Handout 1

1.

Cell type	A	C	G	T	Total
Nonpermissive cells (n)	17	43	39	1	100
Permissive cells (p)	24	72	49	5	150
Difference (n - p)	-7	-29	-10	-4	

(a) Let $n_{n\mathrm{N}}$ and $n_{p\mathrm{N}}$ be the counts of nucleotide N in nonpermissive and permissive cells, respectively. Assume (incorrectly) that the row totals are equal, and write an algorithm to test

$$H_0^{(\text{vague})} : \text{C is not mutated}$$

using test statistic $n_{n\mathrm{C}} - n_{p\mathrm{C}}$.

(b) Let n_n and n_p the total number of nucleotides observed from nonpermissive and permissive cells. To account for a difference in these, note that if $n_p > n_n$, there are $n_p - n_n$ *additional* chances to observed nucleotide C in the nonpermissive cells. Another way to say this: the *ratio* of expected number of C in permissive cells over the expected number of Cin nonpermissive cells should $\frac{n_p}{n_n}$. Use this fact to propose model of the following form involving a single unknown parameter λ.

$$\begin{aligned} N_{n\mathrm{C}} &\sim \text{Poisson}(?) \\ N_{p\mathrm{C}} &\sim \text{Poisson}(?) \end{aligned}$$

(c) Estimate λ from the data (and we'll correct the p-value computed in Part (a)).

(d) The data are actually simulated without an actual signal, so the variation we observe is noise. Why might we be getting a fairly significant result (fairly small p-value)?

2. Suppose you observe the following region 50 bp upstream of the transcript start site of a gene.

```
GCATTGGCCACACATATAAACGGTAGTCAACGTAGGTAACAGAGTCTCGA
  ---    - - -------   -- - --  --  --- - - - -  -
```

Highlighted are the A/T nucleotides. You can observed there is one longer stretch lasting 7 bp. Is this unusual?

(a) What test statistic is sensitive to the vague null hypothesis that there is nothing unusual about this sequence?

(b) What specific null hypothesis could you use to simulate values of the test statistic?

(c) Plan an algorithm:

3. You have been studying a protein that binds DNA, and you know many genomic sites where this protein binds. The binding site is about 7 bp long, with some positions highly predictable. By comparing all known binding sites, you find the following probabilities for each nucleotide at the 7 positions:

	Site						
	1	2	3	4	5	6	7
A	0	0	1	0.333	0.167	1	0
C	0	0	0	0.333	0	0	1
G	1	0	0	0	0	0	0
T	0	1	0	0.333	0.833	0	0

Your goal: identify additional binding sites in the promoters of other genes using purely computational methods.

(a) If $p_{i\mathrm{N}}$ is the probability that base N binds to site i, for example $p_{5\mathrm{A}} = 0.167$, argue that if you observe sequence $\boldsymbol{X} = (X_1, X_2, \cdots, X_7)$, where $X_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, in a promoter, that

$$T(\boldsymbol{X}) = \prod_{i=1}^{7} p_{iX_i}$$

is a statistic and it is sensitive to the vague null that $\boldsymbol{X}$ is not bound by your protein. What values suggest that $\boldsymbol{X}$ *is* bound by the protein?

(b) What specific null hypothesis H_0 could you use to simulate data to determine whether an observed test statistic t_0 is unusual?