Module 4D: 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)

Your pipeline for hypothesis testing in statistics

Step 1

Formulate your null **hypothesis**

- Null hypothesis is ***only*** *hypothesis that is tested*
- Falsification: *want* to reject your null

Step 2

Identify appropriate **test statistic**

- Assumptions of your test

Step 3

Quantify the results of your test

- **P value** or comparison to **critical values**
- How *unusual* is your data?

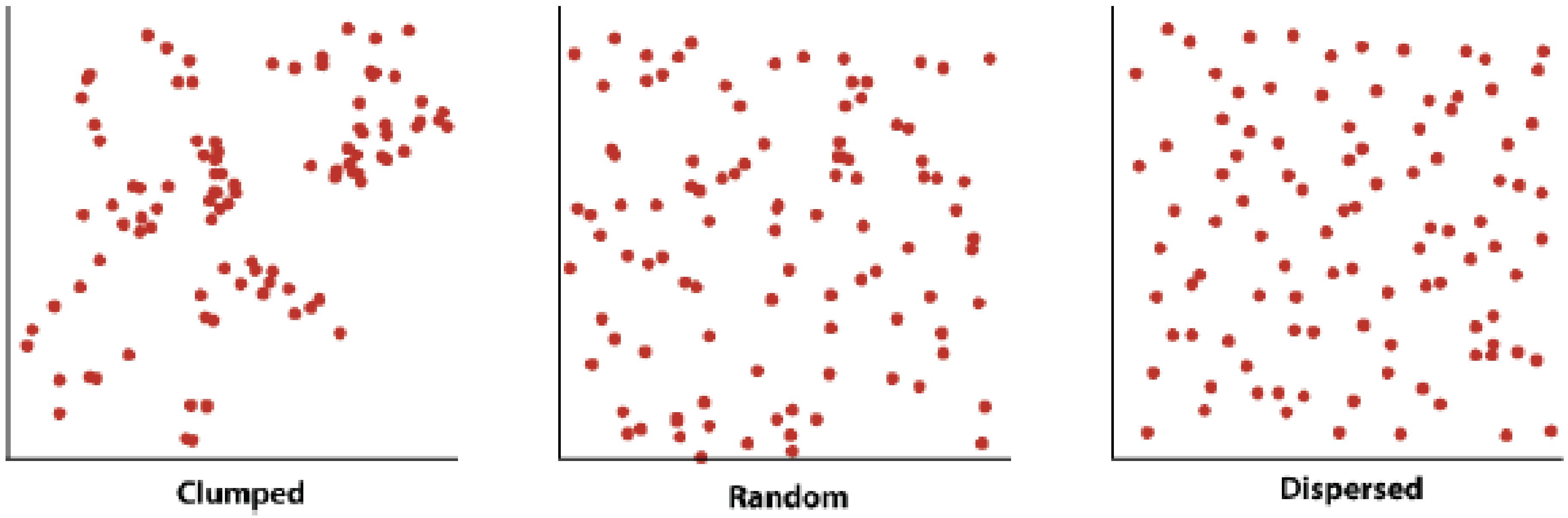
Step 4

Conclude: reject or fail to reject

- based on alpha value
- if appropriate, confidence interval of the parameter

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.



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

Num Extinctions (X)	**Frequency**
12	**0**
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1
>20	0
Total	76

Step 1

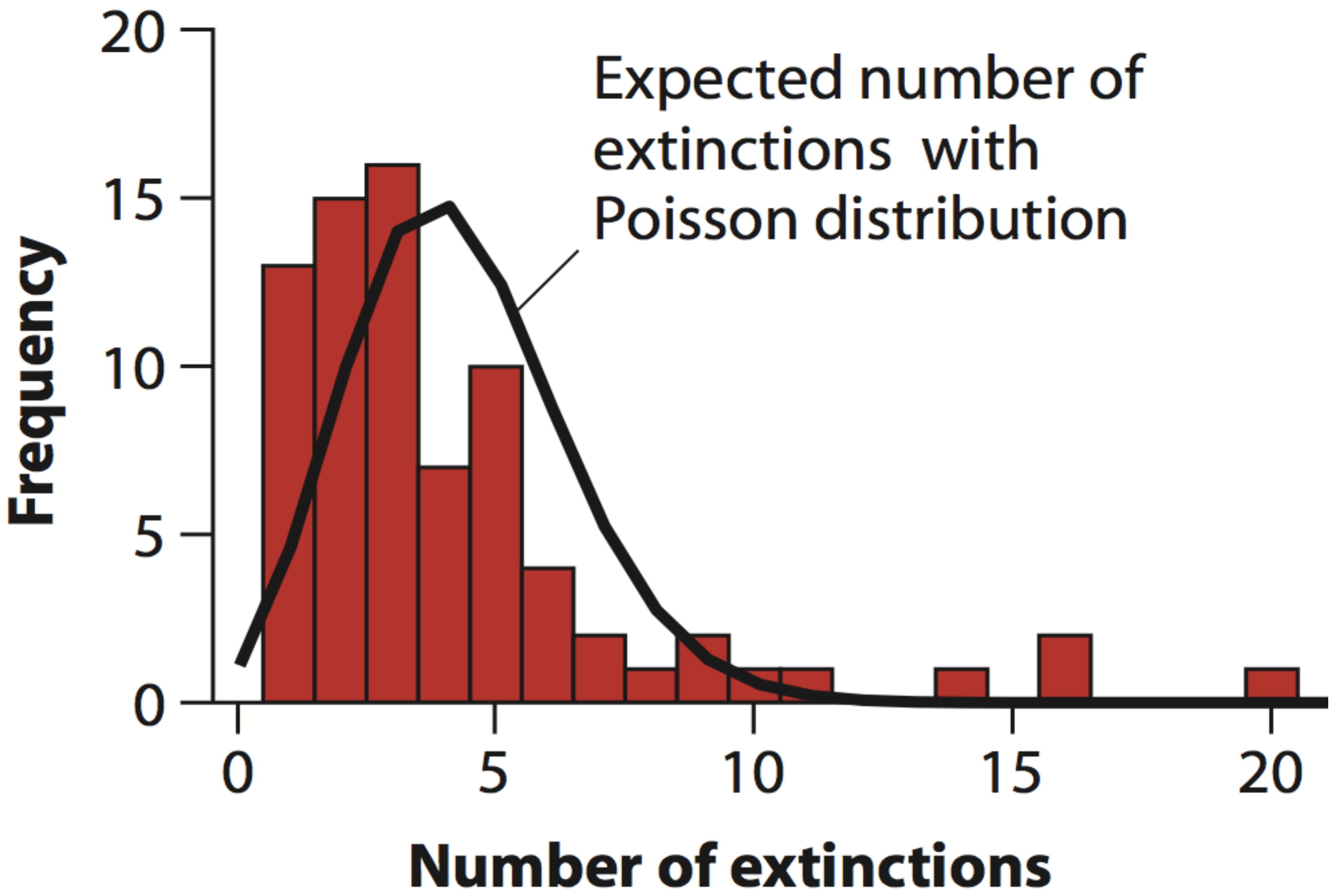
H_0: The number of extinctions per unit of time has a Poisson distribution

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

Step 2

Estimate μ:

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



https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html

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

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

<u>Step 3:</u>

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

<u>Warning:</u>

* *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!*

<u>Step 4:</u>

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 <u>do not</u> 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 12.592
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