Module 4F: Hypothesis Testing

Revisiting Quantitative evidence & uncertainty

Agenda:

- Poisson Distribution of the four steps in Hypothesis Testing
  1. H_O/H_A: Our model of the test universe (the distribution of the variable)
  2. Test & assumptions: are the assumptions met? Is the test valid?
  3. Quantitative evidence: **p-value**, or critical value.
  4. Conclusion & uncertainty/estimation

- Fisher's Exact Test (McDonald-Kreitman)

<u>Fisher's Exact Test:</u>

- 2 x 2 contingency analysis
  – based on **hypergeometric** distribution with four classes

  –Answers the question: given two-way tables with the same fixed margin totals as the observed one, what is the chance of obtaining the observed cell frequencies *a,b,c* and *d and all cell frequencies that represent a greater deviation from expectation? NOTE:* This is a similar to the definition (and spirit) of the P-value (not a coincidence, since Fisher also invented that)!

- No assumptions about size of expectations

- cumbersome to do it by hand (use R)
  > fisher.test(matrix-data)

The total number of ways in which a two-way table with fixed marginal totals can be obtained is:

$$\binom{n}{a+b}\binom{n}{a+c} = \frac{n!}{(a+b)!(c+d)!} \cdot \frac{n!}{(a+c)!(b+d)!}$$

Leads to the probability of obtaining a 2x2 table with the frequencies a,b,c and d:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Example: All but 28 trees of two species of acacia, species A and B, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next 16 different colonies of ants from species X from an area nearby, were brought in and placed equidistant from the 28 acacia trees. The ant colonies had been harvested from cut-down trees of species A.

Example: All but **28** trees of two species of acacia, species **A** and **B**, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next **16** different colonies of ants from species **X** from an area nearby, were brought in and placed equidistant from the **28** acacia trees. The ant colonies had been harvested from cut-down trees of species **A**.

a	b	a+b
<u>c</u>	<u>d</u>	<u>c+d</u>
a+c	b+d	a+b+c+d

Example: All but **28** trees of two species of acacia, species **A** and **B**, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next **16** different colonies of ants from species **X** from an area nearby, were brought in and placed equidistant from the **28** acacia trees. The ant colonies had been harvested from cut-down trees of species **A**.

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

Species	Not Invaded	Invaded	Total
A	2	13	15
B	10	3	13
Totals	12	16	28

Example: All but **28** trees of two species of acacia, species **A** and **B**, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next **16** different colonies of ants from species **X** from an area nearby, were brought in and placed equidistant from the **28** acacia trees. The ant colonies had been harvested from cut-down trees of species **A**.

1	14	15
11	2	13
12	16	28

0	15	15
12	1	13
12	16	28

Species	Not Invaded	Invaded	Total
A	2	13	15
B	10	3	13
Totals	12	16	28

Example: All but **28** trees of two species of acacia, species **A** and **B**, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next **16** different colonies of ants from species **X** from an area nearby, were brought in and placed equidistant from the **28** acacia trees. The ant colonies had been harvested from cut-down trees of species **A**.

Species	Not Invaded	Invaded	Total
A	2	13	15
B	10	3	13
Totals	12	16	28

> fisher.test(acacia)

Fisher's Exact Test for Count Data

data: acacia

p-value = 0.001624

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

 0.003786123 0.425475250

sample estimates:

odds ratio

0.05401494

McDonald- Kreitman Test

- The McDonald- Kreitman (MK) test is a widely used test in population genetics for detecting selection by comparing patterns of polymorphism within species to divergence between species. It aligns sequences of multiple species of an organism and counts the number of two categories of mutation between them. The first category is **synonymous sites**: meaning that due to the degeneracy of the genetic code the nucleotide substitution does not result in a different amino acid. The second category is **non-synonymous sites**: a change in nucleotide does result in the specification of a different amino acid.

- Fisher's exact test can be applied within the MK framework to **assess whether there is a significant excess of non-synonymous substitutions *within* species (polymorphisms) compared to *between* species (fixed differences), which would suggest positive selection acting on the gene of interest.**

- The underlying Ho hypothesis is: "In the absence of natural selection, the ratio of synonymous to nonsynonymous sites should be the same for polymorphisms (PN/PS) and fixed differences (DN/DS)." (Neutrality)

Migration, incomplete lineage sorting, or parallel mutation events can also result in polymorphic sites
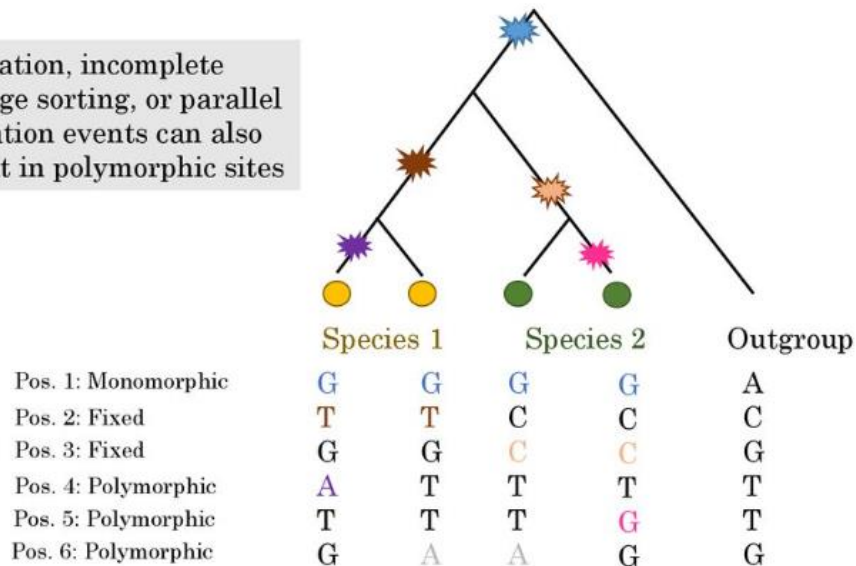
P_N/P_S = DN/DS -> Neutrality

P_N/P_S > DN/DS → Negative/Purifying Selection

P_N/P_S < DN/DS → Positive Selection

McDonald- Kreitman Test

- Fisher's exact test can be applied within the MK framework to **assess whether there is a significant excess of non-synonymous substitutions *within* species (polymorphisms) compared to *between* species (fixed differences), which would suggest positive selection acting on the gene of interest.**

Step 1: The underlying Ho hypothesis is: "In the absence of natural selection, the ratio of synonymous to nonsynonymous sites should be the same for polymorphisms and fixed differences."

Ho: 2/43 = 7/17; under neutrality $D_N/D_S = P_N/P_S$

Step 2:

	Synonymous Polymorphism	Non-synonymous Polymorphism
Fixed Difference	17 = D_S	7 = D_N
Polymorphism	43 = P_S	2 = P_N

Step 3:

Cheated and ran it through R using fisher.test()

Step 4:

This suggests that the null hypothesis of independence

between mutations

P_N/P_S > DN/DS → **Negative/Purifying Selection**

```
Fisher's Exact Test for Count Data

data:  mk
p-value = 0.006653
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.437432 92.388001
sample estimates:
odds ratio
  8.540913
```