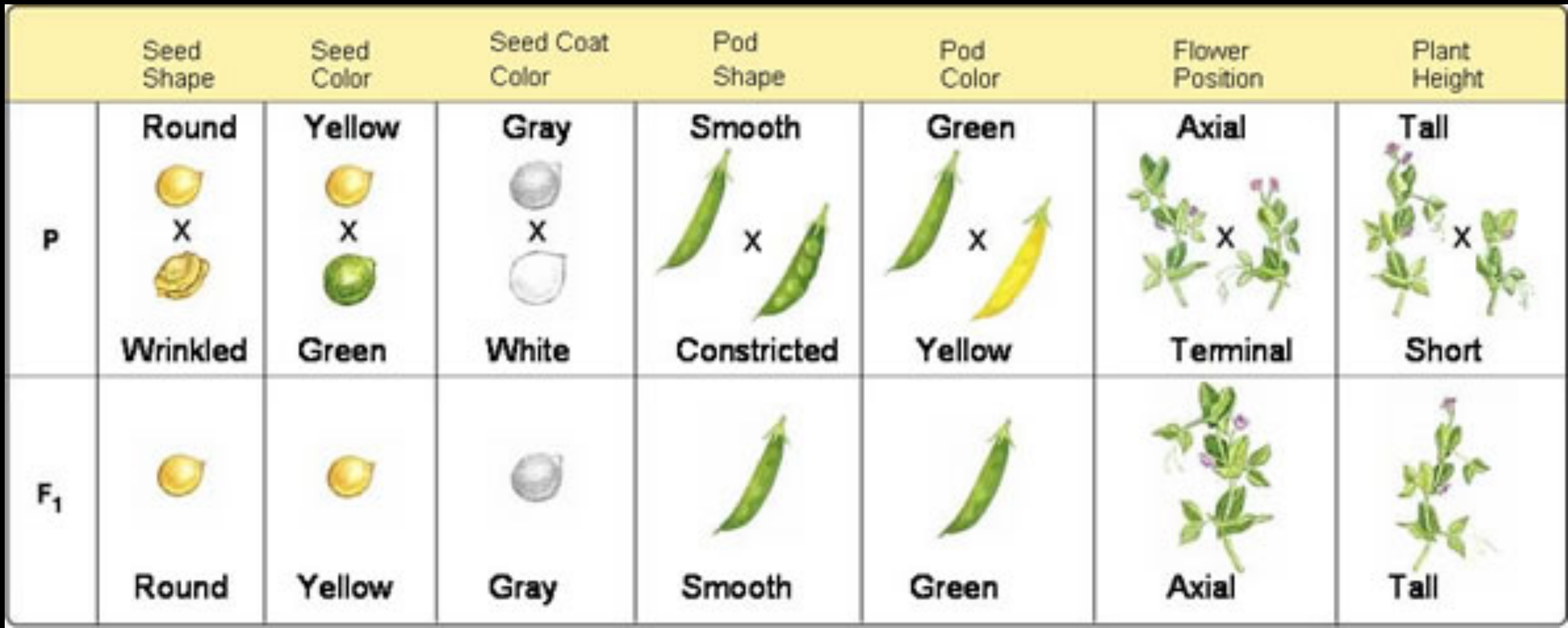
PBL: BINOMIAL AND MULTINOMIAL DATA ANALYSES

DANIEL ORTIZ-BARRIENTOS, 2015

BIOL2006

SPECIFIED CELL PROBABILITIES

EXERCISE 1



SPECIFIED CELL PROBABILITIES

EXERCISE 1

In a genetics experiment on tomatoes, a dihybrid cross was made, with the frequencies of the progeny expected to be in the ratio 9:3:3:1. The following table gives the observed frequencies:

- Round/Yellow, Wrinkled/Yellow, Round/Green, Wrinkled/Green
- N = 100 peas are examined, with the following counts in each category:

RY	WY	RG	WG
56	19	17	8

SPECIFIED CELL PROBABILITIES

EXERCISE 1

What are:

1) The expected proportions?
2) The expected frequencies?
3) The Chi-Square value?
4) The degrees of freedom?
5) The associated *p-value*
6) Can you reject the null hypothesis at alpha=0.05?

THIS IS THE LEAGUE TABLE OF THE MOST EFFECTIVE LEAGUE TABLES
GOVERNMENT DEPARTMENT OF DOING SOMETHING
LEAGUE TABLE
CHRIS MADDEN

EXERCISE 2

1) We can use the X^2 test to analyse cross-classified tabular data.

2) The most common use of X^2 is a test of independence of rows from columns ie a test of association between rows and columns

EXERCISE 2

ASSUMPTIONS

1) The data must satisfy the assumptions of a multinomial experiment
2) Expected cell values should NOT be < 5
   In this case the asymptotic approximation to the $\chi 2$ distribution does not hold.
3) Degrees of freedom for a contingency table = (r – 1)(c - 1)

Table 1

Example contingency table for case and control individuals genotyped at a diallelic marker locus (hypothetical data)

	Genotype			
Affection status	***AA***	***AB***	***BB***	**Row total**
Case	23	47	30	100
Control	12	40	48	100
Column total	35	98	67	200

Here we provide an example of a 2 × 3 (2 rows by 3 columns) contingency table. Each cell represents the number of observed genotypes (*AA*, *AB*, or *BB*) for a given affection status group (case or control). Here we assume that the marker locus has 2 alleles designated A and B.

EXERCISE 2

Expected values

We need to *estimate* the expected values (under the null hypothesis) in each cell. This is done by multiplying the appropriate row and column totals and dividing by the grand total.

$$\widehat{E(n_{ij})} = \frac{r_i c_j}{n}$$

EXERCISE 2

Carrie out the following:

1) Calculate the expected values for each cell?
2) Perform a Chi-Square test on those values
3) Find the the associated *p-value*
4) Can you reject the null hypothesis at alpha=0.05?

$$df = (r-1)(c-1)$$

PRACTICE QUESTION FOR HOME

Question 2. The results of a study suggest that the initial electrocardiogram (ECG) of a suspected heart attack victim can be used to predict in-hospital complications of an acute nature. The study included 469 patients with suspected myocardial infarction (heart attack). Each patient was categorised according to whether their initial ECG was positive or negative and whether the person suffered life-threatening complications subsequently in hospital. The results are summarised in the following table:

	Subsequent in-hospital life-threatening complications	
ECG	No	Yes
Negative	166	1
Positive	260	42

a. Is there sufficient evidence to indicate that whether or not a heart attack patient suffers complications depends on the outcome of the initial ECG? Test using $\alpha = 0.05$

GOODNES-OF-FIT

EXERCISE 3

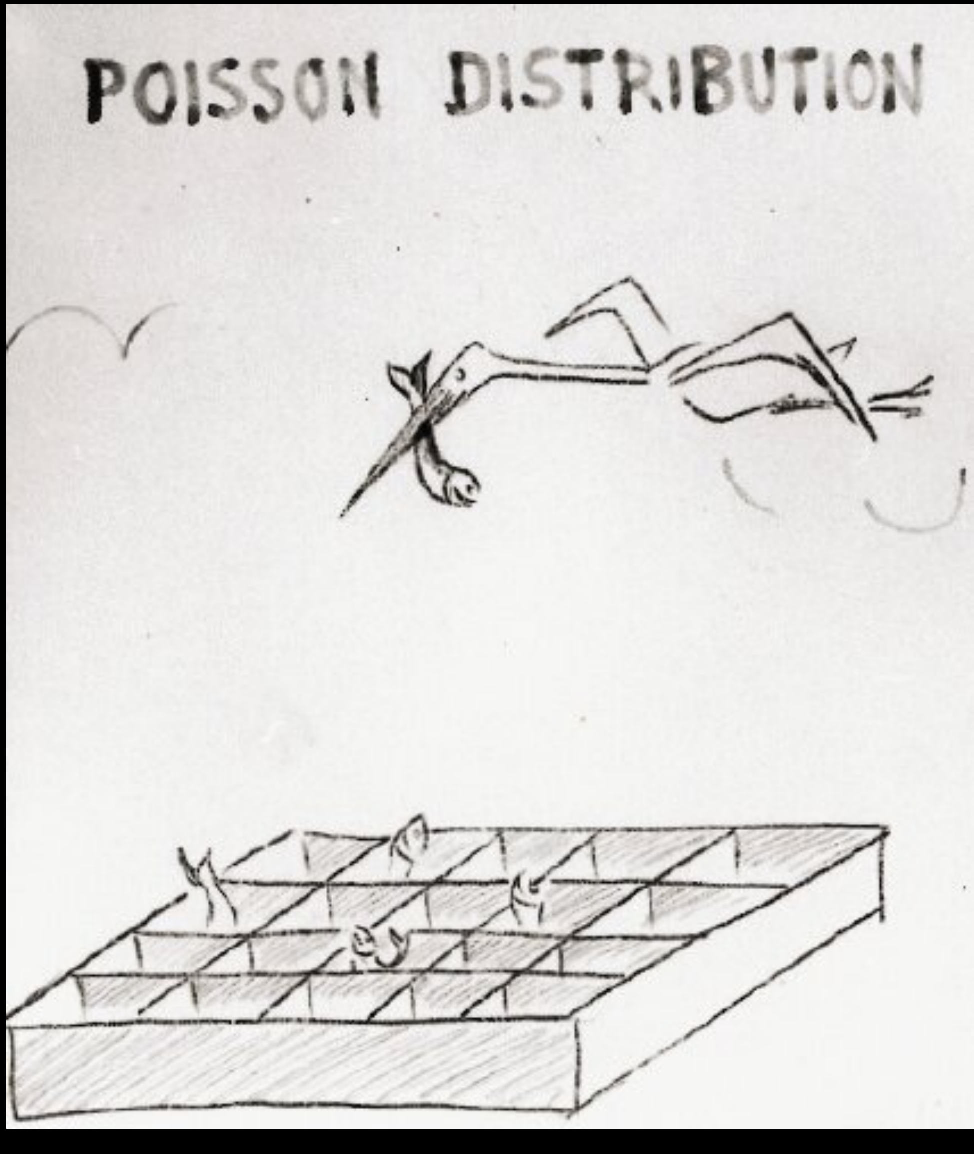
- Useful for testing whether data is consistent with a particular distribution

- Idea: Use X^2 to measure the "fit" of the probability model to the data. Significant lack of fit implies rejection of the null hypothesis of "no departure from the model"

EXERCISE 3

But, what model?

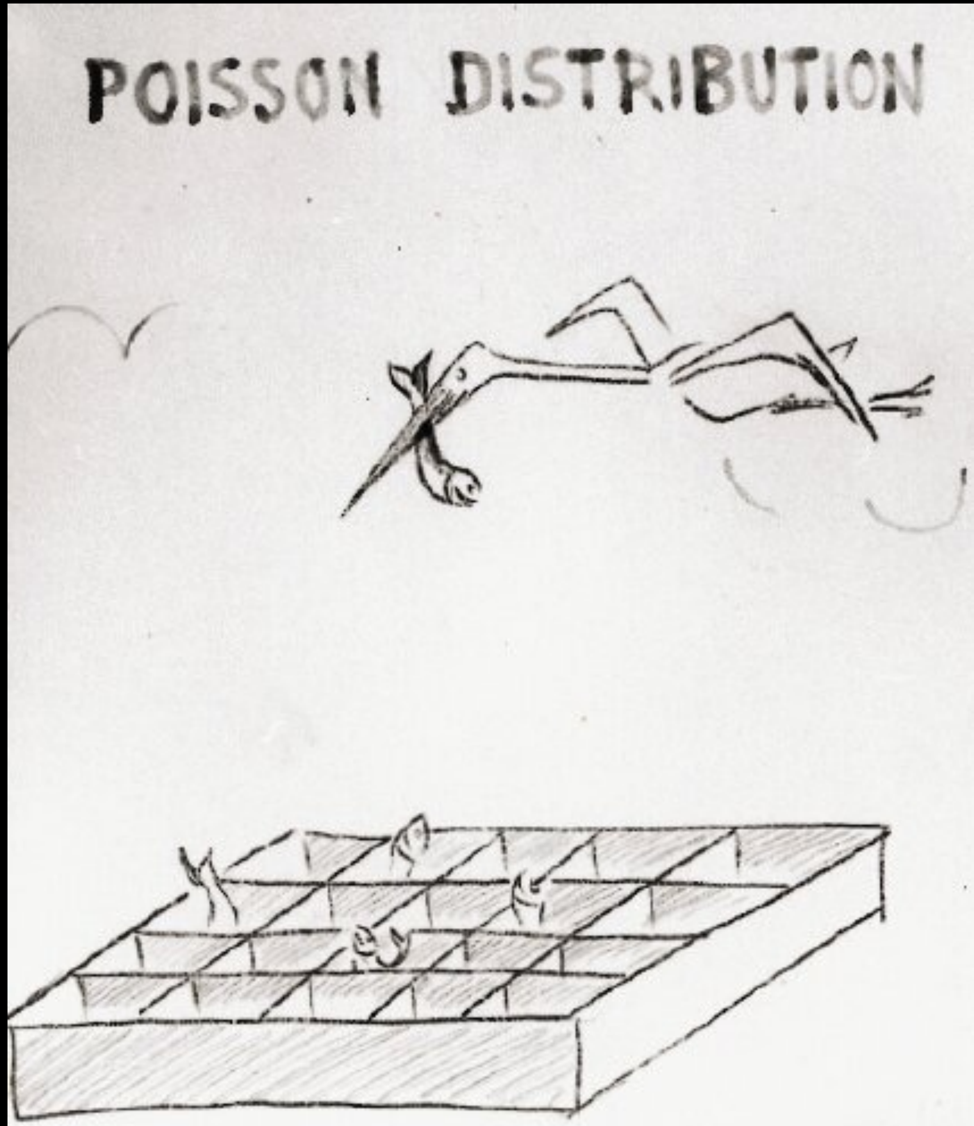
EXERCISE 3

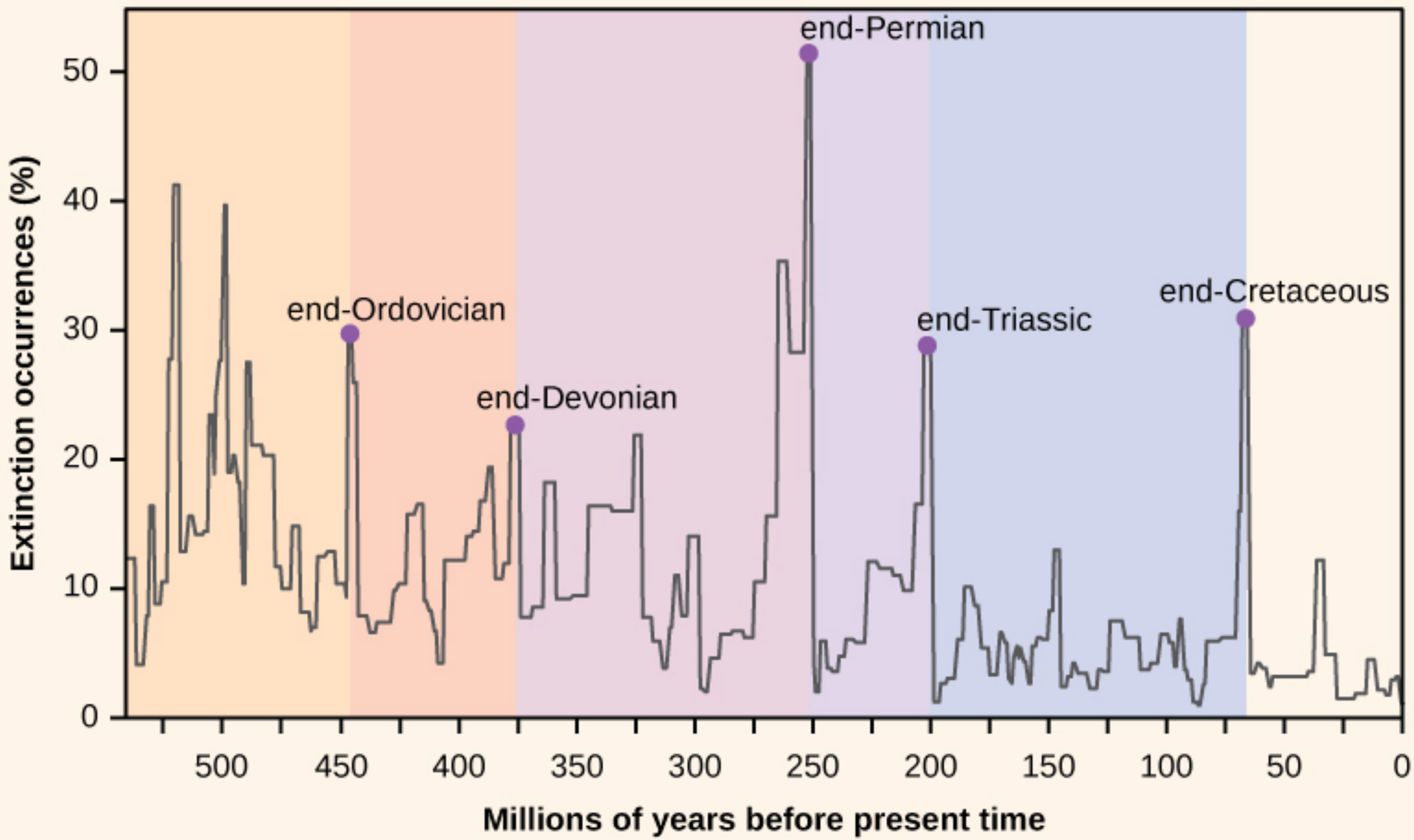


GOODNESS OF FIT

EXERCISE 3

In biology it helps us determine whether patterns in space or time are random





Dr. Susan Shirley, from UBC

Number Of Extinctions	Number Of Time Intervals
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

ARE EXTINCTION EVENTS, AS OBSERVED IN THE FOSSIL RECORD, RANDOM IN TIME, OR DO THEY HAVE CLUSTER (E.G., MASS EXTINCTIONS)?

Number Of Extinctions	Number Of Time Intervals
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

HO: THE NUMBER OF EXTINCTIONS PER TIME INTERVAL HAS A POISSON DISTRIBUTION

HA: THE NUMBER OF EXTINCTIONS PER TIME INTERVAL DOES NOT HAVE A POISSON DISTRIBUTION

Number Of Extinctions	Number Of Time Intervals
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

THE POISSON DISTRIBUTION

$$\Pr\{Y = k \mid \mu\} = \frac{e^{-\mu}\mu^k}{k!}$$

THE KEY PARAMETER IS **MU**. **MU** IS BOTH THE MEAN AND THE VARIANCE OF THE DISTRIBUTION

Number Of Extinctions	Number Of Time Intervals
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

THE POISSON DISTRIBUTION

$$\Pr\{Y = k \mid \mu\} = \frac{e^{-\mu}\mu^k}{k!}$$

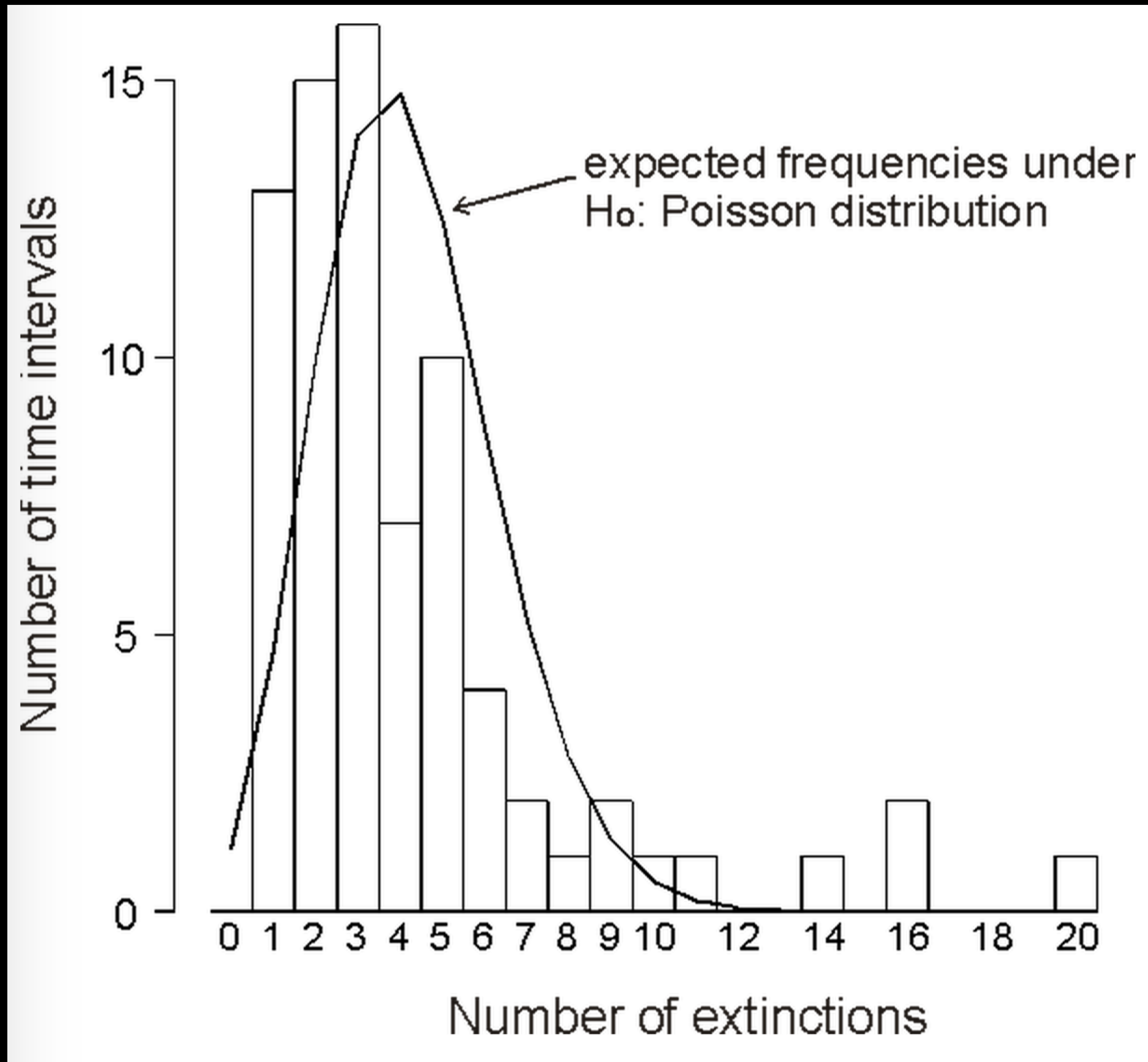
WE FIND **MU** BY AVERAGING THE CONTRIBUTIONS OF THE TWO COLUMNS ON THE LEFT.

$$MU = \frac{(0 \times 0) + (1 \times 13) + (2 \times 15)\ \ldots}{0 + 13 + 15 + 16\ \ldots}$$

MU = 4.21

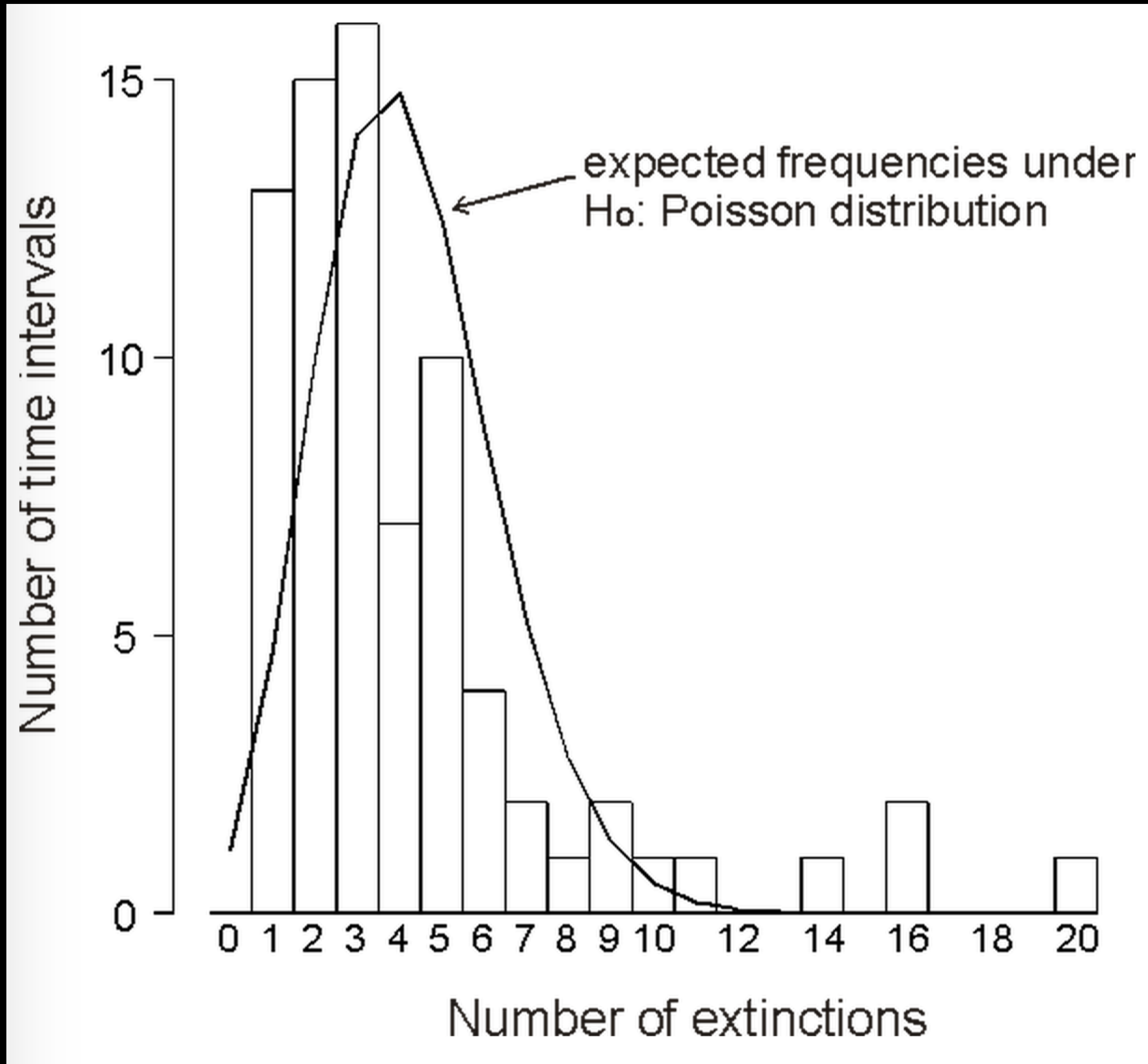
THE POISSON DISTRIBUTION WITH MU=4.21

Number Of Extinctions	Number Of Time Intervals
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



Number Of Extinctions	Number Of Time Intervals
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

DOES THE DATA FOLLOW THE POISSON DISTRIBUTION?



Number Of Extinctions	Number Of Time Intervals
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

DOES THE DATA FOLLOW THE POISSON DISTRIBUTION?

$$\Pr\{Y = k \mid \mu\} = \frac{e^{-\mu}\mu^k}{k!}$$

WE MUST FIND THE EXPECTED PROBABILITIES FOR EACH NUMBER OF EXTINCTIONS.

Number Of Extinctions	Number Of Time Intervals
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

MU=4.21

$$\Pr\{Y = k \mid \mu\} = \frac{e^{-\mu}\mu^k}{k!}$$

$$p_0 = \mu^0 e^{-\mu} = 0$$

$$p_1 = \mu^1 e^{-\mu}/1! =$$

$$p_2 = \mu^2 e^{-\mu}/2! = 0$$

$$p_3 = \mu^3 e^{-\mu}/3! = 0$$

$$p_4 = \mu^4 e^{-\mu}/4! = 0$$

.

.

.

$$p_{20} = \mu^{20} e^{-\mu}/20! = 0$$

Number of extinctions	Observed number of time intervals	Expected number of time intervals	How expected frequencies were computed
X	f_i	$\hat{f}_i$	
0	0	1.13	$(e^{-4.21}\ 4.21^0 / 0!)\ 76$
1	13	4.75	$(e^{-4.21}\ 4.21^1 / 1!)\ 76$
2	15	10.00	$(e^{-4.21}\ 4.21^2 / 2!)\ 76$
3	16	14.03	etc
4	7	14.77	etc
5	10	12.44	etc
6	4	8.72	etc
7	2	5.24	etc
8	1	2.76	etc
9	2	1.29	$(e^{-4.21}\ 4.21^9 / 9!)\ 76$
≥10	6	0.86	76-(1.13+4.75+10.00+…+1.29)
Total	**76**	**76**	

SOME RULES OF THUMB

1) NO EXPECTED FREQUENCIES LESS THAN 1.0
2) NO MORE THAN 20% LESS THAN 5

SOME RULES OF THUMB

1) NO EXPECTED FREQUENCIES LESS THAN 1.0
2) NO MORE THAN 20% LESS THAN 5

Number of extinctions	Observed number of time intervals	Expected number of time intervals
X	f_i	$\hat{f}_i$
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**

GOODNES-OF-FIT

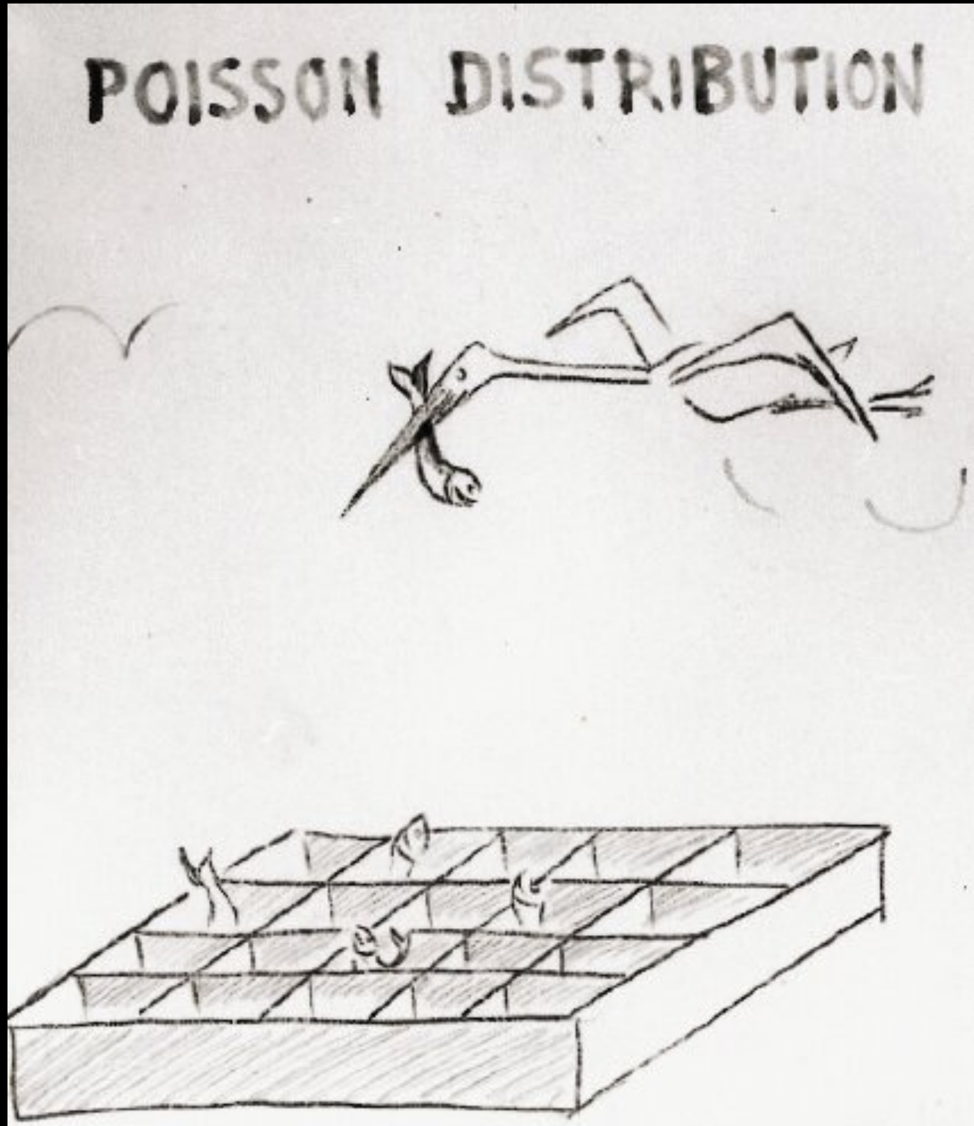
EXERCISE 3

Calculate the Chi-square value for the data before and decide whether extinctions seem to be random over time.

Note: **df** = k-1-1

df = 8 - 1 -1 = **6**,

One from the observations, and one from estimating μ.



GOODNES-OF-FIT

PRACTICE QUESTION FOR HOME

Question 1. The data in the following table are the frequency counts for 400 observations on the number of bacterial colonies within the field of a microscope, using samples of milk film. Is there sufficient evidence to claim that the data do not fit the Poisson distribution? (Use $\alpha = 0.05$). Pool cells where necessary.

Number of Colonies per Field	Frequency of Observation
0	56
1	104
2	80
3	62
4	42
5	27
6	9
7	9
8	5
9	3
10	2
11	0
19	1
	400

Data source: Bliss, C. A. and R. A. Fisher (1953). Fitting the negative binomial distribution to biological data. *Biometrics* **9:** 176-200

GOODNES-OF-FIT

EXERCISE 4 - OPTIONAL HOMEWORK

Question 3. According to the genetic model for the relationship between sex and colour blindness, the four categories, male and normal, female and normal, male and colour blind, female and colour blind, should follow a Multinomial distribution with probabilities given by $p/2, (p^2/2)+pq, q/2, q^2/2$, respectively, where $q=1-p$. A sample of 2000 people revealed 880, 1032, 80, and 8 in the respective categories. Do these data agree with the model? Use $\alpha = 0.05$. Use maximum likelihood to estimate p. Hint: The probability mass function for the Multinomial distribution with four categories is:

$$P(n_1, n_2, n_3, n_4 | p_1, p_2, p_3, p_4) = \frac{n!}{n_1! n_2! n_3! n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}$$

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38