Module 4B : Hypothesis Testing

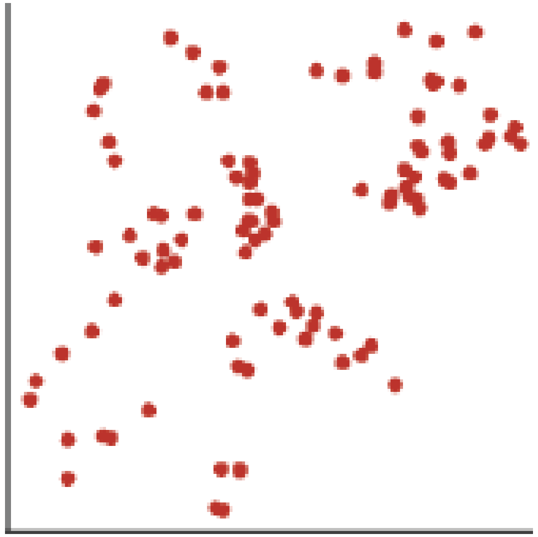
Revisiting Quantitative evidence & uncertainty

Agenda:

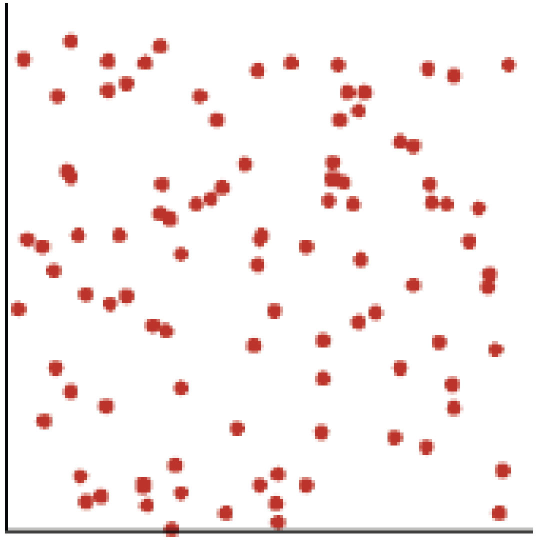
- Poisson Distribution of the four steps in Hypothesis Testing
  1. H_O/H_A: Our model of the test universe (the distribution of the variable)
  2. Test & assumptions: are the assumptions met? Is the test valid?
  3. Quantitative evidence: **p-value**, or critical value.
  4. Conclusion & uncertainty/estimation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

Fitting the Poisson Distribution:

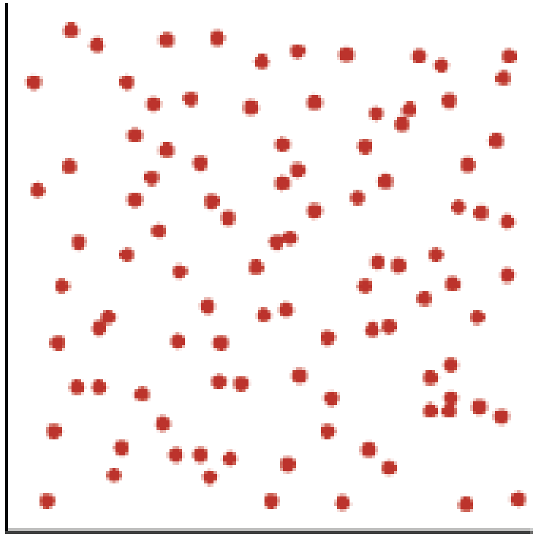
The Poisson Distribution describes the probability of getting X successes in a block of time or space when the successes happen independently of each other and occur with equal probability at every point in time or space.



Clumped



Random



Dispersed

Poisson Distribution:

$$P[X] = \frac{e^{-\mu}\mu^{X}}{X!}$$

Example: Mass extinctions random or concentrated in periods of time? Fossil Marine invertebrates' families' extinctions in 76 blocks of time of similar duration (Raup Sepkoski, 1982).

If extinction is random, then the number of extinctions per block of time will be Poisson.

If not, then they could be either clumped or dispersed.

Num Extinctions (X)	**Frequency**
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1
12	**0**
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1
>20	0
Total	76

Step 1

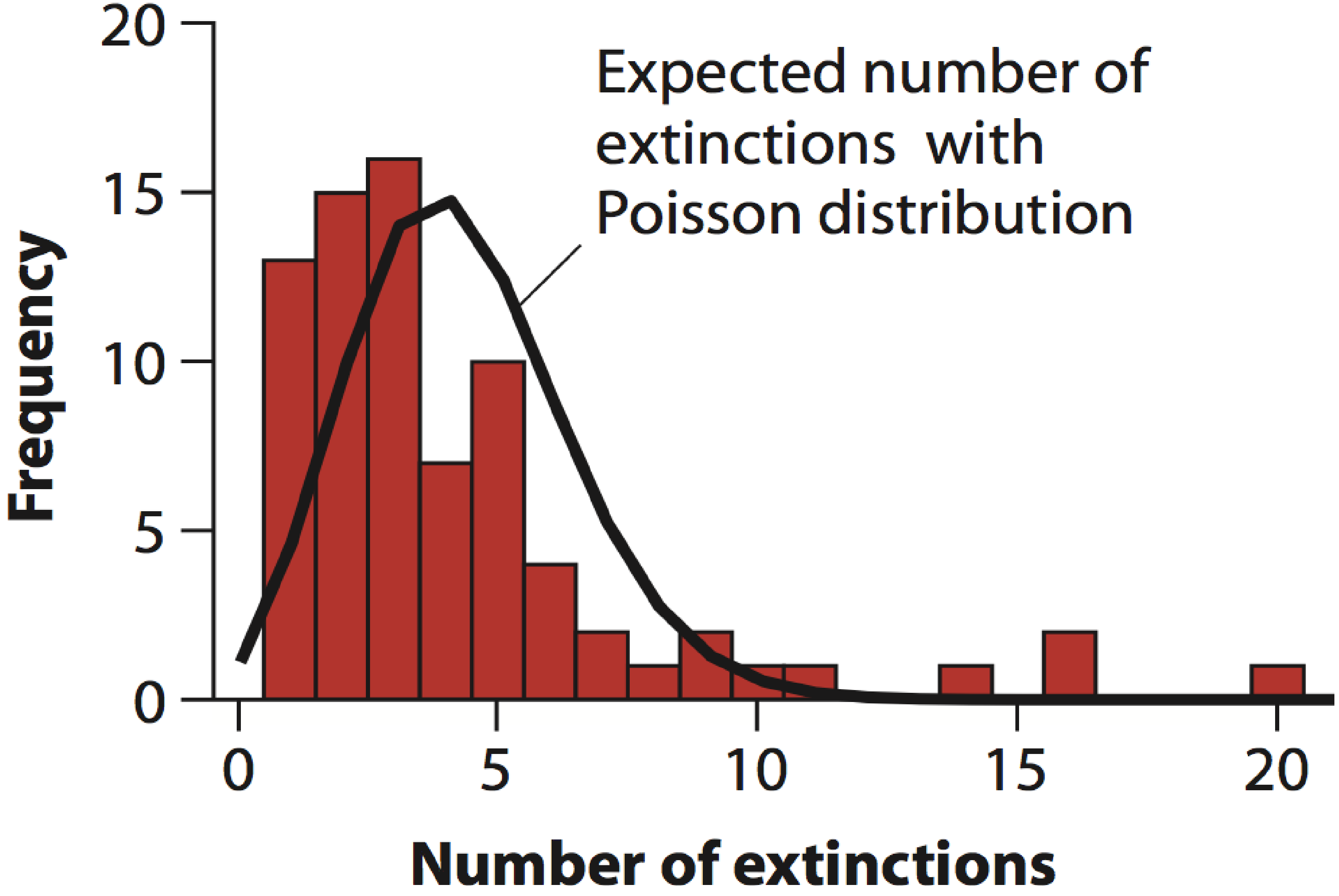
H_0: The number of extinctions per unit of time has a Poisson distribution

H_A: The number of extinctions per unit of time does *NOT* have a Poisson distribution

Step 2

Estimate μ:

$$\bar{X} = \frac{(0X0)+(13X1)+(15X2)+\ldots}{76} = 4.21$$



Num Extinctions(X)	Observed Frequency	Expected Frequency
0	0	1.13
1	13	4.75
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
8	1	2.76
9	2	1.29
≥10	6	0.86
Total	76	76

Num Extinctions(X)	Observed Frequency	Expected Frequency
0 or 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
>8	9	4.91
Total	76	76

$$\chi^2 = \frac{(13-5.88)^2}{5.88} + \frac{(15-10.00)^2}{10.00} + \ldots = 23.93$$

Step 3:

DoF = 8 categories **– 1 – 1 estimate = 6** degrees of freedom

Warning:

** When you 're-bin' your data to ensure that the assumptions of the χ^2 gof test are met, you might need to update your degrees of freedom since they are based on the number of categories!*

Step 4:

Critical value for χ^2 is given in statistical table found at: https://www.math.arizona.edu/~jwatkins/chi-square-table.pdf In fact, P-value < 0.001. Therefore, we can reject the null hypothesis and conclude that the extinction record for these fossils do not fit a Poisson distribution. **BUT THERE IS MORE WE CAN SAY….**

Variance = Mean:

If Variance > Mean, then CLUMPED

- visual hint: histogram is ‘u-shaped’

If Variance < Mean, then DISPERSED

- points are spread uniformly in space or time

- This may be a bit confusing if you are familiar with molecular genetics, because we refer to the “over dispersed molecular clock” which is really saying that variance > mean number of substitutions. Sometimes, terminology is ambiguous!

$$\chi^2 = \frac{(13-5.88)^2}{5.88} + \frac{(15-10.00)^2}{10.00} + \ldots = 23.93$$

Critical value for χ^2 is given in statistical table as 15.507
In fact, P -value < 0.001. Therefore, we can reject the null hypothesis and conclude that the extinction record for these fossils do not fit a Poisson distribution.

Since the sample variance is 13.72, we can also say that not only do we reject the null hypothesis that extinction patterns follow the Poisson distribution (and so we can reject that they occur randomly), we can also say that extinction events are clumped

Rejecting a null hypothesis of a Poisson distribution of successes implies that

A- Success are not independent

B- The probability of a success occurring is constant over time or space.

C-The probability of a success occurring is NOT constant over time or space.

D- A and B

E- A and C

Contingency Analysis

Contingency: *allows us to determine if two categorical variables are associated (some contingency tests will allow us to quantify the degree of association as well but not all do this).*

Major tests:

- **χ^2 Contingency Test** –> similar but not exactly as the same χ^2 Goodness of fit test. You can think of it as a subset of χ^2 Goodness of fit tests with some calculation differences. Basis of test is Multiplication rule with the assumption of independence. Degrees of freedom are calculated differently!
- **Odds ratio** –> Ho: OR=1. Challenge: transforming the sampling distribution of OR so that it is normally distributed.
- **Relative** Risk -> like OR but accounts for proportion of (rare) event in the population
- **Fisher’s Exact test** -> exact calculation. You can think of it as the contingency version of calculating a p-value

Contingency Analysis:

Review prompt: Associations between categorical variables

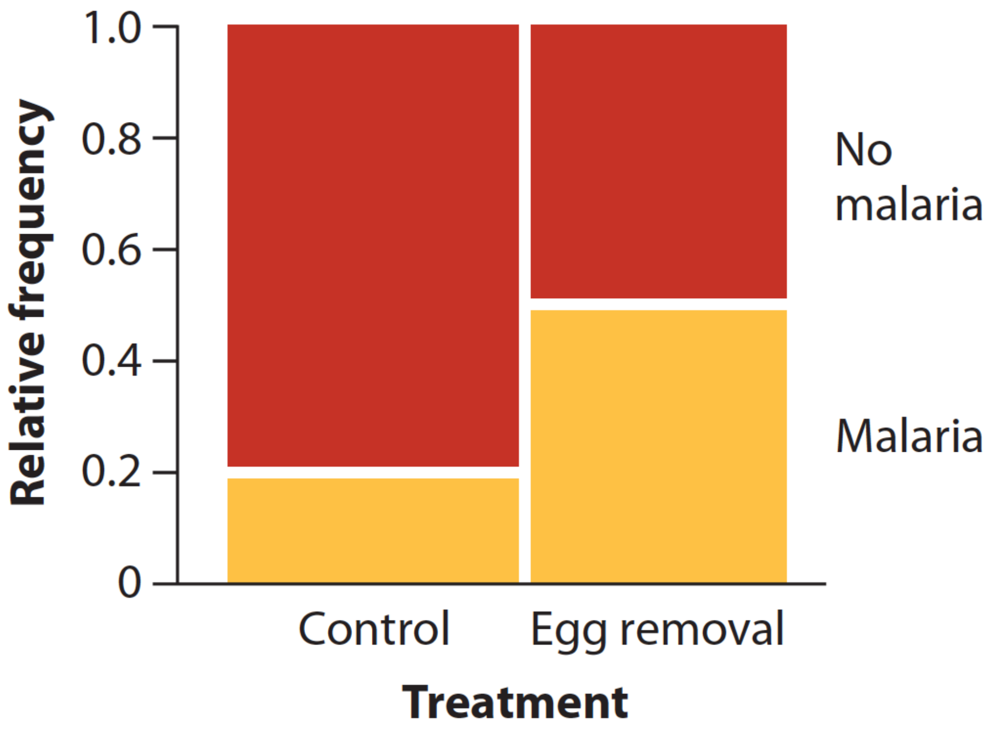
- Test the independence of two or more categorical variables

	Control Group	Egg-Removal Group	Row Total
Malaria	**7**	**15**	**22**
No Malaria	**28**	**15**	**43**
Column Total	**36**	**30**	**65**

Contingency Analysis:

- Associations between categorical variables
- Test the independence of two or more categorical variables

	Control Group	Egg-Removal Group	Row Total
Malaria	**7**	**15**	**22**
No Malaria	**28**	**15**	**43**
Column Total	**36**	**30**	**65**



Reminder: Multiplication Rule

Multiplication rule: P[A and B] = P[A|B]P[B]

IFF INDEPENDENT, this collapses to: P[A and B] = P[A]P[B]

χ^2 Contingency Test:

- Tests goodness-of-fit to the data of the null hypothesis of independence of variables

- Two categorical variables but, unlike the Odds Ratio, each variable can have more than 2 categories

- Assumptions:
  - The value of the cell ***expected values*** should be 5 or more in at least 80% of the cells
  - No cell should have an **expected value** of less than one

- Description of χ^2 Contingency Test:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

A chi-squared test statistic in a test of a contingency table that is equal to zero means:

A. The two nominal variables have values consistence with independence.
B. The two nominal variables have values that are consistent with equality.
C. The two nominal variables have the same proportions listed in Ho.
D. All these choices.

When is it appropriate to use Chi-Squared tests?

----

a. When you are determining if two categorical variables are associated.

b. When you are directly comparing proportions

c. When your number of independent data points is less than 5

d. When you are looking for an exact P value.

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	**3220**
No Cancer	1422	7325	1092	406	**10245**
Column Total	**1742**	**9542**	**1555**	**626**	**13465**

STEP 1: Formulate null hypothesis

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

Step 1:

H_0: The development of breast cancer is ***independent*** of the age at first birth

H_A: The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence. Assumptions: no cells less than 5 so both assumptions are met.

With independence,

P[Age at first birth AND breast cancer] = ?

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	**3220**
No Cancer	1422	7325	1092	406	**10245**
Column Total	**1742**	**9542**	**1555**	**626**	**13465**

Step 1:

H_0: The development of breast cancer is ***independent*** of the age at first birth

H_A: The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence

With independence,

P[Particular Age at first birth AND breast cancer] = P[Particulat Age at first birth]P[Breast cancer]

Calculating the expectations under H_0:

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	**3220**
No Cancer	1422	7325	1092	406	**10245**
Column Total	**1742**	**9542**	**1555**	**626**	**13465**

$$P[Age < 20 Birth] = \frac{1742}{13465} = 0.13$$

$$P[Cancer] = \frac{3220}{13465} = 0.24$$

$$P[NoCancer] = \frac{10245}{13465} = 0.76$$

If H_0 is true, then:

P[< 20 Age at first birth AND breast cancer] = 0.13*0.24 = 0.031

Calculating the expected **COUNTS** under H_0:

EXPECTED values Under Ho	<20	20-29	30-34	>=35	Row total
Cancer	**416.6** 320	**2281.9** 2217	**371.9** 463	**149.7** 220	3220
No Cancer	**1325.6** 1422	**7260.2** 7325	**1183.2** 1092	**477** 406	10245
Column Total	1742	9542	1555	626	13465

χ^2 Contingency Test

Step 2:

$$\chi^2 = \sum_{i} \frac{(Observed_i - Expected_i)^2}{Expected_i} = 104.76$$

$$= \frac{(416.6-320)^2}{416.6} + \frac{(2281.9-2217)^2}{2281.9} + \frac{(371.9-463)^2}{371.9} + \frac{(149.7-220)^2}{149.7} + \frac{(1325.6-1422)^2}{1325.6} + \frac{(7260.2-7325)^2}{7260.2} + \frac{(1183.2-1092)^2}{1183.2} + \frac{(477-406)^2}{477}$$

Step 3:
Degrees of Freedom:

dof = (row - 1)(column - 1)

For the Birth age/cancer example: **dof = (2-1)(4-1)=3**

Step 4, Conclusion:

$$\chi^2 = 104.76 \quad >> \quad \chi_3^2 = 7.81$$

We reject the null hypothesis of independence with a significance level of >= 0.05 and say that the age of first birth was not independent on whether breast cancer eventually developed.

What would a chi-square contingency test resulting in a significance value of P > 0.05 suggest?

A. We cannot reject the hypothesis of independence between the two variables

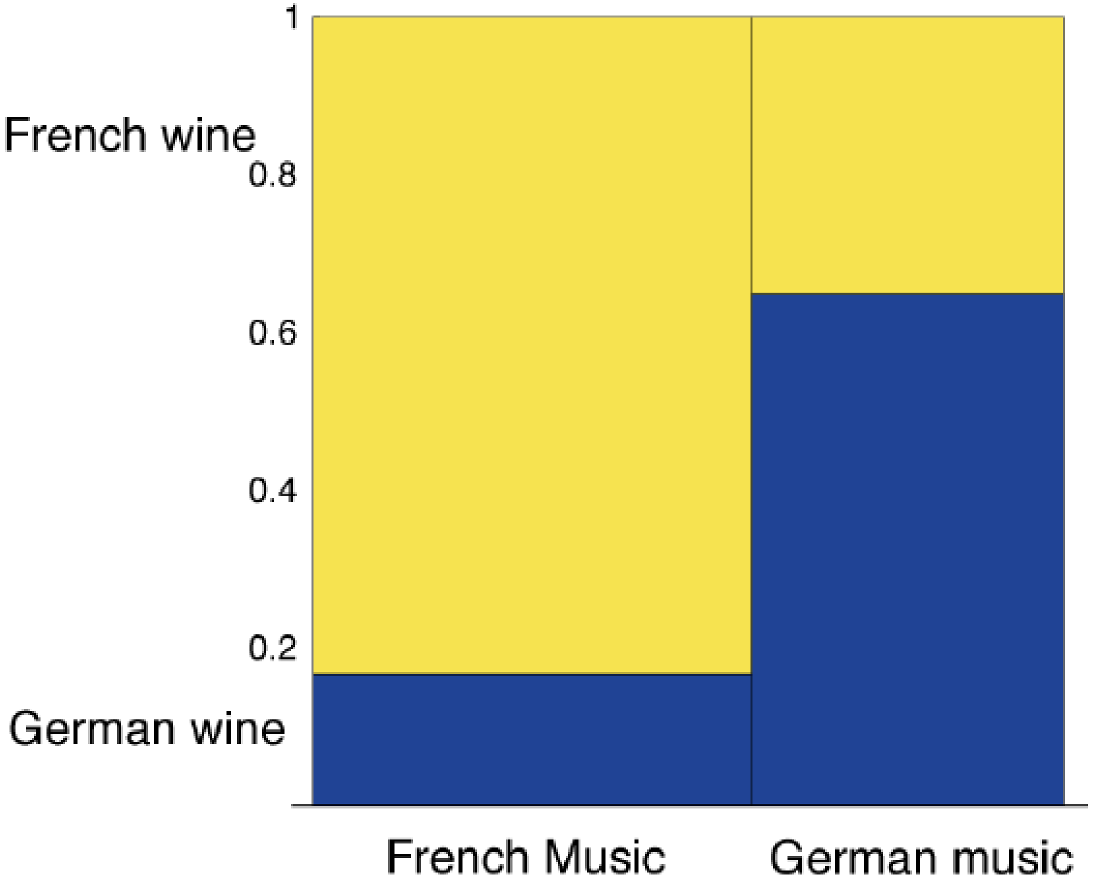
B. We cannot reject the hypothesis of dependency between the two variables

C. There is a significant relationship between the two variables

D. We can reject the hypothesis of dependency between the two variables

Example: Does the nationality of background music effect the nationality of wine that is bought?

Observed	**French Music**	**German Music**	**Row Totals**
Bottles of French Wine	40	12	**52**
Bottles of German Wine	8	22	**30**
Column Totals	**48**	**34**	**82**



Example: Is there an influence of the following three SES on preterm delivery rates?

Socio-Economic status	Preterm Birth	Normal Birth
Upper/Upper-middle	25	85
Middle	33	64
Lower/Lower-middle	112	149

A. Yes, we reject the null hypothesis
B. No, we fail to reject the null hypothesis
C. Yes, we fail to reject the null hypothesis
D. No, we reject the null hypothesis

Fisher's Exact Test:

- 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:

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

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**
c	**d**	**c+d**
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**

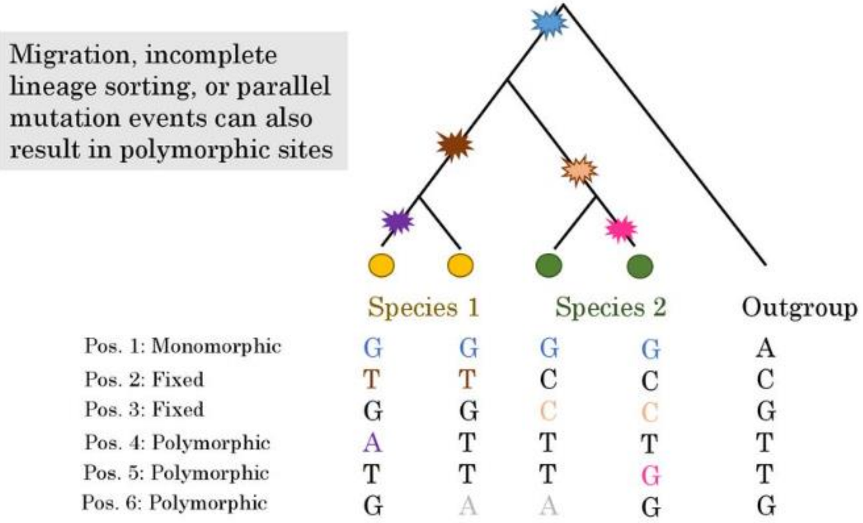
Example: All but **28** trees of two species of acacia, species **A** and **B**, were cleared from an area in Central America. These trees were un-infested (no ant colonies). Next **16** different colonies of ants from species **X** from an area nearby, were brought in and placed equidistant from the **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 that: ” In the absence of natural selection, the ratio of synonymous to nonsynonymous sites should be the same for polymorphisms and fixed differences.”



Migration, incomplete lineage sorting, or parallel mutation events can also result in polymorphic sites

	Species 1		Species 2		Outgroup
Pos. 1: Monomorphic	G	G	G	G	A
Pos. 2: Fixed	T	T	C	C	C
Pos. 3: Fixed	G	G	C	C	G
Pos. 4: Polymorphic	A	T	T	T	T
Pos. 5: Polymorphic	T	T	T	G	T
Pos. 6: Polymorphic	G	A	A	G	G

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 that: ” 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 $F_N/F_S = P_N/P_S$

Step 2:

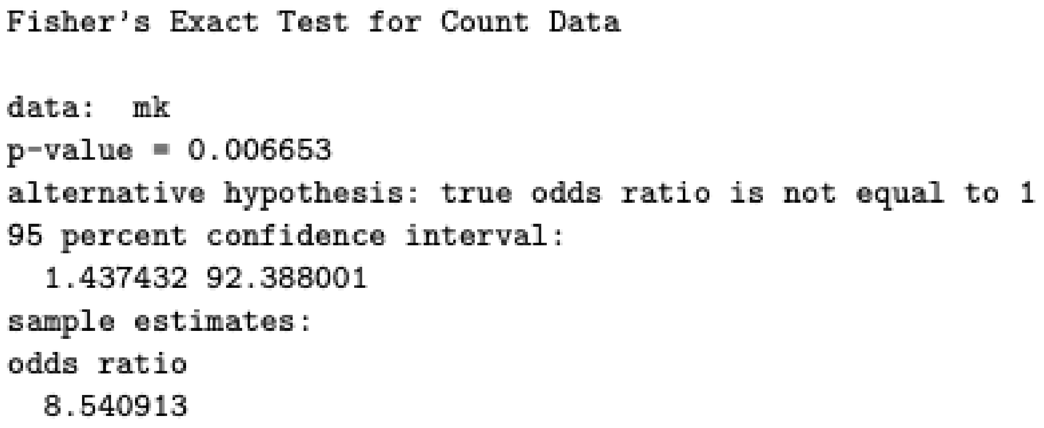
	Synonymous Polymorphism	**Non-synonymous Polymorphism**
Fixed Difference	17 = F_S	7 = F_N
Polymorphism	43 = P_S	2 = P_N

Step 3:

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

Step 4:

This suggests that the null hypothesis of independence between mutations



```
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
```

Upload to Canvas as part of your Certification of Completion

Module 4B Questions:

Finish up and submit the questions we covered in lecture (in orange).