Lab 7

Probability and Bayesian

2019,11,7

Zhiwei Li

Probability and logic

- Suppose you know the probability of event A and event B happening to be *P(A)* and *P(B)*. How do you describe the probability of:
  - P(A and B)
  - P(A or B)
  - P(either A or B)?
- And which of these 3 number is the smallest? Biggest?

Probability and logic

- Suppose you know the probability of event A and event B to be *P(A)* and *P(B)*. How do you describe the probability of:
  - P(A and B): P(A)*P(B) Conjunction
  - P(A or B): P(A)+P(B) - P(A and B) Disjunction
  - P(either A or B): P(A or B) - P(A and B)
    - Only if A and B are independent!

Conjunction fallacy

- You flip a coin for 6 times. Which of the following sequence is most likely given the coin is fair?
  - THTTT
  - HTHTTT
  - THHHHH

Source of conjunction fallacy?

- You flip a coin for 6 times. Which of the following sequence is most likely given the coin is fair?
  - THTTT? 1/2^5
  - HTHTTT? 1/2^6
  - 1 head in 5 flips?
  - 2 heads in 6 flips?

Source of conjunction fallacy?

- You flip a coin for 6 times. Which of the following sequence is most likely given the coin is fair?
  - THTTT? 1/2^5
  - HTHTTT? 1/2^6
  - 1 head in 5 flips? 1/2^5 *5
  - 2 heads in 6 flips? 1/2^6*15

Prosecutor’s fallacy

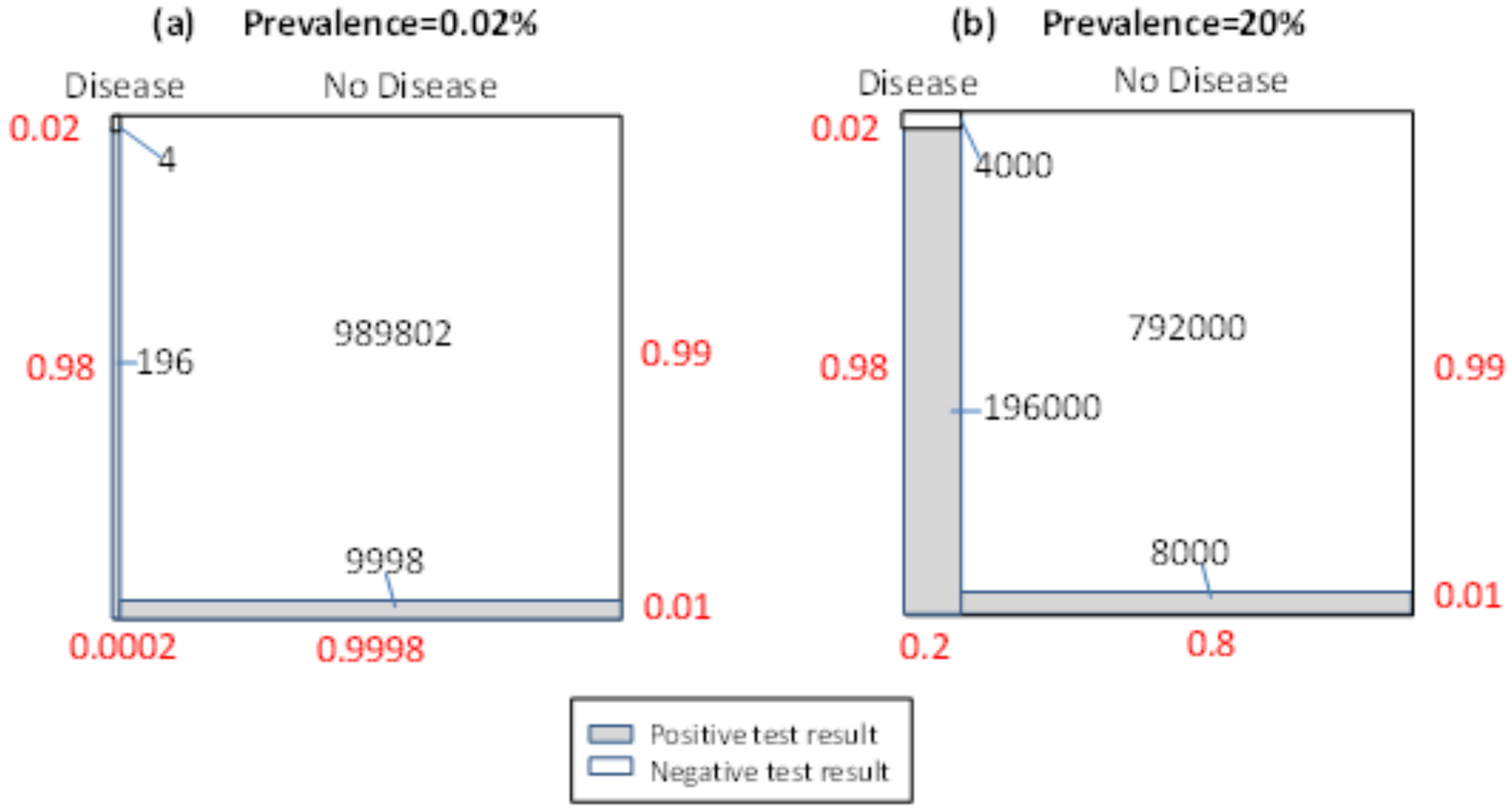
- She hasn’t replied my message for 3 days, therefore she mustn’t love me any more.

- My gene test result is positive for this deadly disease and this test is sensitive to 98% of the patients — I will die!!

1000000 people doing test on two different diseases. One rare and one common.
Both tests have sensitivity rate of .98
false positive rate of .01

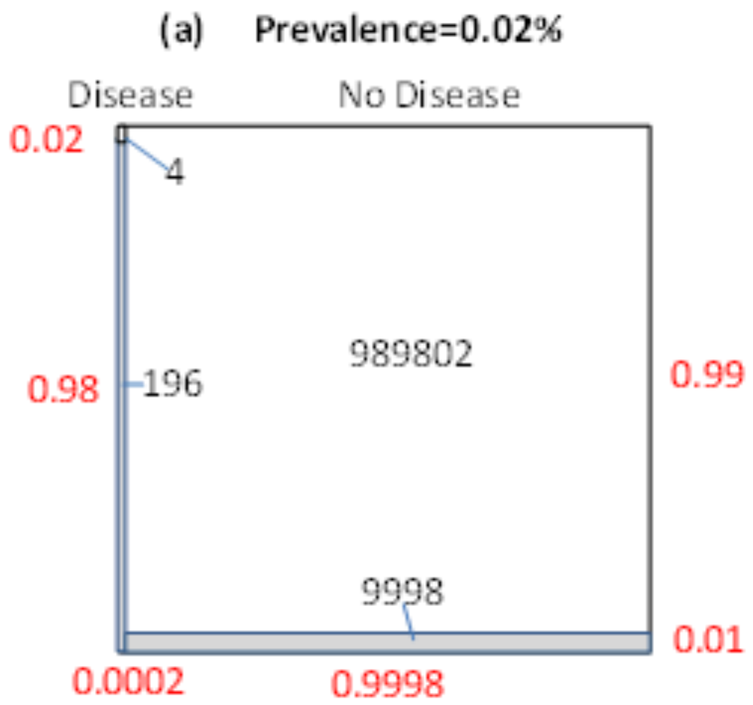
	disease	no disease
test negative	0.02	0.99
test postivie	0.98	0.01

1000000 people doing test on two different diseases. One rare and one common.
Both tests have sensitivity rate of .98
false positive rate of .01



P(disease|test positive) = P(test pos|disease) * P(disease)/P(test pos)

.98 .0002 ?



probability

	disease .0002	no disease .9998
test negative	0.02	0.99
test postivie	0.98	0.01

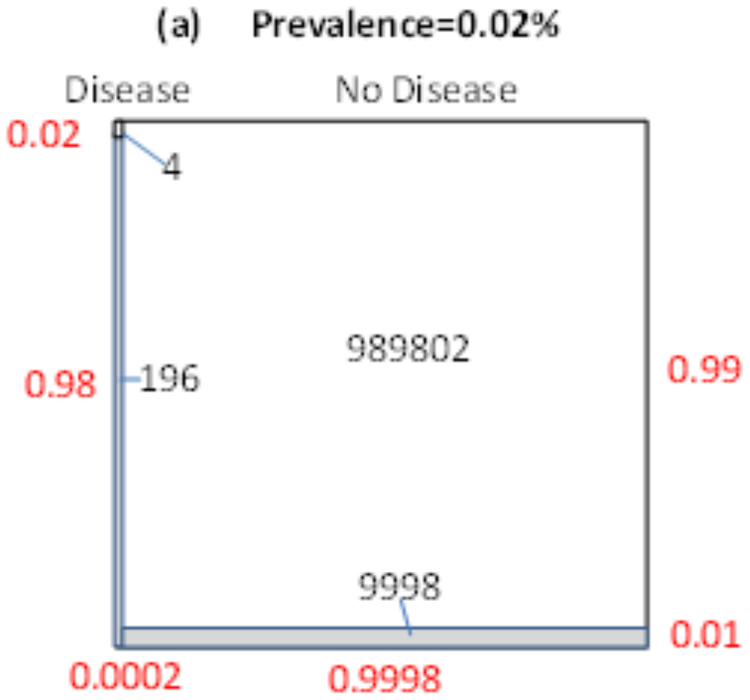
frequency

	disease	no disease	marginal
test negative	4	989802	
test postivie	196	9998	
marginal			1000000

$$P(\text{disease}|\text{test positive}) = P(\text{test pos}|\text{disease}) * P(\text{disease})/P(\text{test pos})$$

.98 .0002 0.0102

$$P(\text{test pos}) = P(\text{test pos}|\text{disease})P(\text{disease}) + P(\text{test pos}|\text{no disease})P(\text{no disease})$$



	disease	no disease	marginal
test negative	4	989802	989806
test postivie	196	9998	10194
marginal	200	999800	1000000

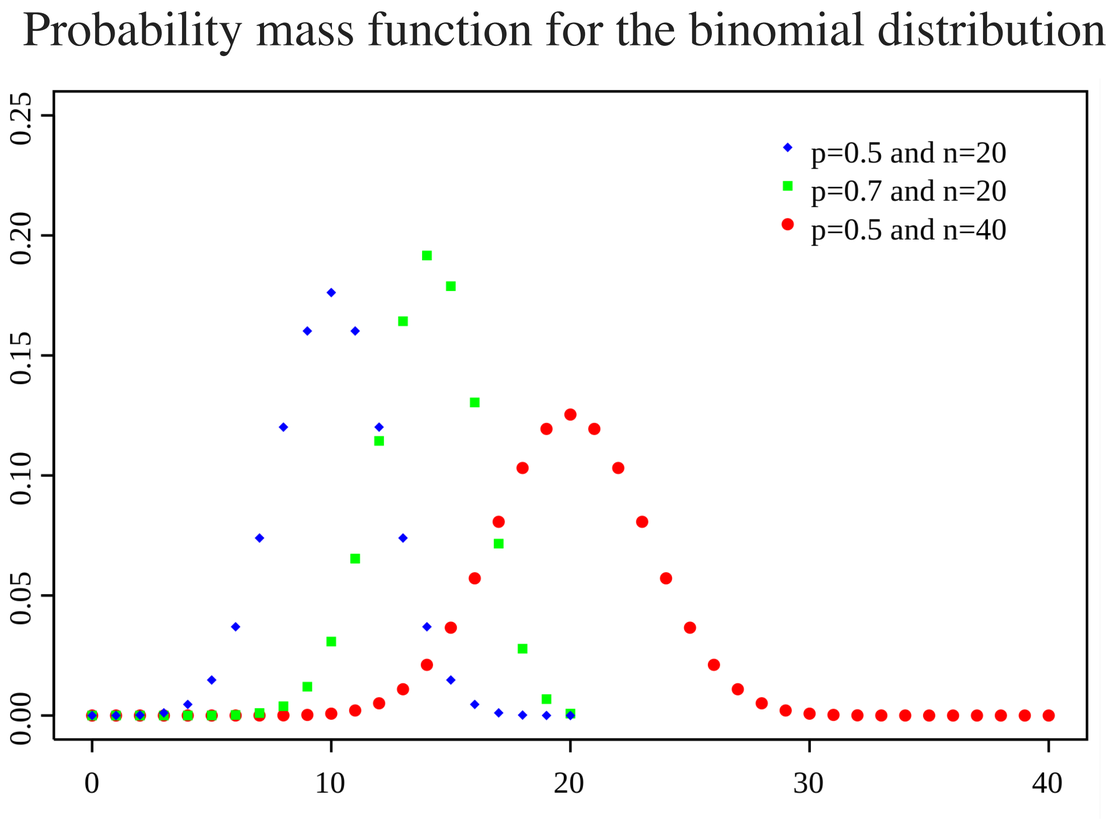
Reply to Prosecutor’s fallacy

- Your gene test result is positive for this deadly disease and this test is sensitive to 98% of the patients — please start panicking.
  - but the disease base rate is so low. And false positive rate cannot be ignored.
- She hasn’t replied my message for 3 days, therefore she mustn’t love me any more.
  - but the probability she hates me is so low. And her not checking message bc doing mathtool homework cannot be ignored.

How many students have letter “e” in their name?

- Letter "e" has as frequency of 12% in English words. How likely is its appearance in names? Now let's ask 10 students in the room.
  - How many student's names do you expect to have letter "e"?
  - What’s the variance of this number?

Binomial distribution



https://en.wikipedia.org/wiki/Binomial_distribution

- Letter "e" has as frequency of 12% in English words. How likely is its appearance in names? Now let's ask 10 students in the room.
  - Bino(N,p) where N = 12, p = 5/12

Bernoulli distribution

~ Binomial(p, n=1)

$$\text{mean} = p$$

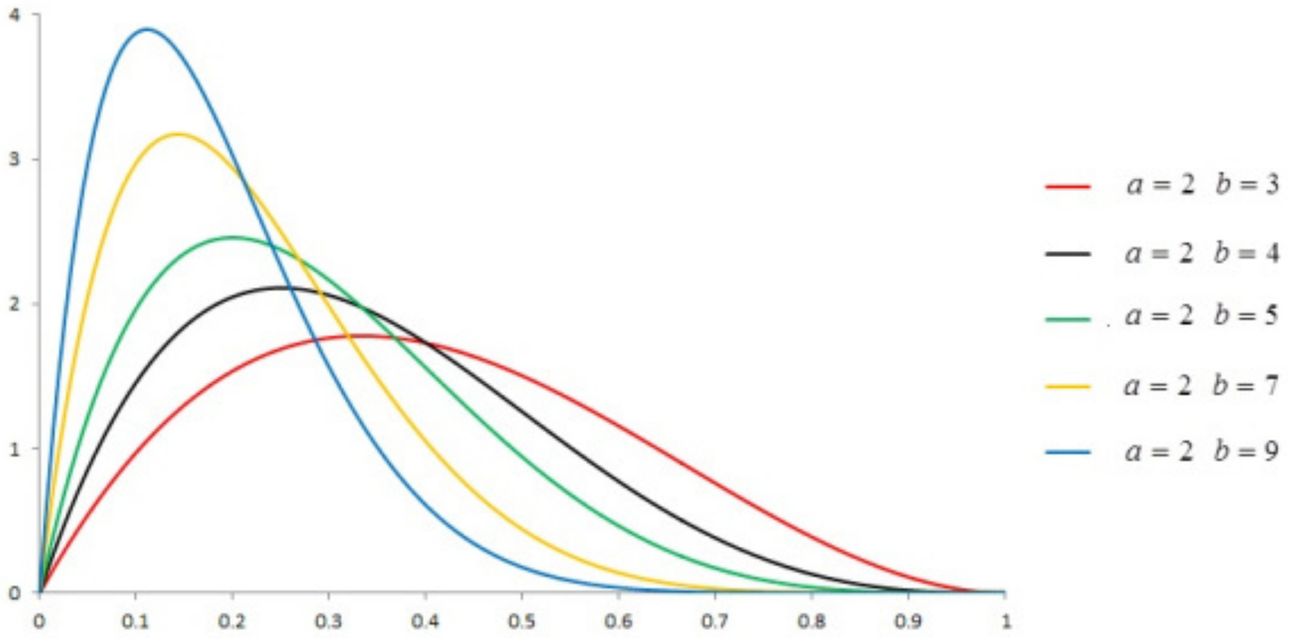
$$\text{var} = p(1-p) \, .$$

- Letter "e" has as frequency of 12% in English words. How likely is its appearance in names? Now let's ask 12 students in the room.
  - X~Bern(p).
  - Y = 12*X. [sum of 10 variables from the same X distribution]
  - E(Y)?
  - std(Y)? (thinking in terms of variance may be easier)

Posterior of coin flip

- Now let’s ask more students to update our estimation.
- How do you estimate the probability of “e” appearing in a name?

Beta distribution

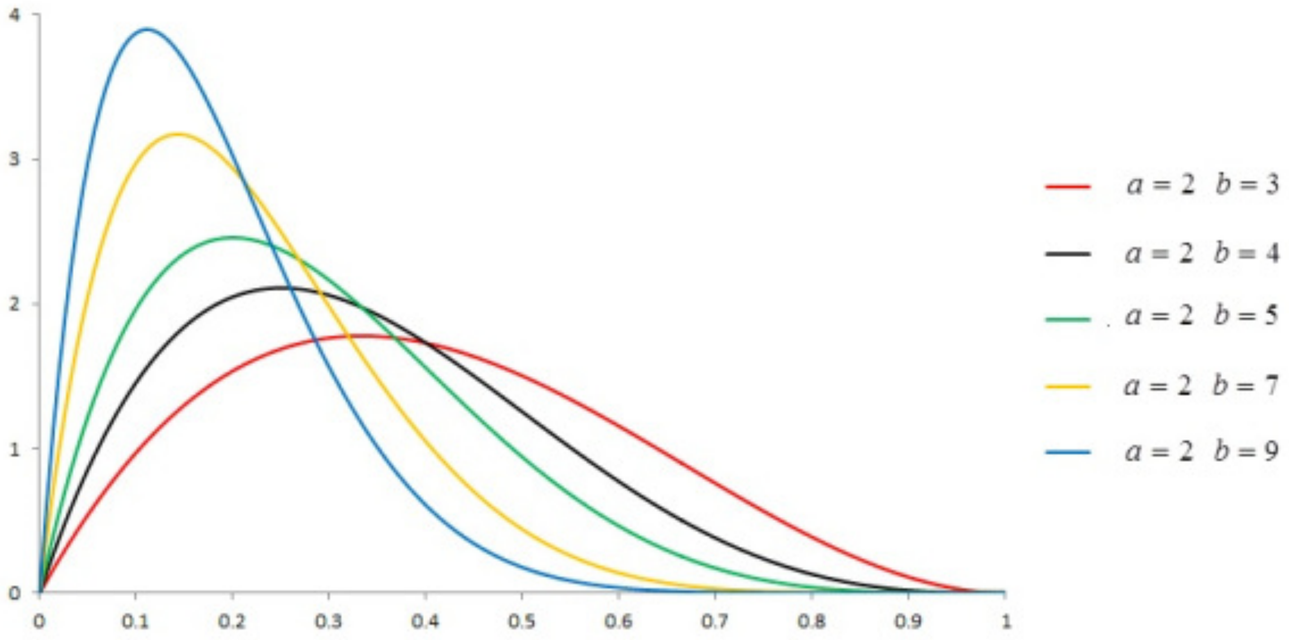


- Support (range of x): [0,1]
- Interpreted as the Bernoulli distribution parameter *p*.
  - "Conjugate prior" of Bernoulli distribution.
- Parameter: a,b => Beta(p|a,b)

posterior = prior * likelihood

Beta = Beta*Bernoulli

Beta distribution intuition



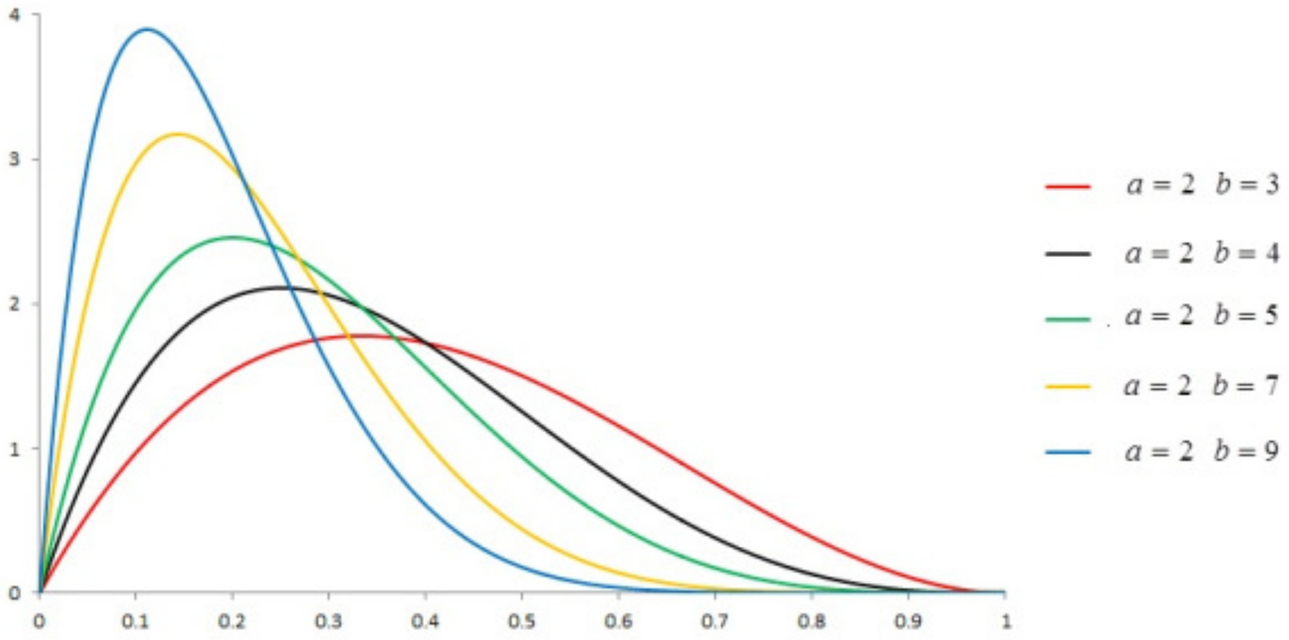
Beta(p|a,b):

- a=number of “hit”
- b=number of “miss”
- a+b = N = total number of events.

posterior = prior * likelihood

Beta = Beta*Bernoulli

Beta distribution update



What if we add 5 more data points that “hit” (by just asking people who I know has “e” in their names)?

- How will the distribution move?

- What should be the mode of the posterior?

posterior = prior * likelihood

Beta = Beta*Bernoulli

Conjugate prior

- Beta is the conjugate prior of Bernoulli.
- Normal is the conjugate prior of itself!

- Test to convince yourself!