

Probability/Statistics Review Algorithms

Analysis of Biological Sequences

140.638

Role of probability and statistics

- You need to know
 - What to expect from an experiment
 - When you get a result, how likely is it that it was an accident, versus an interesting and new behavior?
- What is a p-value and why is it useful?
- What does a p-value really measure?

Basic concepts

Null hypothesis: outcome that you would expect if your data come from the null distribution

offspring height is determined in part by the parents' height

null hypothesis: sons' height is average of parents' height plus $2x \text{ (dad-mom)}/3$

sample of 3 sons: two are 3 inches shorter than the mother. Is this unusual??

Basic concepts

Null hypothesis: outcome that you would expect if your data come from the null distribution

sample of 3 sons: two are 3 inches shorter than the mother. Is this unusual??

methods:

- 1) probability, calculated from a distribution
- 2) simulations to calculate a p-value
- 3) compare to an alternative model

Basic concepts

1) probability, calculated from a distribution

calculate the distribution of heights under the model that is specified by the equations.

where are these kids?



Basic concepts

2) simulations to calculate a p-value

From the null distribution, draw three value at random. In how many trials are two or more values smaller than three inches less than the mother's height?

Do this 10,000 times. If in only 2 cases is the simulation more extreme than the kids' heights, there is only a $2/10,000$ chance that the kids' heights come from the null distribution. $p < 0.0002$.

Basic concepts

3) compare to an alternative model

nonpaternity rate: about 10%

$p(\text{height}|\text{paternity})$

Basic concepts

Example: I have a pocket full of candy. If the candy is half chocolate and half licorice and there are 5 pieces of each,

what are the chances of picking one of each type if I only pick two pieces? (probability)

If I pick 5 pieces and they are all licorice, what are the odds that this is an honest setup? (statistics)



“number sense”

- With small numbers, humans do fine
 - Family with 6 boys
 - All Brazil nuts at the top
- In fact, maybe we're too good at seeing connections
 - Superstitions
 - Horoscopes
- With big numbers, human brains fall flat
 - Lottery
 - Microarray experiments
 - Sequencing experiments

Role of probability and statistics



Common terms

Mean:

“average”—assign numbers to each outcome, add up all numbers and divide by # of outcomes

1 dog has 4 legs, 1 human has 2 legs. Mean # legs for 1 dog and 2 humans is $(4+2*2)/3 = 8/3 \sim 2.67$

Expectation: generalized mean, most likely outcome

(2 legs!)



Common terms cont'd

Median: value in the middle | | | 7 7 9 9

less sensitive than mean to extreme values:

| | | 7 7 9 200000000 median=7 mean=40,000,005

| | | 7 7 9 median=4 mean=4.33

Mode: most common value

Variance: measure of the range or spread of the data

-2 0 2 0 median=mean=0

-1 1 median=mean=0

Common terms cont'd

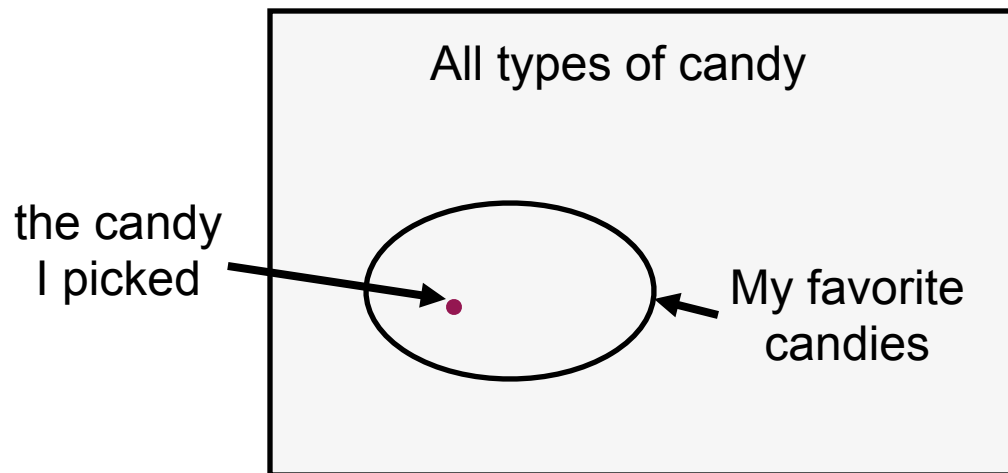
Sample space: all conceivable outcomes

Sample point: one specific outcome

Event: a set of outcomes from the sample space

Random variable: a quantity that takes its value from the sample space with some degree of randomness

Probability of event A: how likely is it that event A will occur?



C = type of candy
chosen on a single trial

Want to know
 $P(C \text{ is a candy I like})$

Definitions

Union

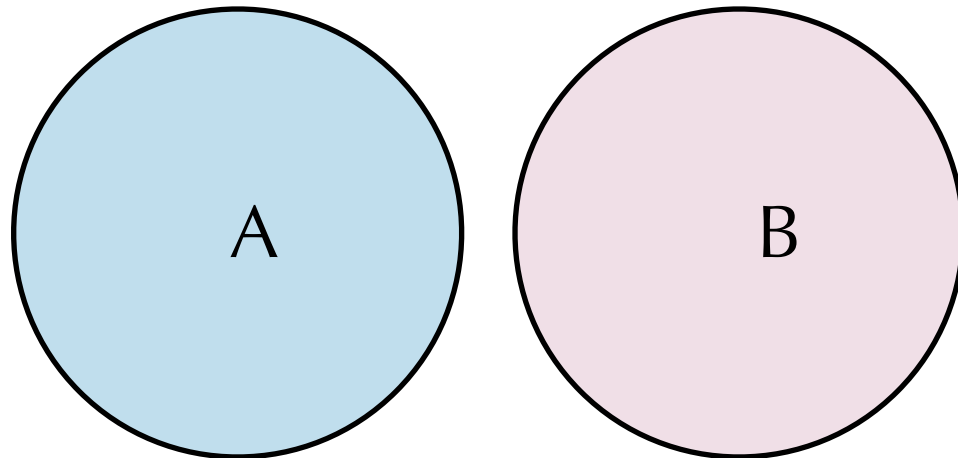
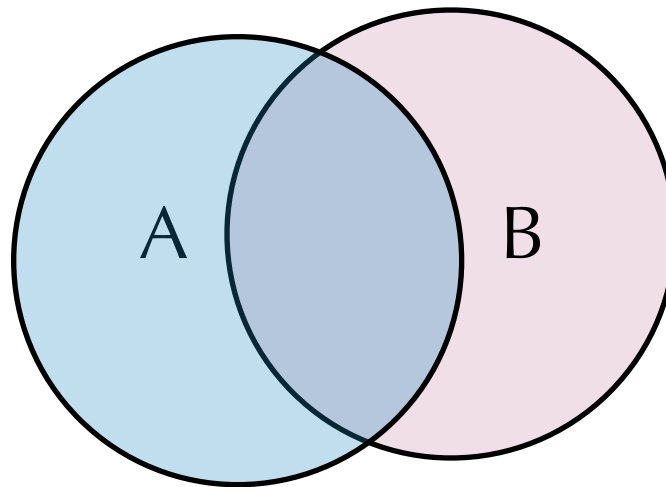
Intersection

Complement

Mutually

Exclusive

Exhaustive



Definitions

Relative frequency

look at 10,000 flowers:

6721 white, 3279 yellow

relative frequency of white
flowers is 67.21%

Probability

if I look at an infinite # of
flowers what will I see?



Basic probability

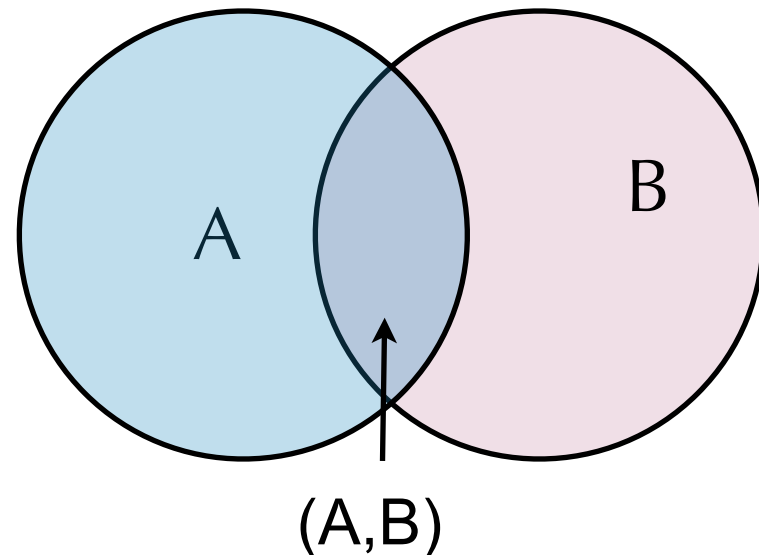
$p(A)$ = probability that event A will occur (picked a chocolate candy)

$p(B)$ = probability that event B will occur (picked a candy shaped like an animal)

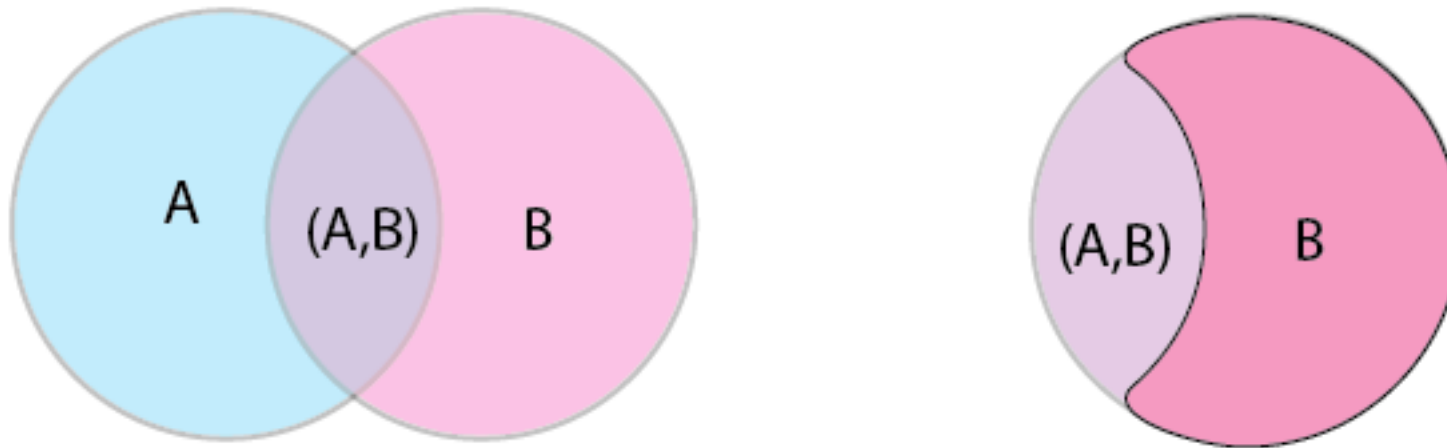
$p(A,B)$ = probability that both A and B will occur

$p(A,B) = p(B,A)$

From this picture you don't know how likely either A or B is



Conditional probability (Bayes' theorem)

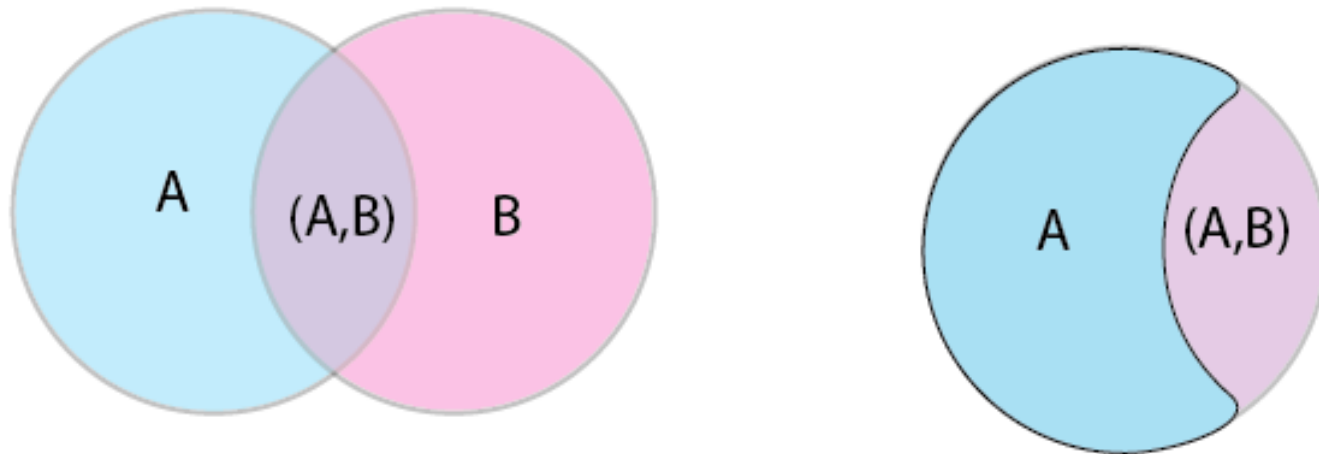


$P(A|B)$ = probability that A will occur, given that B has already happened

$$P(A|B) = P(A,B)/P(B)$$

rearrange: $P(A,B) = P(A|B)P(B)$

Conditional probability (Bayes' theorem)



$P(B|A)$ = probability that B will occur, given that A has already happened

$$P(B|A) = P(A,B)/P(A)$$

rearrange: $P(A,B) = P(B|A)P(A)$

Conditional probability

$$P(A,B) = p(A)p(B|A) \text{ (Bayes' formula)}$$

$$P(A,B) = p(B)p(A|B)$$

$$p(A)p(B|A) = p(B)p(A|B)$$

$$p(A) = p(B)p(A|B) / p(B|A)$$

Independence

If A and B are two events from the same sample space S
then A and B are independent if $p(A|B) = p(A)$
(definition)




Two events are independent if knowing the outcome of
one does not alter the expectation for the outcome of
the other.

(microarray??)

Permutations and Combinations

Permutation: number of ways to arrange r out of n possible objects

example: choose the gold, silver and bronze finishers from 10 contestants

1 2 3 4 5 6 7 8 9 10	\longrightarrow choose	6 8 9	\longrightarrow arrange	 8  9  6	$\frac{n!}{(n-r)!}$ $\frac{10!}{(10-3)!}$
					$\frac{3628800}{5040} = 720$

Permutations and Combinations

Combinations: number of ways to choose r out of n possible objects

Example: choose 3 candidates from a list of 10 genes

$${}_n C_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}$$

$$\frac{10!}{(10-3)!}$$

permutation

$$\frac{10!}{(10-3)!3!}$$

combination

$$\frac{3628800}{5040*6} = 120$$

Discrete probability distributions

- Bernoulli trial: repeat an experiment x times, with 2 possible outcomes; the probability of getting one outcome or the other does not change during the experiment

- e.g rolling dice: $P(x_i) = P(\text{roll } 1 \text{ or } 2) = \frac{2}{6} = \frac{1}{3}$

$$P(x_j) = P(\text{roll } 3, 4, 5, \text{ or } 6) = \frac{4}{6} = \frac{2}{3}$$

Binomial distribution

Trials must be Bernoulli trials

The number of trials that will be performed, n , is fixed in advance (not dependent on the outcomes of the trials)

Then it's possible to predict outcomes. For this example, x_i is getting a 1 or 2 on a roll of a dice, and x_j is getting a 3, 4, 5, or 6

$$P(x_i=r) = \binom{n}{r} P(x_i)^r P(x_j)^{(n-r)}$$

Binomial distribution

P(roll dice 10 times and get 4 results that are 1 or 2)

$$n = 10$$

$$p_i = 1/3$$

$$p_j = 2/3$$

$$r = 4$$

$$P(x_i=r) = \binom{n}{r} P(x_i)^r P(x_j)^{(n-r)}$$

$$P(x_i=4) = \binom{10}{4} (1/3)^4 (2/3)^6 = 22\%$$

Poisson Process

Outcomes are discrete

The number of successes in any one interval is independent of the number of successes in any other interval

Probability of success is sufficiently small (often restated as: probability of getting two successes in a sufficiently small interval is essentially zero)

Time-interval or space-interval statistics—extremely useful in sequence biology

Poisson Distribution

- The Poisson distribution is the limiting form of the binomial distribution as $n \rightarrow \infty$ and $p \rightarrow 0$ but $np = \lambda$ remains finite
- λ is the expected number of occurrences in a given interval (mutations over time, # binding sites per length interval, etc)
- The probability of an event is constant over time (!!)

$$p(n \text{ events}) = \frac{e^{-\lambda} \lambda^n}{n!}$$

Poisson Distribution

Used to model rare events like

radioactive decay

mutations: if there are 5×10^{-9} mutations per human cell per year, what is the probability that a 95-year-old colonic stem cell has at least one mutation?

$$p(0) = \frac{e^{-\lambda} \lambda^0}{0!} = 0.99999525$$

$$p(1) = 4.75 \times 10^{-6}$$

$$p(3) \sim 10^{-17}$$

But there are 75000000000 colonic stem cells in a typical human!

Continuous probability distributions

- N is very big
- Probability of any one event is vanishingly small
- Look at the probability of a range of events

- Normal distribution, exponential distribution

Useful distributions

Geometric	discrete	# trials before the first failure
hypergeometric	discrete	Sampling without replacement
Binomial	discrete	2 outcomes (Bernoulli trials)
Multinomial	discrete	General form of binomial (>2 outcomes)
Poisson	discrete	Distribution of rare events
Normal	continuous	Binomial for large n
Exponential	continuous	Time to next event (distance between mutations)

patterns in biological sequences

looking for similarities, motifs, binding sites, coding regions etc. is a big part of computational biology!!

Same question: what is the probability that this is a real result, given the background distribution of nucleotides?

Typical computation:

probability that the result came from the match model

probability that the result came from the random model

patterns in biological sequences

for example:

malaria genome has 40% A, 40% T, 10% C, 10% G

motif ATAA is found at 1/3 of my 1000 4-bp ChIPseq sites

$$p(\text{ATAA} \mid \text{random}) = 0.4 * 0.4 * 0.4 * 0.4 = 0.0256$$

$$p(\text{not ATAA} \mid \text{random}) = 1 - 0.0256 = 0.9744$$

- 1) try probability ${}_{1000}C_{333} = 5.8 \times 10^{274}$... not helpful
- 2) statistics? chi square: $p < 0.00001$
- 3) simulation: 1000 times, randomly pick 4 nucleotides from the distribution above
result: get > 333 ATAA zero times. $p < 0.001$

Entropy

Measure of how close to uniform the distribution is

(~unpredictability) ... Like variance in a way but not the same thing

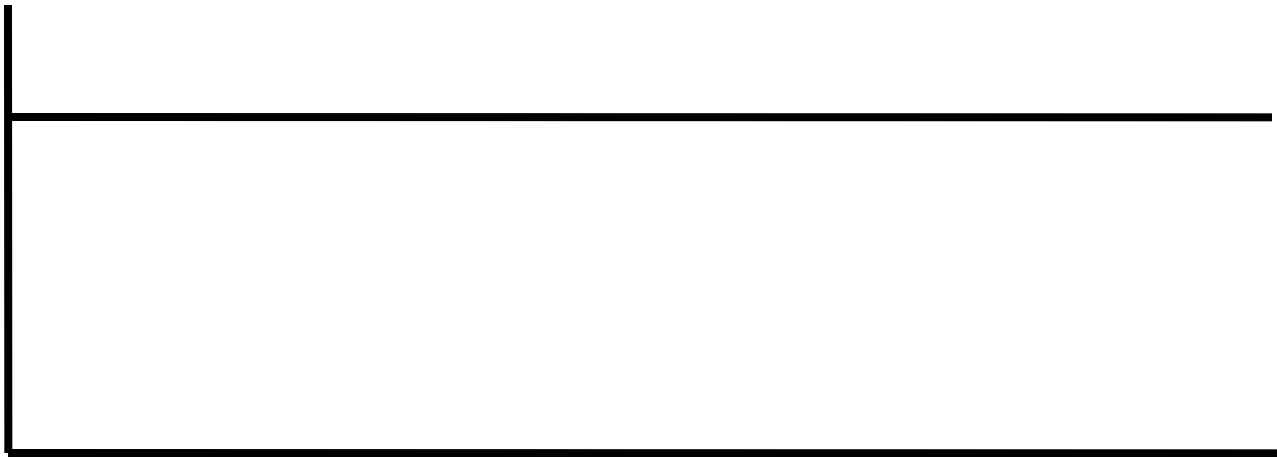
Entropy of random DNA (Wootton and Federhen definition):

$$S = \left(\frac{1}{N}\right) \log_k \left(\frac{N!}{\prod_{i=1}^k n_i!} \right)$$

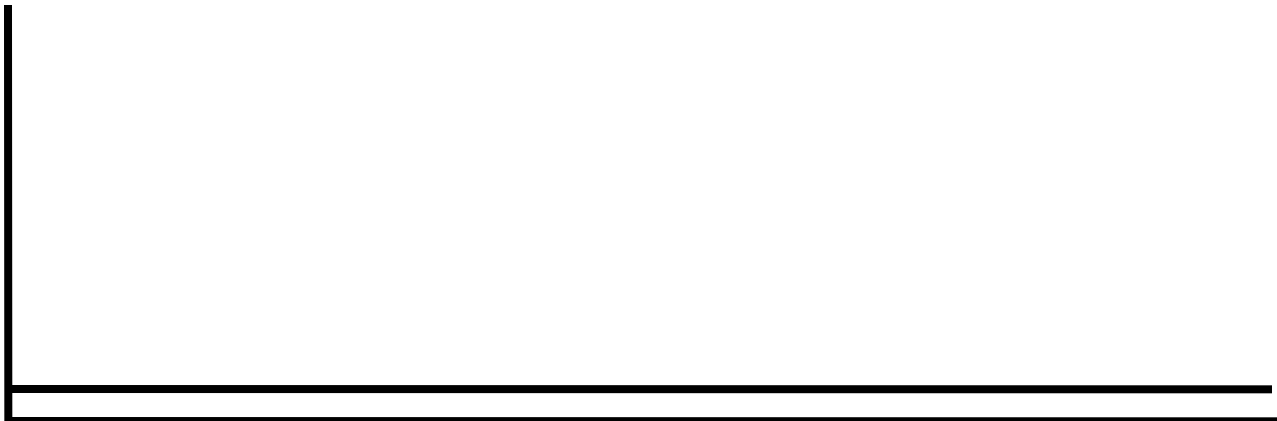
$$\text{ACAGGTTTCT} \quad S = \left(\frac{1}{10}\right) \log_4 \left(\frac{10!}{2!2!2!4!} \right)$$

$$\text{AAAAAAAAAA} \quad S = \left(\frac{1}{10}\right) \log_4 \left(\frac{10!}{10!} \right)$$

Entropy



ACTGACTGATCGACGTACGTACGTACGTACGT



AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Entropy

Computing in windows is critical to assessing landscape



Algorithms

- An algorithm is just a procedure for doing something
 - Recipe
 - Lab or medical protocol
 - Driving directions
- Some algorithms are better than others (time, space, ease of implementation)
- Choice of algorithm depends on data, computer, intent

Algorithms

- Greedy (multiple alignment)
- Exhaustive
- Branch-and-bound (phylogeny)
- Dynamic programming (sequence alignment)
- Binary search
- Depth-first search (short read alignment)
- Breadth-first search (short read alignment)

Time and space complexity

- Add all numbers from 1 to n:

$1+2+3+ \dots + n-1 + n$ (linear time and space)

$(n+1)*n/2$ (constant time, constant space)

- $O(n), O(n^2), O(n^3) \dots O(x^n) \dots$

Algorithm notation (pseudocode)

- Just be organized!
- if there are numerical parameters or limits, use “for” or “while” constructions (with counters)
- specify stopping points, special conditions

example:

```
print ("I'm going to count to three:")  
i=1  
while (i < 4)  
    print ("i")  
    i = i+1
```