Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

February 9, 2016

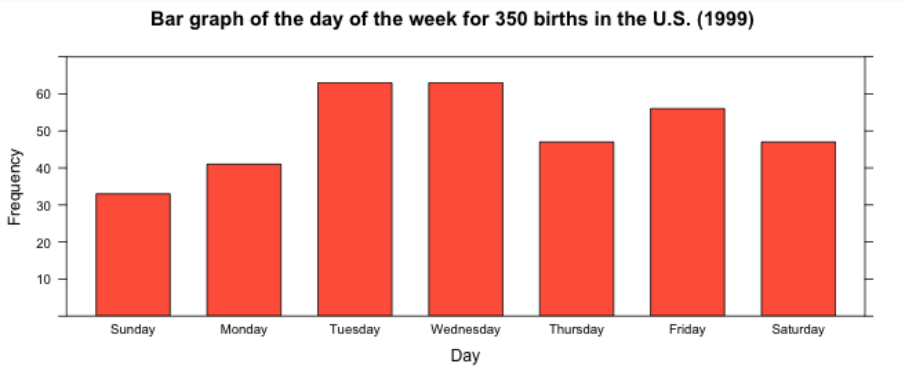
The Voinovich School of Leadership and Public Affairs

Table of Contents

Probability Models for Frequency Data

Probability Models

- Thus far we have used the binomial distribution, which works well for binary outcomes
- Now we move on to situations where we have frequency data on proportions of more than two outcomes



Day	No. of births
Sunday	33
Monday	41
Tuesday	63
Wednesday	63
Thursday	47
Friday	56
Saturday	47

The χ^2 goodness-of-fit test

H_0: Proportions are all the same; H_A: Proportions are *not* all the same

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

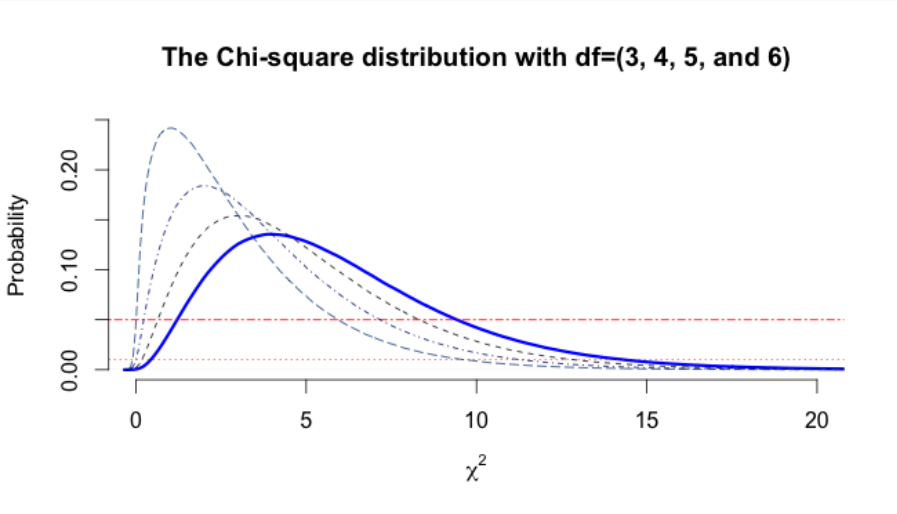
χ^2 distributed with $(\text{no. of categories} - 1)$ degrees of freedom (df)

Reject H_0 if $p - value \leq \alpha$; Do not reject H_0 otherwise

As $df \to \infty$ you need a larger χ^2 to Reject H_0 at the same α

Assumptions of the χ^2 test

1. No category should have expected frequency < 1
2. No more than 20% of categories should have expected frequencies < 5



An Example

We have four health campaigns that air. Null hypothesis is that each is recalled by identical proportion of viewers.

- $H_0 : P_a = 0.25; P_b = 0.25; P_c = 0.25; P_d = 0.25$
  H_A : Proportions are different
- $e_a = 0.25(300) = 75; e_b = 0.25(300) = 75;$
  $e_c = 0.25(300) = 75; e_d = 0.25(300) = 75$

Category	f_i	e_i	$(f_i - e_i)$	$(f_i - e_i)^2$	$(f_i - e_i)^2/e_i$
a	85	75	10	100	1.3333
b	95	75	20	400	5.3333
c	50	75	-25	625	8.3333
d	70	75	-5	25	0.3333
$\chi^2_{df=3}$	300	300			15.3333

- $p - value < 0.005$; Reject H_0; The Proportions are different

Another Example

M&M/MARS polled consumers as to their favorite $M\&M^{®}$ colors. Traditional distribution of colors and that found in a sample of 506 M&Ms is shown below. Do sampled proportions match tradition?

Category	f_i	e_i	$(f_i - e_i)$	$(f_i - e_i)^2$	$(f_i - e_i)^2/e_i$
Brown (30%)	177	151.8	25.2	635.04	4.1834
Yellow (20%)	135	101.2	33.8	1142.44	11.2889
Red (20%)	79	101.2	-22.2	492.84	4.8700
Orange (10%)	41	50.6	-9.6	92.16	1.8213
Green (10%)	36	50.6	-14.6	213.16	4.2126
Blue (10%)	38	50.6	-12.6	158.76	3.1375
$\chi^2_{df=5}$	506				29.5138

- $p - value < 0.005$; Reject H_0; Data do not support expected percentages so we have a problem with quality control

Days of the Week and No. of Births

H_0: Proportion of births are distributed equally across days of the week
H_A: Proportion of births are not distributed equally across days of the week

Set $\alpha = 0.05$

Day	No. of births	Expected	χ_i^2
Sunday	33	49.863	$\frac{(33-49.863)^2}{49.863} = 5.70$
Monday	41	49.863	$\frac{(41-49.863)^2}{49.863} = 1.58$
Tuesday	63	49.863	$\frac{(63-49.863)^2}{49.863} = 3.46$
Wednesday	63	49.863	$\frac{(63-49.863)^2}{49.863} = 3.46$
Thursday	47	49.863	$\frac{(47-49.863)^2}{49.863} = 0.16$
Friday	56	50.822	$\frac{(56-50.822)^2}{50.822} = 0.53$
Saturday	47	49.863	$\frac{(47-49.863)^2}{49.863} = 0.16$
Total	365	365	15.05

Calculated $\chi_6^2 = 15.05$ and its $p - value < 0.05$ so we Reject H_0; the data provide insufficient evidence to conclude that births are distributed equally across days of the week.

The Binomial Distribution Revisited

Gene content of the X chromosome revisited

Sex chromosomes are inherited in a very different pattern from that of the other chromosomes, which is known to affect their evolution in many ways. Are sex chromosomes unusual in other ways as well? For example, are there as many human genes on the X chromosome as we would expect from its size? The Human Genome Project has found 781 genes on the human X chromosome, out of a total of 20,290 genes found so far in the entire genome. The X chromosome represents 5.2% of the DNA content of the whole human genome. Under the proportional model, then, we would expect 5.2% of the genes to be on the X chromosome.

Is this what we observe?

H_0: Percentage of human genes on the X chromosome is $= 5.2\%$
H_A: Percentage of human genes on the X chromosome is $\neq 5.2\%$

Chromosome	Observed	Expected
X	781	1,055
Not X	19,509	19,235
Total	20,290	20,290

We could use the Binomial but why do that; much easier to use χ^2 ...
$\chi^2_1 = \frac{(781-1055)^2}{1055} + \frac{(19509-19235)^2}{19235} = 75.1$
The associated $p-value < 0.05$ so we can easily Reject H_0; the data provide insufficient evidence to conclude that the percentage of human genes on the X chromosome is 5.2%

The Binomial Test revisited

Does the number of boys in families with 2 children *follow the binomial distribution*?

H_0: No. of boys in families with 2 children follows the binomial distribution

H_A: No. of boys in families with 2 children does not follow the binomial distribution

Data come from the NLYS, with number of families $= 2,444$. Of the 4888 children in the sample only 1332 +1164 are boys; $\hat{p} = \frac{2496}{4888} = 0.5106$

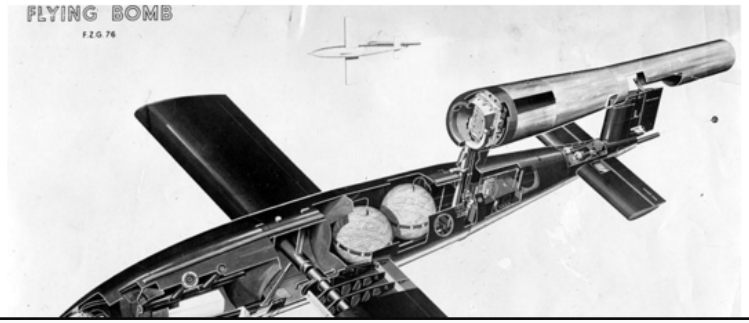
Boys	Families	Children	$P[\text{X successes}\|n=2]$	Expected Families	χ^2
0	530	1060	$P[\text{0 boys}] = 0.2395124$	$2444 \times 0.2395124 = 585.3682$	5.237111
1	1332	2664	$P[\text{1 boy}] = 0.4997753$	$2444 \times 0.4997753 = 1221.4508$	10.005421
2	582	1164	$P[\text{2 boys}] = 0.2607124$	$2444 \times 0.2607124 = 637.1810$	4.778773
Total	2444	4888	1	2444	20.02131

Note df$= 3-1-1 = 1$ (WHY?); and $p-value < 0.05$ so we Reject H_0. The no. of boys in families with two children does not follow the binomial distribution.

The Poisson Distribution

AATISH BHATIA SCIENCE 12.21.12 4:48 PM

WHAT DOES RANDOMNESS LOOK LIKE?



The Poisson Distribution

The Poison distribution is a discrete probability distribution for the counts of events that occur in a given space or time interval. For e.g.,

- The number of cases of a disease in different towns
- The number of particles emitted by a radioactive source per second
- The number of births per hour during a given day
- The number of highway fatalities per mile driven
- The number of shark attacks in a year

$$P(X) = \frac{e^{-\mu}\mu^{X}}{X!}; \text{where } X = 0, 1, 2, 3, \ldots, n; \text{and Mean } = \text{Variance } = \mu$$

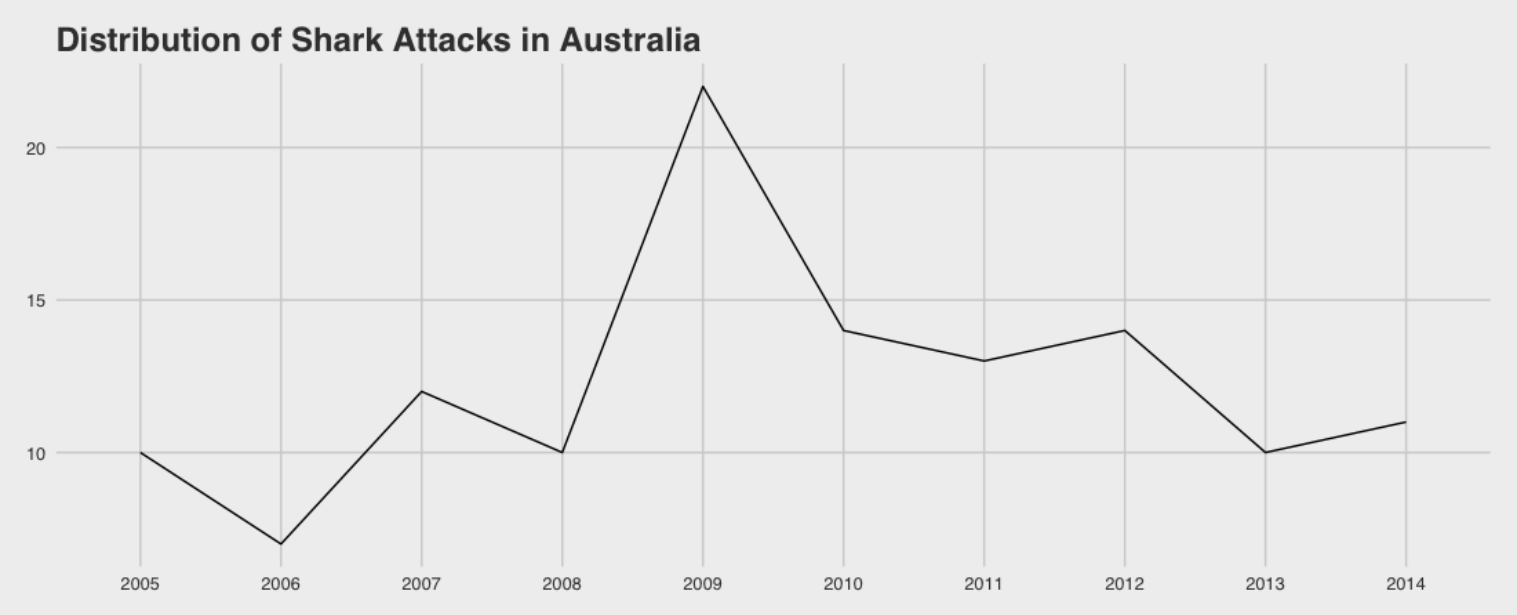
where $X =$ the number of events in a given time interval or space; $\mu =$ the mean number of events per time interval or space; and $P(X) =$ the probability of observing exactly X events in a given interval.

Example

Hospital births occur on average at 1.8 births per hour. What is $P(X = 4)$?

$$P(X = 4) = \frac{e^{-1.8}(1.8)^{4}}{4!} = 0.0723$$

Shark Attacks



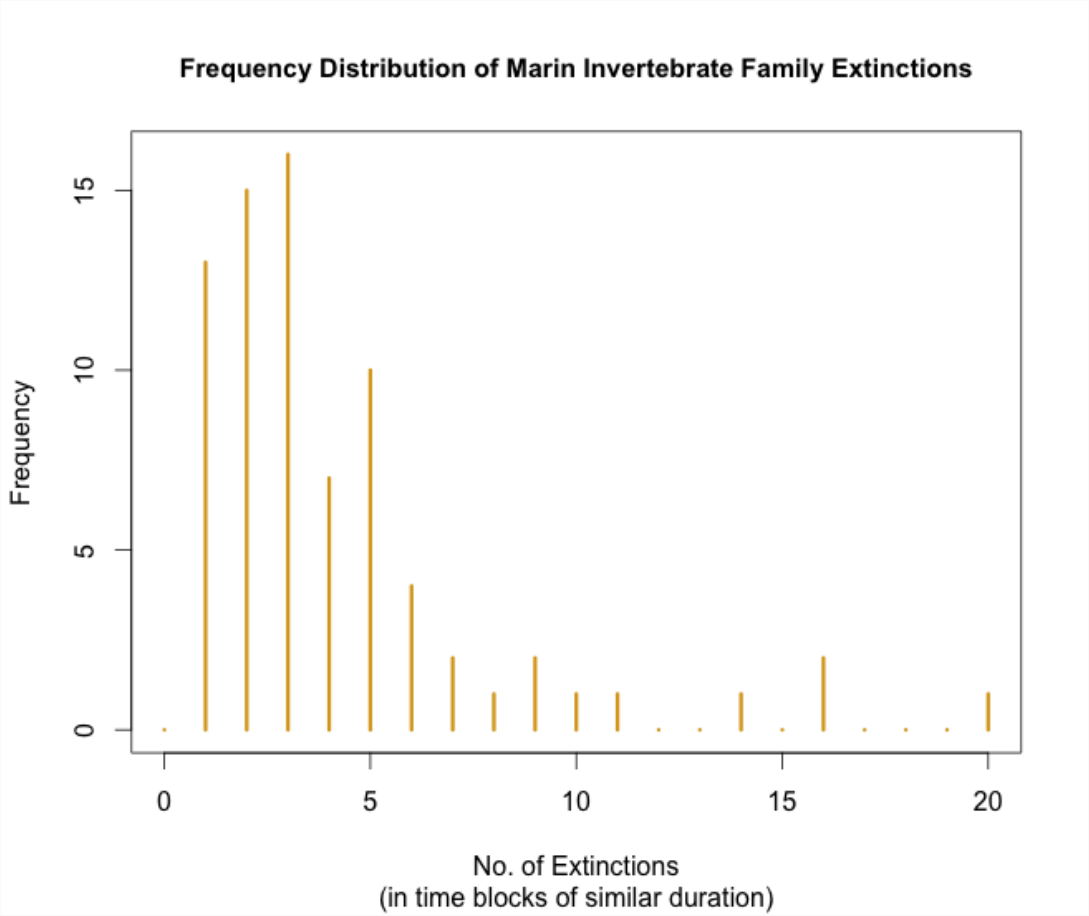
Are shark attacks random or caused by climate change, etc? Does their distribution mimic a Poisson process?

If $\mu = 2$, what is $P(X = 22)$? Practically 0.

What about $P(X = 0)$? About 0.1353353.

Testing Randomness with the Poisson

No. of 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



If extinctions are randomly distributed then a Poisson distribution should capture that flow of events rather well.

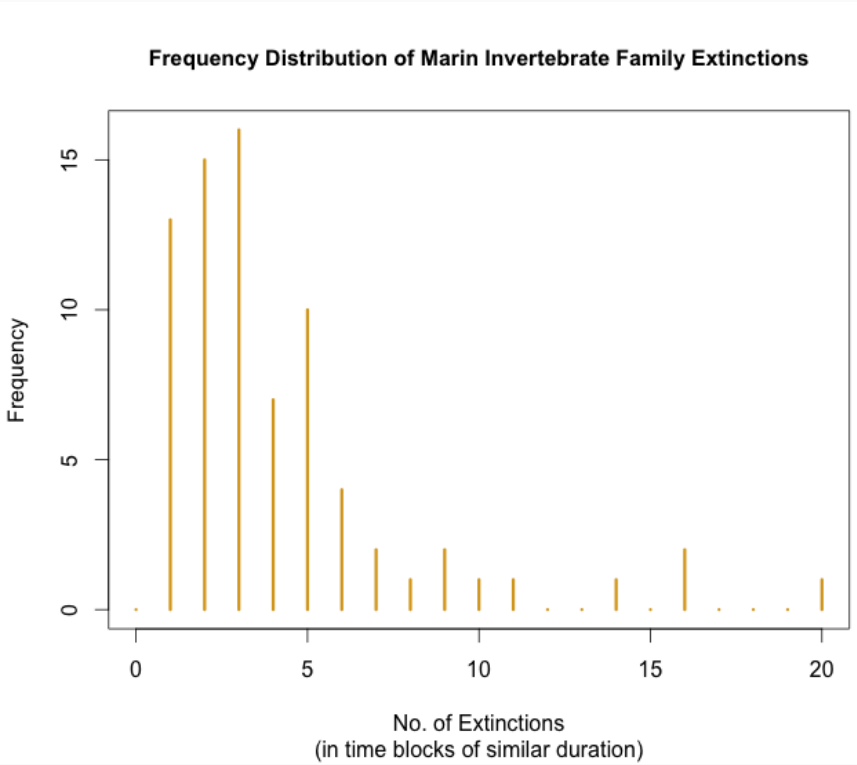
H_0: No. of extinctions per time interval follow a Poisson distribution
H_A: No. of extinctions per time interval do not follow a Poisson distribution

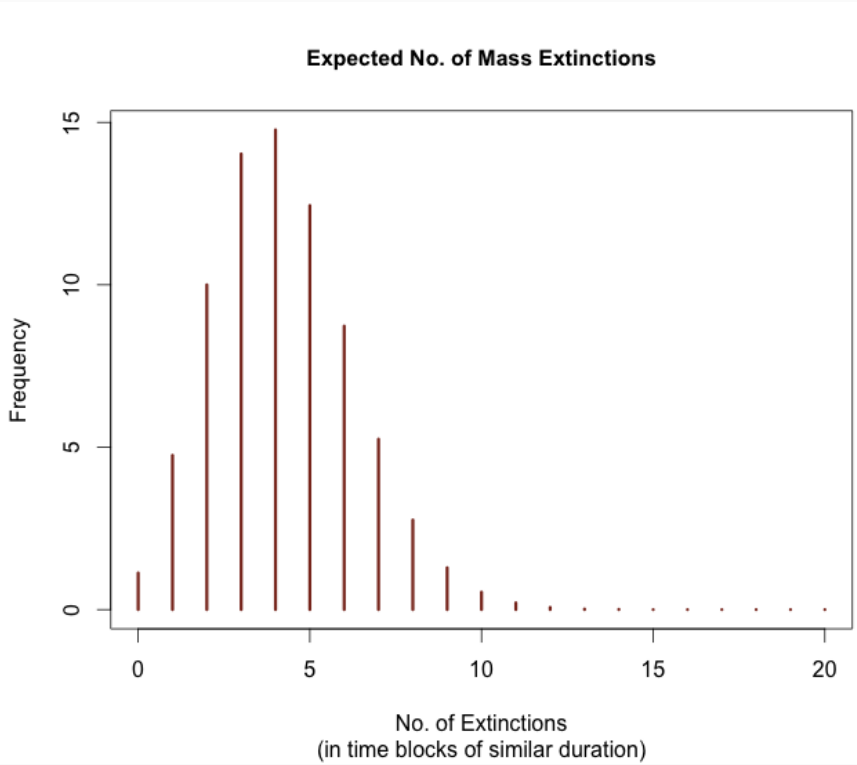
Since we do not know μ we will have to use $\bar{X} = 4.210526$ as our estimate of μ. Now, if extinctions are $\sim Poisson\,(\mu = 4.210526)$ then what would be the expected counts of $0, 1, 2, 3, \ldots, 20$ extinctions?

We can calculate these expected frequencies via R; they are shown below:

```
[1]  1.13  4.75 10.00 14.03 14.77 12.44 8.73 5.25 2.76
[10] 1.29  0.54  0.21  0.07  0.02  0.01  0.00  0.00  0.00
[19] 0.00  0.00  0.00
```

Observed vs. Expected No. of Extinctions





Because several categories have expected frequencies < 1 and 15 of the 21 categories have expected frequencies < 5 we can recode the categories to be: 0 & 1, 2, 3, 4, 5, 6, 7, 8 or more.

Extinctions (X)	Observed	Expected	χ^2
0 or 1	13	5.88	8.6215
2	15	10.00	2.5000
3	16	14.03	0.2766
4	7	14.77	4.0875
5	10	12.44	0.4786
6	4	8.72	2.5549
7	2	5.24	2.0034
8 or more	9	4.91	3.4069
Total	76	76	23.93

The p-value for $\chi^2_6 = 23.93$ with $\alpha = 0.05 = 0.0005381$
Since this p-value is < 0.05 we Reject H_0; the data provide insufficient evidence to conclude that mass extinctions follow the Poisson distribution. Recall that for the Poisson distribution the Mean $=$ Variance. In this particular case we have $Mean = 4.21$ and $Variance = 13.72$. This tells us extinctions occurred more often in particular time intervals than others.

Clumping versus Dispersion in Poisson

Clumping

- Variance is $>$ Mean
- Events occur closer together (in space and/or time) than would be expected by chance (for e.g., contagious diseases)
- One "success" increases the chance of another successes occurring soon/nearby

Dispersion

- Mean is $>$ Variance
- Events occur farther apart (in space and/or time) than would be expected by chance (for e.g., territorial animals)
- One "success" decreases the chance of another success occurring soon/nearby

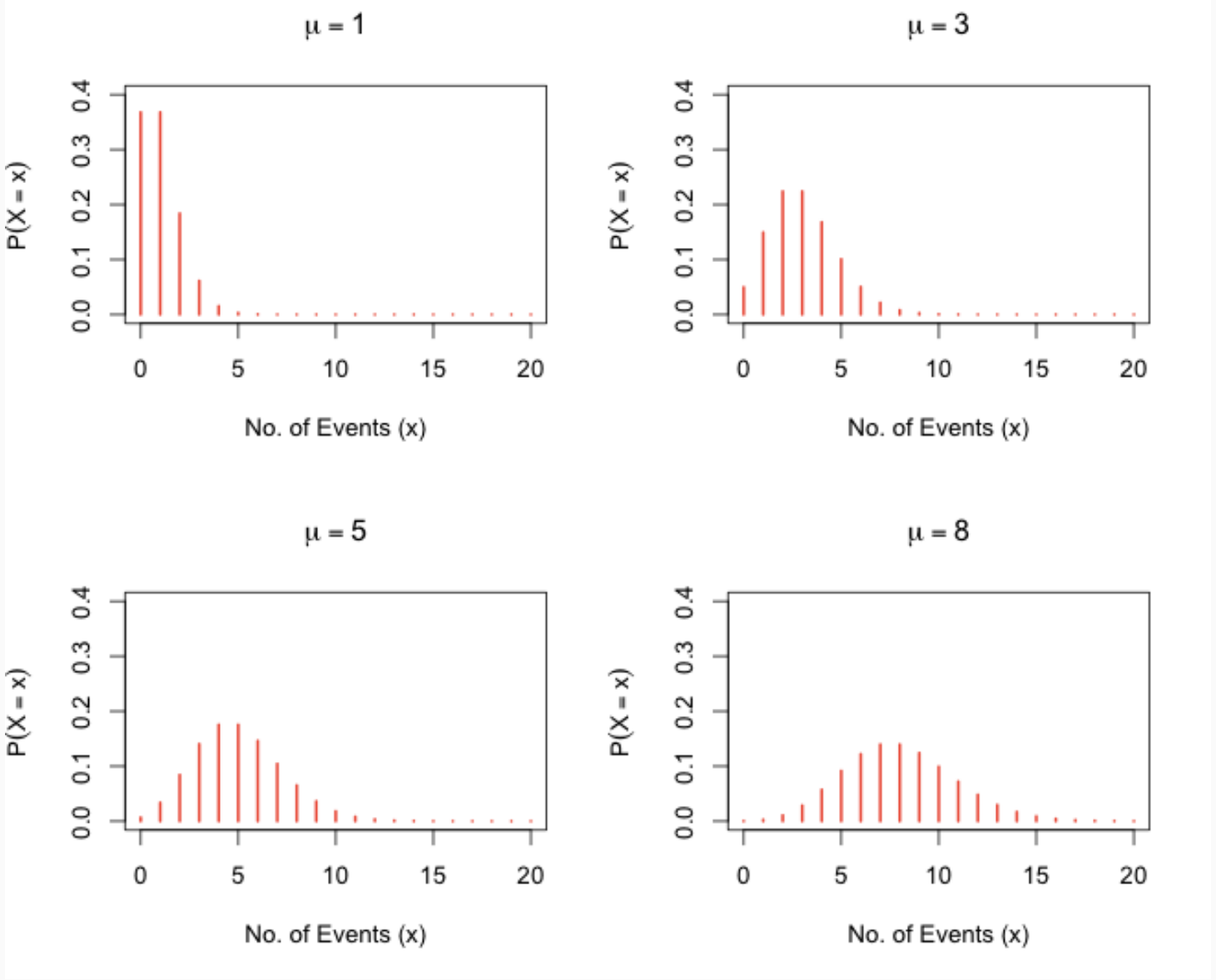
Alternatives: (a) Negative-Binomial; (b) Zero-Inflated Poisson; (c) Zero-Inflated Negative-Binomial; (d) Hurdle Models

Assumptions of the Poisson Distribution

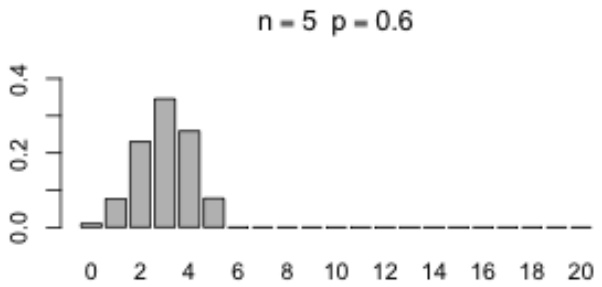
1. The probability of observing a single event over a small time interval (or space) is approximately proportional to the size of that time interval (or space) .
2. The probability of two events occurring in the same narrow time interval (or space) is negligible.
3. The probability of an event within a certain time interval (or space) does not change across different time intervals (or space).
4. The probability of an event in one time interval (or space) is independent of the probability of an event in any other non-overlapping time interval (or space) .

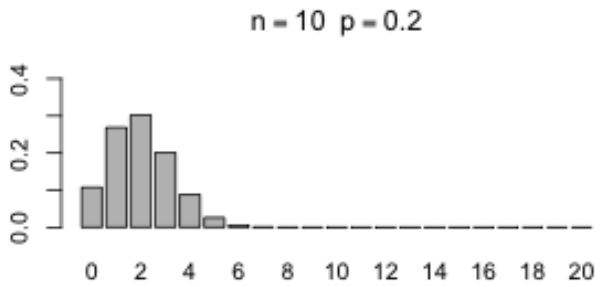
When (a) $n \to \infty$, and (b) $p \to 0$, the Poisson distribution approximates the Binomial distribution. Much easier to calculate the probability of a specific number of "rare" successes via Poisson than if we used the Binomial approach. As the mean $\to \infty$ the Poisson resembles the Normal distribution.

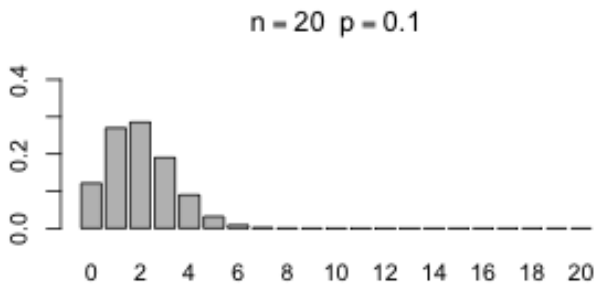
Some Poisson Distributions

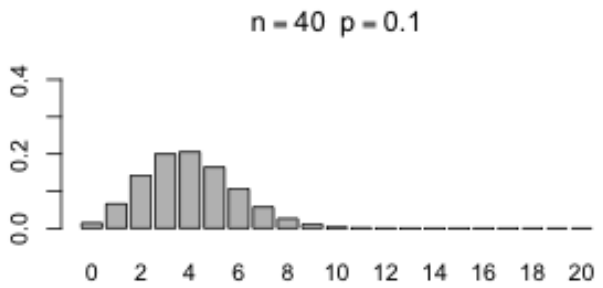


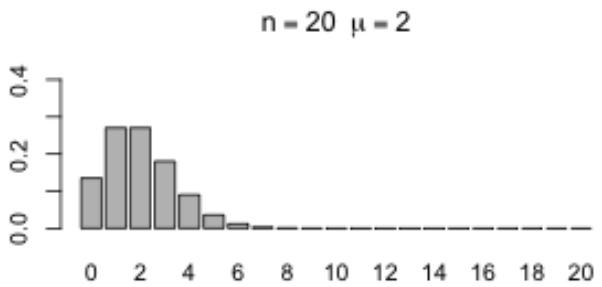
Binomial $\to$ Poisson as $n \to \infty$ and $p \to 0$

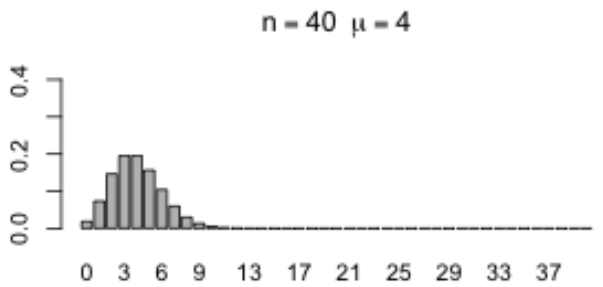












Poisson $\to$ Normal as $\mu \to \infty$