1 Python

Here is an example of the sort of question that we plan to ask:

1. Write down the output that would be produced by the following snippet of Python code.

```python
s = 0
for i in range(0,3):
    s += i*i
print s
```

We would expect to see something like this as an answer:

Program prints "5".

You only need to know about Python constructs that have been used in the course, either in JEPy or in one of the labs. We went through the code in those projects and made a series of short snippets like the one above. Here are a few more, just to give you the flavor:

2. `print "a %s b %d:%3.1f" % ("hey",12,1.2)`
3. `def f(x, y):
   return x*y
print f(3,2)`
4. `x = 5
print "%d %s" % (x, x**"*")`
5. Here’s one that often comes in handy when doing calculations with gene genealogies:

```python
print sum([1.0/i for i in range(1,3)])
```

This last snippet is the easy way to calculate $\sum_{i=1}^{2} 1/i$. What would have happened if I had written 1/i instead of 1.0/i in the Python code?

6. from random import random

```python
samp_size = 100000000
n = 0
for i in xrange(0,samp_size):
    if random() < 0.7:
        n += 1
p = float(n)/samp_size
print p
```

The last snippet above is the longest one in our list.

2 Probability

7. Make a table showing the relative frequencies of the values in these data: [A, A, T, A, G].

8. Here is the probability distribution of two variables, X and Y:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Pr[X,Y]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>


9. Be able to manipulate the properties of expectations listed at the bottom of the first column of page 6, in JEPr. We’ll ask you for things like $E[4X]$, $E[Y/2]$, and $E[X + Y]$. You should be able to get such answers quickly, using the expectations you calculated in the previous item.

10. An urn contains 40 black balls and 60 red balls. You choose a ball at random, put it back, and choose another at random. What are the possible outcomes and their probabilities?

11. If we choose 2 students from this class at random without replacement, what is the probability that both are women?

12. Suppose that, in a class of 40 students, 20 are women. If we choose a student at random from the class, what is the probability that this student is a woman?

13. If we choose 2 students from this class at random without replacement, what is the probability that both are women?
14. Now you select 3 students at random with replacement. The number of men in this sample is a random variable, which may equal 0, 1, 2, or 3. What is the probability distribution of this random variable? (You may answer either by listing the probability of each outcome or by writing down a formula. Don’t bother with the calculator.)

15. JEPr gives three formulas for the variance: \( E[(X - E[X])^2] \), \( E[X^2] - E[X]^2 \), and \( E[X(X - E[X])] \). Be able to calculate a variance using all three.

16. The preceding item is about the variance as an expected value. But we also calculate variances from data, and in that case the variance is a statistic rather than an expected value. These are distinct quantities, even though we use the same word for them. Make sure you understand the difference between a statistic and an expected value, and make sure you can calculate either sort of variance.

17. JEPr also gives three formulas for the covariance: \( E[(X - E[X])(Y - E[Y])] \), \( E[XY] - E[X]E[Y] \), and \( E[X(Y - E[Y])] \). Be familiar with these too, both as statistics and as expected values.

18. JEPr discussed the following probability distributions: (1) binomial, (2) Bernoulli, (3) Poisson, (4) uniform, (5) exponential, and (6) normal. You’re expected to know the mean, variance, and distribution function of each distribution.

19. Under what circumstances would each distribution be plausible? For example, which distribution would you use to model the number of clicks emitted by a Geiger counter during some fixed interval of time? Suppose you spin a bottlecap 30 times and count the number of times it lands “concave-side-up.” Which distribution would best model that? Which distribution would make sense as a model of the length of a coalescent interval? (This last question won’t make sense until you’ve read the chapter on Gene Genealogies in Rogers’s lecture notes.)

20. For which distribution is the mean equal to the variance? For which is the mean equal to the standard deviation?

3 Population Genetics

21. Discuss the two inconsistent sets of definitions of “locus,” “gene,” and “allele.”

22. A locus with two alleles is called “biallelic.” Data from such a locus may be in the form of a list of genotypes like this: 


or in the form of genotype counts like this:

\[ \text{AA: 2, AT: 2, TT: 1} \]

Given either type of data, you should be able to calculate (1) genotype frequencies, (2) allele frequencies, and (3) genotype frequencies expected at Hardy-Weinberg equilibrium.

23. Define genetic drift. What events in the lives of real organisms contribute to drift?

24. Describe the Urn Model. What do the balls in the urn represent? What do the balls drawn from the urn represent? How does the urn model behave?

25. Under the influence of drift alone, heterozygosity \( (H) \) declines according to

\[ H_t = H_0(1 - 1/2N)^t \]

Here are a few examples of questions that we might ask about this formula:

(a) We might ask you to complete some step or steps of the derivation, so be sure you know how it goes.

(b) Describe it in plain English. What do the symbols \( H_t \), \( N \), and \( t \) mean?

(c) Genetic drift has its largest effect in small populations. How is this fact reflected in the formula?

26. At equilibrium between mutation and genetic drift, and assuming the infinite alleles model of mutation, the expected heterozygosity is

\[ H = \frac{4N\mu}{4N\mu + 1} \]

There are several kinds of questions that we might ask about this formula. For example:
(a) What model of mutation does this formula assume?

(b) We might ask you to complete some step or steps of the derivation, so be sure you know how it goes.

(c) What is the distinction between this formula and the formula $H = 2p(1 - p)$ for heterozygosity at Hardy-Weinberg equilibrium? Both formulas predict heterozygosity at equilibrium, so why are they different? (Hint: What is meant by “equilibrium” in the two contexts?)

(d) What is genetic drift, and how does it affect variation within and among populations?

(e) Ditto for mutation?

(f) How are these two effects captured by the equation above?

(g) The formula above predicts the average heterozygosity. Values at individual loci vary widely at any given time, from 0 up through levels of heterozygosity far in excess of the average. Why?

(h) Suppose that (a) the average heterozygosity equals 0.1, and (b) that $\mu = 10^{-6}$. What can you conclude about $N$? (Assume that the assumptions that went into the formula above are all correct.)