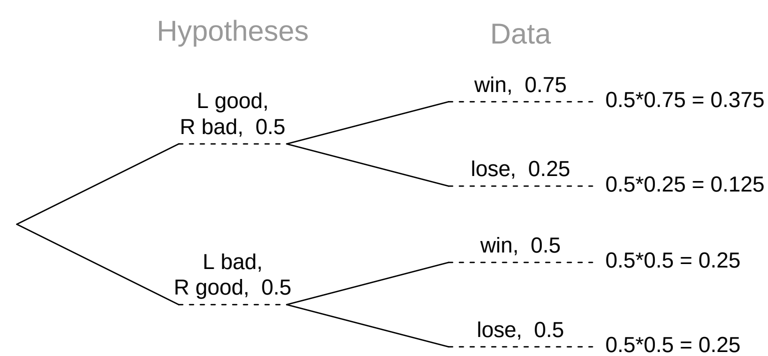
What is the probability, based on the outcome of the first roll, that L is the good die (and R is the bad die)?

$$\text{P(L good | outcome of } 1^{st} \text{ roll)} = \frac{\text{P(L good AND outcome of } 1^{st} \text{ roll)}}{\text{P(outcome of } 1^{st} \text{ roll)}}$$

Probability tree - roll 1 (cont.)

What is the probability, based on the outcome of the first roll, that L is the good die (and R is the bad die)?

$$\text{P(L good | outcome of } 1^{st} \text{ roll)} = \frac{\text{P(L good AND outcome of } 1^{st} \text{ roll)}}{\text{P(outcome of } 1^{st} \text{ roll)}}$$



Posterior probability

- The probability we just calculated

  P(L is good | outcome of 1^{st} roll)

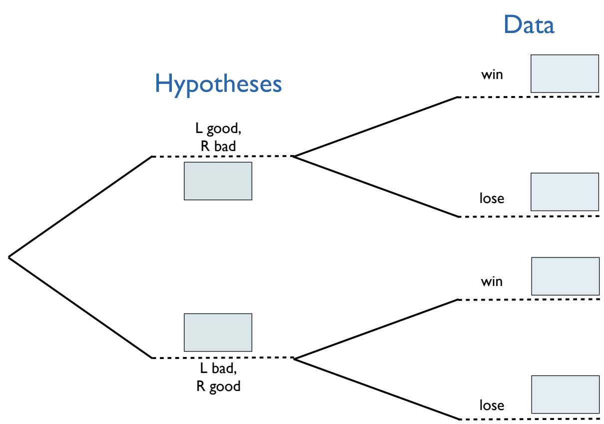
  is also called the *posterior probability*.
- Posterior probability is generally defined as P(hypothesis | data). It tells us the probability of a hypothesis we set forth, given the data we just observed. It depends on both the prior probability we set and the observed data.
- This is different than what we calculated at the end of the randomization test on gender discrimination – the probability of observed or more extreme data given the null hypothesis being true, i.e. P(data | hypothesis), also called a *p-value*. (We're going to be seeing a lot more of these!)

Updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data.
- In the next iteration (roll) we get to take advantage of what we learned from the data.
- In other words, we *update* our prior with our posterior probability from the previous iteration.

Probability tree - roll 2

What is the probability, based on the outcome of the <u>second</u> roll, that L is the good die (and R is the bad die)? This time we update our prior for "L is good", based on what we found in the previous stage.



Rolls 2 through *n*

Calculating the posterior probabilities by hand for each iteration (roll) is tedious and not very efficient. We can use computation instead.

Recap: Bayesian inference

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive Frequentist definition of a p-value as the P(observed or more extreme outcome | H_0 is true). Instead base decisions on the posterior probability, P(hypothesis is true | observed data).
- *Watch out!*
  A good prior helps, a bad prior hurts, but the prior matters less the more data you have.
- More advanced Bayesian techniques offer flexibility not present in Frequentist models.

Breast cancer screening

- American Cancer Society estimates that about 1.7% of women have breast cancer.
  http://www.cancer.org/cancer/cancerbasics/cancer-prevalence
- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.
  http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html
- An article published in 2003 suggests that up to 10% of all mammograms are false positive.
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940

Note: *These percentages are approximate, and very difficult to estimate.*

Calculating the posterior

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer, i.e. what is the posterior probability of having cancer if mammogram yield a positive result?

Setting a prior when retesting

Poll

Suppose a woman who gets tested once and obtains a positive result wants to get tested again. What should be the prior probability that this woman has cancer?

(a) 0.017
(b) 0.12
(c) 0.0133
(d) 0.88

Calculating the posterior when retesting

Poll

What is the posterior probability of having cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52