Stat 536 HW 1 Solutions

1. *In a population of a* 100 *individuals,* 25 *have genotype* AA *and* 75 *have genotype* aa *at a locus.* 5% *of the* AA *type are susceptible to anemia while only* 2% *of the* aa *type suffer from this condition. If an individual is picked at random from the population and she is susceptible to anemia, what is the probability that she has genotype* AA*?*

   Solution:
   Let AA denote the event of carrying an AA genotype and aa, the event of being genotype aa. Let CB denote the event of an individual being Color Blind.

$$
\begin{array}{rcl}
 & & P(AA) = 0.25 \qquad P(aa) = 0.75 \\
 & & P(CB \mid AA) = 0.05 \quad P(CB \mid aa) = 0.02 \\
P(AA|CB) & = & P(CB|AA)P(AA)/P(CB) \quad \text{(Applying Bayes' Rule)} \\
P(CB) & = & P(CB|AA)P(AA) + P(CB|aa)P(aa) = 0.05 * 0.25 + 0.02 * 0.75 = 0.0275 \\
P(AA|CB) & = & 0.0125/0.0275 = 0.455
\end{array}
$$

2. *A* 500 *bp long sequence has* 150 *purines and* 350 *pyrimidines. A point mutation at any position can cause a purine to pyrimidine change with probability* 1/4 *and pyrimidine to purine change with probability* 1/3*.*

   (a) If a purine is observed at a particular location, what is the probability that it was a purine before the last mutation event?

   (b) If a dinucleotide purine-purine is observed, what is the probability distribution of the four possible dinucleotide sequences that could have resulted in the observed dinucleotide. (Note: dinucleotide refers to two successive nucleotides in the sequence of interest)

   Solution:
   Assumptions:

   (a) Proportions of purines and pyrimidines given apply to the sequence *before* the last mutation.

   (b) Sites evolve independently.

   Let the event of observing a purine at a position be denoted by pu_{am} and of observing a pyrimidine be py_{am}. Let pu_{bm} and py_{bm} denote events of a nucleotide being a purine or a pyrimidine before mutation.
   We know that,

$$
\begin{array}{ll}
P(pu_{bm}) = 0.3 & P(py_{bm}) = 0.7 \\
P(py_{am} \mid pu_{bm}) = 0.25 & P(pu_{am} \mid pu_{bm}) = 0.75 \\
P(pu_{am} \mid py_{bm}) = 0.33 & P(py_{am} \mid py_{bm}) = 0.67
\end{array}
$$

(a)

$$P(pu_{bm} \mid pu_{am}) = \frac{P(pu_{am} \mid pu_{bm})P(pu_{bm})}{P(pu_{am})} \quad \text{(By Bayes' Rule)}$$
$$P(pu_{am}) = P(pu_{am} \mid pu_{bm})P(pu_{bm}) + P(pu_{am} \mid py_{bm})P(py_{bm})$$
$$= 0.75 * 0.3 + 0.33 * 0.7 = 0.225 + 0.233 = 0.458$$
$$P(pu_{bm} \mid pu_{am}) = 0.225/0.458 = 0.49$$

(b) As in part a. we can compute $P(py_{bm} \mid py_{am}) = 0.51$
Since sites evolve independently, the probability of a pu-pu dinucleotide arising out of a pu-pu dinucleotide is simply the product of the probability for each site. The probability distribution looks like this:

Before	After	Probability
pu-pu	pu-pu	$(0.49)^2 = 0.2401$
pu-py	pu-pu	$(0.49)(0.51) = 0.2499$
py-pu	pu-pu	$(0.51)(0.49) = 0.2499$
py-py	py-py	$(0.51)^2 = 0.2601$

3. *Mendel performed very simple early genetics experiments by crossing varieties of peas. The characteristics he selected (e.g. wrinkled vs. smooth peas) were obvious to the eye and turned out to be what we now call Mendelian traits (i.e. controlled by a single locus) with just two alleles, labeled A and a. In the first season, Mendel selected an AA plant and an aa plant. (In fact, he did not know the plants' genotypes, but selected plants based on phenotypes that had been selected by breeders for many generations.) He meticulously crossed AA with aa and collected the seeds. The next season he planted only the sampled seeds from the previous generation and crossed them with each other. Again he collected the resulting seeds.*

(a) *What is the sample space of possible genotypes for the seeds collected from the second generation?*
Solution:
Sample space is $\Omega = \{AA, Aa, aa\}$

(b) *Applying the* ***Law of Segregation****, derive the probability mass function for the second generation seed genotypes.*
Solution:

M/F	A	a
A	AA	Aa
a	Aa	aa

Therefore, the probability mass function can be written out as:

$$P(AA) = \frac{1}{4}$$
$$P(Aa) = \frac{2}{4}$$
$$P(aa) = \frac{1}{4}$$

(c) *For some traits, Mendel observed a* $3:1$ *ratio of phenotypes in the second generation. Explain how the phenotype pmf derives from the genotype pmf.*
Solution:
Homozygotes of the dominant allele and heterozygotes will all have the dominant phenotype resulting in a phenotype pmf. Let DP denote the dominant phenotype and dp denote the recessive phenotype. From part b above,

$$P(DP) = P(AA) + P(Aa)$$
$$P(dp) = P(aa)$$

Thus yielding the phenotype pmf of

$$P(DP) = \frac{3}{4}$$
$$P(dp) = \frac{1}{4}$$

corresponding to the $3:1$ ratio observed by Mendel.

(d) *To show that the larger class actually consisted of two distinct types of plants, Mendel sampled 100 plants with the dominant phenotype, selfed them (crossed these plants with themselves) and typed 10 offspring. What is the probability that all 10 offspring have the dominant phenotype?*
Solution:
Let DP denote the event of an offspring showing dominant phenotype and dp be the event that the offspring shows the recessive phenotype. Also, let AA denote the even that the parental genotype is AA and Aa denote the event that the parental genotype is Aa.
From the previous part we know that $P(AA) = 0.33$ and $P(Aa) = 0.67$.

$$P(DP|AA) = 1 \quad P(DP|Aa) = 0.75 \quad \text{(from part b,phenotype pmf of 3:1)}$$

Let 10DP denote the event that all 10 offspring have dominant phenotype DP

$$P(10DP) = P(10DP|AA) + P(10DP|Aa)$$
$$= (1^{10})(\frac{1}{3}) + (0.75^{10})(\frac{2}{3}) = 0.3708757$$

(e) *If all 10 offspring shared the dominant phenotype, Mendel called the parent type I, else type II. He theororized (hypothesized) that the ratio of type I to type II was 1:2. He was actually slightly wrong. Why?*
Solution:
From the previous part we get a ratio 1 : 1.7 which is slightly lesser than the 1 : 2 predicted by Mendel and this is because Mendel did not take into account the small probability of Aa parents being classified as type I .Since 10 is a small sample number, this probability (0.056) contributes significantly.

(f) *Test whether the following data shows evidence against the correct ratio (not exactly 1:2).*

Type I	*Type II*
28	*72*

Use either a t-test or a chi-square test. There is not enough evidence to reject the null hypothesis that there is no difference between the two ratios 1 : 1.7 *and* 1 : 2. Solution:
The data has a Binomial sampling distribution with proportion of Type I p. We test $H_0 : p = 0.371$ against $H_A : p \neq 0.371$. We estimate $\hat{p} = \frac{28}{100}$. Assuming, $\hat{p}$ has an approximate normal distribution (by CLT), then the approximate z-statistic, assuming H_0 and letting $p_0 = 0.371$, is

$$z = \frac{\hat{p} - 0.371}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and the p-value for the two-tailed test is

```
> d <- c(rep(1, 28), rep(0, 72))
> p.0 <- 1/3+(3/4)^10*2/3
> p.hat <- mean(d)
> z <- (p.hat - p.0)*sqrt(n)/sqrt(p.0*(1-p.0))
> 2*pnorm(z)
[1] 0.05992689
```

We may not be comfortable with the normal assumption used in computation of z. Instead, we could use a t test

```
> n <- length(d)
> t <- (p.hat - p.0)*sqrt(n)/sd(d)
> 2*pt(t, df=n-1)
[1] 0.04674232
```

although it still achieves its sampling distribution asymptotically. So far, we have conflicting conclusions. Still uncomfortable about assumptions, we could compute probabilities using the binomial sampling distribution directly

```
> 2*pbinom(28, size=100, prob=p.0)
[1] 0.07159217
```

in which case we would not reject the H_0 at a traditional type I error rate of $\alpha = 0.05$.

Finally, one of you wanted to see a LR test applied to this problem.

```
> lr <- p^28*(1-p.0)^72 / p.hat^28 / (1-p.hat)^72
> pchisq(-2*log(lr), df=1, lower.tail=F)
[1] 0.05478934
```

and we reject again.

[Please note, all calculations were done before rounding, though they are written rounded here.]

4. *You have discovered a new gene, 768 base pairs long, that encodes for a protein you think interacts with the HIV virus and makes people more or less susceptible to the disease. You sample 11 individuals from an African population where the virus has been prevalent and active for a long time. Your excitement about a potential discovery increases when you detect variation, i.e. multiple alleles, in your sample. You hypothesize that though HIV is a relatively new human pathogen, that it has exerted extreme selection on the population you study. If mutants of this gene can make people less susceptible to the disease, then you theororize that these mutations will have been highly selected in the population you study. Looking more closely at your data, you observe that 14 of the 768 sites in the gene are segregating two nucleotides and 10 of these mutations are nonsynonymous. Use the genetic code to formulate a simple model that assumes there has bee! n no selection (assume selection acts at the protein level only). Use your model to test whether there is evidence of selection acting on this gene. Critique your model.* Solution:

Assume that all 61 non-stop codons are equally likely in the protein. Assume that all mutations, except those resulting in stop codons, are equally likely. Assume that all 14 mutated sites are in separate codons.

I wrote a small perl script to compute the probability of a nonsynonymous change given all of the above assumptions. It is

$$p_N \approx 0.745$$

Many of you were even more draconian and assumed of the 9 possible mutations at a codon, 3 were synonymous (the 'wobble' position) and 6 nonsynonymous. This assumption lea ds to $p_N = 2/3$, which is not terribly far from the more exact calculation above.

As for problem 3, the data is binomially distributed, and I compute the p-value using `R`'s `pbinom`. For a one-sided test, i.e. testing that nonsynonymous mutatio ns are over-represented in the data, the p-value is 0.505.

The model is foolhardy for many reasons. First, there is no protein in which codons are equally likely. Second, it is well known that transitions are more likely than transvers ions, and in fact, models that consider different rates for all possible mutations tend to be supported when there is enough data. Third, it is foolish to test the null hypothesis of no selection becau se if this gene encodes a functioning protein, there *must* be extensive purifying selection (i.e. selection against nonsynonymous change) so that the protein can retain its function. In general , a surprising abundance of nonsynonymous mutations can suggest selection, but the number of nonsynonymous changes must be evaluated relative to other proteins or a much more sensible null model of prot ein evolution. In fact, observing 10/14 nonsynonymous mutations is *very* unusual and would be highly suggestive of something strange governing the evolution of this protein.