In-class exercises: Probability

Kris Gunsalus

9/24/2020

Practice problems

1. Co-occurrence of infection: Venn diagams

Among women voluntarily tested for sexually transmitted diseases in one university (modified from Tábora et al. 2005):

- 18% tested positive for human papilloma virus (HPV) only,
- 2% tested positive for Chlamydia only, and
- 3% tested positive for both HPV and Chlamydia.

Use the following steps to calculate the probability that a woman from this population who gets tested would test positive for either HPV or Chlamydia.

a. Write the goal of the question using probability notation (e.g. P(something) or Pr[something]).

What is P(HPV OR Chlamydia)?

b. Write the general addition rule, as applied to this example.

P(HPV OR Chlamydia) = P(HPV) + P(Chlamydia) - P(HPV AND Chlamydia)

c. First, calculate the total frequency of HPV and of Chlamydia in the population.

```
# probabilities
p.hpv = 0.18
p.chl = 0.02
p.hpv.chl = 0.03

p.hpv.tot = p.hpv + p.hpv.chl
p.chl.tot = p.chl + p.hpv.chl
p.hpv.tot
```

```
## [1] 0.21
```

```
p.chl.tot
```

```
## [1] 0.05
```

d. Calculate the probability that a randomly sampled woman would test positive for at least one of these, using the total probabilities and the joint probability.

```
# P(HPV OR Chlamydia)
p.hpv.tot + p.chl.tot - p.hpv.chl
```

```
## [1] 0.23
```

```
# this is the same as adding up the original numbers
p.hpv + p.chl + p.hpv.chl
```

```
## [1] 0.23
```

e. Is the occurrence of these infections independent? Explain.

```
# Independence: P(HPV AND Chlamidia) = P(HPV) * P(Chlamydia)
indep.test = p.hpv.tot * p.chl.tot
indep.test
```

```
## [1] 0.0105
```

```
indep.test == p.hpv.chl
```

```
## [1] FALSE
```

2. Allele frequencies: Addition and multiplication rules

Many gene loci have a major allele and a number of minor alleles in the population. Let's say there are 5 alleles for a particular locus, that individuals mate randomly with respect to the genotype at this locus, and that allele frequences for A1-A5 are 0.82, 0.06, 0.05, 0.04, and 0.03 respectively.

For each of the questions below, write out the probability equations (e.g. Pr[Ai] or P(Ai) for each term) and then compute the answers numerically.

a. What is the probability that any single allele chosen at random from the population is either A4 or A5? Write out the probability statement as a comment below and then computer the answer.

```
# Allele frequencies
A1 = 0.82
A2 = 0.06
A3 = 0.05
A4 = 0.04
A5 = 0.03

# Probability of either A4 or A5 = ?

# P(A4 OR A5) = P(A4) + P(A5)
A4 + A5
```

```
## [1] 0.07
```

b. What is the probability that an individual carries two A2 alleles?

```
# P(A2 AND A2) = P(A2)*P(A2)
A2*A2
```

```
## [1] 0.0036
```

c. What is the probability that someone does NOT carry two A2 alleles?

```
#P(not (A2 AND A2) = 1 - P(A2 AND A2)
1 - A2^2
```

```
## [1] 0.9964
```

d. What is the probability that someone is heterozygous for A1 and A3?

```
# P(A1 AND A3) = 2 * P(A1)*P(A3)
2*A1*A3
```

```
## [1] 0.082
```

e. What is the probability that *neither* of two random individuals in the population would carry two A1 alleles?

```
# P(2 individuals are NOT A1;A1) = (1 - P(A1;A1))^2
(1-A1*A1)^2
```

```
## [1] 0.1073218
```

f. What is the probability that 3 random individuals carry no A4 or A5 alleles at all? (Remember that each individual carries two alleles.)

```
# P(A4 or A5) = P(A4) + P(A5)
# P(NOT (A4 or A5)) = 1 - P(A4 or A5)
# P(3 are not A4 or A5) = ( 1 - P(A4 or A5) )^3
p4or5 = A4 + A5
pNOT4or5 = 1 - p4or5
pNOT4or5^6 # 3 invididuals * 2 alleles each
```

```
## [1] 0.6469902
```

3. Nucleic acids: Sampling, permutations and combinations

a. Restriction enzymes recognize specific sequences in DNA and cut the DNA within or near those sites. How many possible restriction sites of length 6 are there?

```
4^6
```

```
## [1] 4096
```

```
# There are 4 possibilities for each position, and they are drawn independently,
# so there are 4*4*4*4*4*4 = 4^6 possibilities.
```

b. At what frequency would you expect to find a binding site for a particular restriction enzyme that recognizes a 6 bp sequence? *Note: Restriction enzymes usually recognize a palindromic sequence, so you don't need to worry about looking on both strands.*

```
1/4^6
```

```
## [1] 0.0002441406
```

```
# Since there are 4096 possible 6-mers, you would expect to find a specific one
# around 1/4096th of the time.
# This is 2.4 sites per 10,000 bases, or one site every 4096 bases.
```

c. What is the average length of a fragment produced by a 6-cutter?

```
4^6
```

```
## [1] 4096
```

```
# One cut every 4096 bases on average translates to an expected fragment size of 4096.
```

d. The EcoR1 enzyme cuts the sequence "GAATTC". How many different sequences could be made out of this set of nucleotides (assume that you treat each A and T as distinct individuals)?

```
# nPerm = n!
factorial(6) # 720
```

```
## [1] 720
```

d. You are interested in synthesizing a bunch of random oligonucleotides for a SELEX experiment. How many different oligos of length 22 could you synthesize?

```
4^22 # 1.76e+13
```

```
## [1] 1.759219e+13
```

e. How many ways could you make a sequence of 6 nt by grabbing them at random from a bag of 12 nucleotides (assuming each base is treated as a distinct unique entity, and you can only pick each nt once)? This is an example of "random sampling without replacement".

```
# nPerm = n!/(n-k)!
factorial(12)/factorial(12-6) # 665280
```

```
## [1] 665280
```

f. How many combinations of 6nt are there in a set of 12 random nucleotides? (That is, pick any group of 6 nt, where each one is treated as unique, and you don't care about the sequence in which you pick them.)

```
# nComb = nPerm/k! = n!/k!(n-k)!
factorial(12)/( factorial(6) * factorial(12-6) ) # 924
```

```
## [1] 924
```

g. How are your answers in parts e and f related?

The number of combinations is smaller than the number of permutations by a factor of k! since all the permutations of the k-mer sequence get collapsed into one combination.

4. Wnt signaling: Binomial proportions

Proliferation of embryonic stem cells is important for early development and is promoted by Wnt signaling, which promotes cell cycle progression through the transcriptional activator beta-catenin. In a study of the developing mouse, it was found that only 11% of cells in a region of the brain that is rich in neuronal progenitors was responsive to Wnt (even though Wnt is required for stem cell expansion).

Note: You may find the `choose()` function to be useful for parts of this question.

a. If you were to perform FACS on cells from this region that are labeled with a fluorescent antibody against the Wnt receptor, what is the probability of observing the following outcomes among a sample of 6 cells? Below, let's call Wnt-responsive cells = W and Wnt-insensitive cells = I.

```
W = 0.11
I = 0.89

# 1. A sequence of WWIIII
W*W*I*I*I*I
```

```
## [1] 0.007591811
```

```
# 2. A sequence of IIWWII
W^2 * I^4
```

```
## [1] 0.007591811
```

```
# 3. A string of WWWWWW
W^6
```

```
## [1] 1.771561e-06
```

```
# 4. At least one W cell out of 6
1 - I^6
```

```
## [1] 0.5030187
```

b. How many ways are there to get four Wnt-insensitive cells out of 6 total? (This is the same as the number of ways to get two W cells out of 6.)

```
choose(6,4)
```

```
## [1] 15
```

c. What is the total probability of getting exactly 4 I cells and 2 W cells in your sample, taking into account all the different ways that this outcome can be obtained?

```
choose(6,4)*W^2 * I^4
```

```
## [1] 0.1138772
```

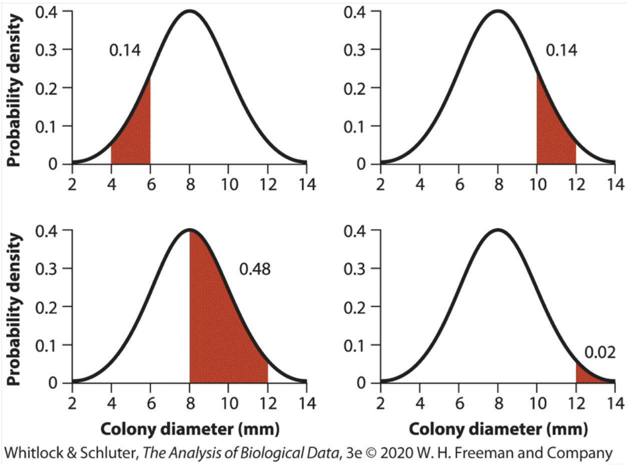
d. What would be the probability of getting 2 or less W cells in a sample? (Consider the probability of getting either 0, 1, or 2 W cells.)

```
choose(6,0) * W^0 * I^6 +
choose(6,1) * W^1 * I^5 +
choose(6,2) * W^2 * I^4
```

```
## [1] 0.9794064
```

5. Normal distribution: Probability density and cumulative probability

Answer the questions below, based on Question 31 from the Assignment Problems in Whitlock, Chapter 5. The questions are based on the diagrams in the accompanying figure, which has been uploaded to the Resources page on Ed in the Exercises section and embedded in the HTML version of this document:



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Figure 1: Normal density

The figure shows the probability density of colony diameters (in mm) in a hypothetical population of *Paenibacillus* bacteria. The distribution is continuous, so the probability of sampling a colony within some range of diameter values is given by area under the curve. Numbers next to the curve indicate the area of the region indicated in red. Consider the case in which a single colony is randomly sampled from the population.

a. Are the events "diameter is between 4 and 6" and "diameter is between 8 and 12" mutually exclusive? Explain.

b. What is the probability that a randomly chosen colony diameter is between 4 and 6 or between 8 and 12?

```
0.14 + 0.48
```

```
## [1] 0.62
```

c. What is the probability that a randomly chosen colony diameter is greater than or equal to 10?

```
0.14 + 0.02
```

```
## [1] 0.16
```

d. What is the probability that a randomly chosen colony diameter is between 8 and 10?

```
0.48-0.14
```

```
## [1] 0.34
```

e. What is the probability that a randomly chosen colony diameter is NOT between 4 and 12?

```
2*0.02
```

```
## [1] 0.04
```

```
1 - 2*0.48
```

```
## [1] 0.04
```
