Lecture 7 – Proportion & Frequency 比例&频数

- Outline for today
  - Recall Hypothesis testing 假设检验
  - The binomial test 二项检验
  - Chi-square test 卡方检验
  - Poisson distribution 泊松分布
  - Summary
  - R Lab & Discussion

第二次课后作业

- 发布日期： 2024年11月1日
- 截止日期：2024年11月8日11:59pm

1. 假设检验

- 假设检验的步骤
  - 1. 提出假设 (state the H_0 & H_A)
  - 2. 利用数据计算检验统计量 (compute the test statistic with data)
  - 3. 确定 P 值 (determine the P-value)
  - 4. 得出适当的结论 (draw the appropriate conclusions)
  - 5. 汇报结果 (report the results)

1. 假设检验

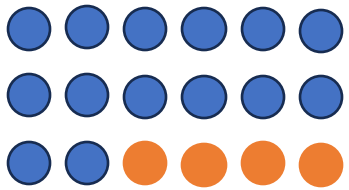
- 假设检验的例子
  - 蟾蜍是否有惯用手的偏好?
  - 总体中右撇子和左撇子出现的频数相同，还是像人类那样一种类型比另一种类型更常见?

$H_0: p_{left} = p_{right} = 0.5$

$H_A: p_{left} \neq p_{right}$

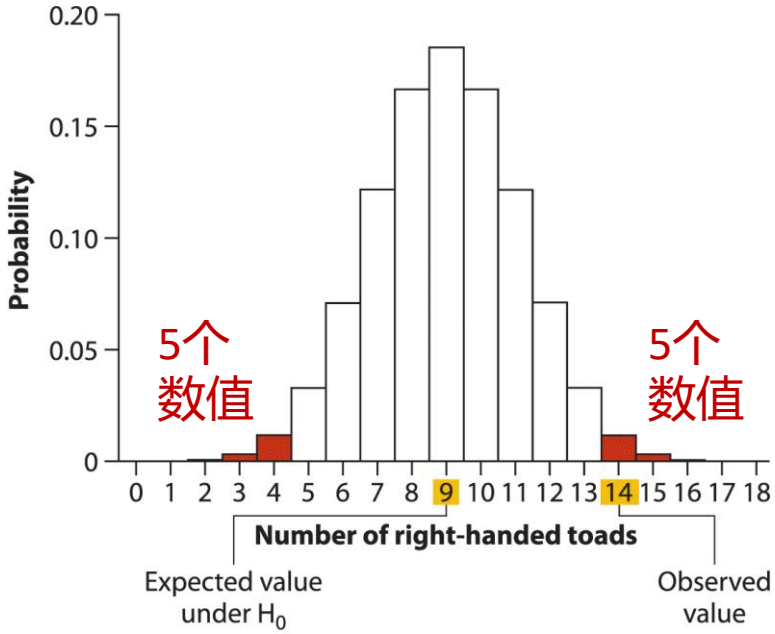
- **描述假设检验的过程?**

Hintau Aliaksei/ Shutterstock.com



观察结果

- 14 right-handed
- 4 left-handed



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

1. 假设检验

- I 类错误和 II 类错误 Type I and Type II errors
  - I 类错误：拒绝为真的零假设 (rejecting a true null hypothesis) ——假阳性
  - II 类错误：没有拒绝为假的零假设 (failing to reject a false null)——假阴性

- 小测试 Q4: 把显著性水平α的值从 0.05 降到 0.01， 对以下各项有怎样的影响？ (Ch6-Q6)
    - a. 犯第一类错误的概率？ - 降低
    - b. 犯第二类错误的概率？ - 增加
    - c. 检验的功效？ - 降低
    - d. 样本大小？ - 无关

1. 假设检验

- 选择合适的检验方法（回答四个问题）
  - 检验涉及一个变量，还是检验两个或更多变量之间的关系?
  - 变量是分类的还是数值的?
  - 数据是否以配对形式出现?
  - 检验的假设是什么，数据是否符合这些假设?

- 参考：INTERLEAF 7 （Page 958）

表 1. 一个变量的检验方法 (一组数据)

Data type 数据类型	Goal	目标	Test	检验方法
Categorical 类型变量 （分类变量）	Use frequency data to test whether a population proportion equals a null hypothesized value	使用频数数据检测总体中的比例是否等于零假设中的值	Binomial test (7) χ^2 Goodness-of-fit test with two categories (use if sample size is too large for the binomial test) (8)	二项检验 (7) χ^2 拟合度检验，应用于变量分为两类（如果样本量过大，无法进行二项检验，则使用该检验） (8)
	Use frequency data to test the fit of a specific population model	使用频数数据检测特定总体模型的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
Numerical 数值变量	Test whether the mean equals a null hypothesized value when data are approximately normal (possibly only after a transformation) (13)	当数据近似正态分布（或经过转换后符合）时，检测平均值是否等于零假设中的值 (13)	One-sample t-test (11)	单样本t检验 (11)
	Test whether the median equals a null hypothesized value when data are not normal (even after transformation)	当数据不符合正态分布（即使经过转换），检测中位数是否等于零假设中的值	Sign test (13)	符号检验 (13)
	Use frequency data to test the fit of a discrete probability distribution	使用频率数据测试离散概率分布的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
	Use data to test the fit of the normal distribution	检测数据是否符合正态分布	Shapiro-Wilk test (13)	Shapiro-Wilk 检验 (13)

表 2. 两个变量相关性的检验方法

		Type of explanatory variable 解释变量			
		Categorical	类型变量	Numerical	数值变量
Type of response variable 响应变量	Categorical 类型变量	Contingency analysis (9)	独立性检验 (9)	Logistic regression (17)	逻辑斯蒂回归 (17)
	Numerical 数值变量	t-tests, ANOVA, Mann-Whitney U-test, etc. [See Table 3 for more details.]	t检验 方差分析 U检验等 [更多细节见表 3]	Linear and nonlinear regression (17) Linear correlation (16) Spearman's rank correlation (when data are not bivariate normal) (16)	线性和非线性回归 (17) 线性相关 (16) Spearman 秩相关（当数据不是二元正态分布时） (16)

表 3. 两个变量相关性的检验方法及其前提假设（或检验多组数据的差异）

Number of treatments 处理组数量	Tests assuming normal distribution	假设数据符合正态分布的检验	Tests not assuming normal distributions	假设数据不符合正态分布的检验
Two treatments (independent samples) 两独立样本	Welch's t-test (12) Two-sample t-test (use when variance is equal in the two groups) (12)	Welch's t检验 (12) 双样本t检验（两组方差相等时使用） (12)	Mann-Whitney U-test (Wilcoxon rank-sum test) (13)	U检验（秩和检验）(13)
Two treatments (paired data) 两配对样本	Paired t-test (12)	配对t检验 (12)	Sign test (13)	符号检验 (13)
More than two treatments 超过两组设置	ANOVA (15)	方差分析 (15)	Kruskal-Wallis test (15)	Kruskal-Wallis 检验 (15)

R commands summary

Contingency table

First categorical variable — Second categorical variable

```
sex_survive_table <- table(titanicData$sex, titanicData$survive)
```

Mosaic plot

```
mosaicplot(sex_survive_table)
```

Odds ratio

Estimate the odds ratio

```
fisher.test(sex_survive_table)$estimate
```

```
fisher.test(sex_survive_table)$conf.int
```

Give 95% confidence interval of odds ratio

χ^2 contingency analysis

Calculate the expected values

```
chisq.test(sex_survive_table)$expected
```

```
chisq.test(sex_survive_table, correct = FALSE)
```

Fisher's exact test

```
fisher.test(sex_survive_table)
```

Confidence interval for a proportion

Use the Agresti-Coull method

```
binom.confint(x = 30, n = 87, method = "ac")
```

Number of "successes" — Sample size

Binomial test

Proportion specified by null hypothesis

```
binom.test(x = 14, n = 18, p = 0.5)
```

Number of "successes" — Sample size

Frequency table

Name of frequency table — Categorical variable

```
MMtable <- table(MMlist$color)
```

χ^2 Goodness of fit test

Vector of expected proportions

```
chisq.test(MMtable, p = expected_proportions)
```

Poisson distribution

Number of successes — Mean number of successes

```
dpois(x = 3, lambda = 4.21)
```

P-value from χ^2

Observed χ^2 — Degrees of freedom

```
pchisq(q = 23.939, df = 6, lower.tail = FALSE)
```

2. 针对一维类型变量：比例/频数

- 学习目标：比例/频数及相关概率模型检验
  - 如何最佳地利用随机样本估计总体中某一类事件的比例/频数
  - 如何计算其置信区间以及P值
  - 针对比例/频数如何最佳地检验假设
    - 二项分布、卡方分布、泊松分布

2. 针对一维类型变量：比例/频数

- 在总体中共享某一特征的个体比例也是从该总口中随机抽取的个体具有该属性的概率。
  - 比例的范围：[0, 1]
- 一些概念
  - 比例 proportion
  - 概率 probability （比例是对概率的一种估计）
  - 频数 frequency（count，计数）
  - 频率 rate (或relative frequency 相对频数 = 频数/总量)

2.1 对比例进行检验：二项分布 (the binomial distribution)

- 一般化 (generalization)
  - 基于对个体进行的测量，将它们分为两个互斥的组：成功 vs 失败
    - 例如成功或失败、活着或死亡、左撇子或右撇子、糖尿病患者或非糖尿病患者
    - 将两个个类别分布称为“成功”和“失败”是为了方便，而非价值判断。
  - 如果我们从这个人口中随机抽取n个个体，那么落入“成功”类别的个体数量的抽样分布可以用二项分布来描述。
    - “二项” 的含义：只有两个（"bi-"）可能的结果，两者都有名称（"-nomial"）
  - 二项分布提供了在一定数量的独立试验中获得“成功”次数的概率分布，其中每次试验中“成功”的概率相同。

2.1 二项分布 (the binomial distribution)

- 二项分布公式

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X} \qquad \text{其中，} \binom{n}{X} = \frac{n!}{X!(n-X)!}$$

  - 二项式公式给出了在n次试验中获得 X 次成功的概率，其中任何单次试验的结果要么是成功要么是失败。二项分布假设：
    - 试验次数（n）是固定的
    - 各次试验是独立的
    - 每次试验中成功的概率（p）都是相同的
  - 其中X是介于0和n之间的整数。
  - 在右侧，有二项系数 $\binom{n}{X}$ 为“n中选择X”。这表示在n次试验中产生X次成功的所有可能序列的数量。

2.1 二项分布 (the binomial distribution)

- 概率分布：不需要考虑success事件的顺序
  - 二项系数

$$\binom{n}{X} \qquad \Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$\binom{5}{3}$$

SSSFF SSFSF SSFFS SFSSF SFSFS
SFFSS FSSSF FSSFS FSFSS FFSSS

2.1 二项分布 (the binomial distribution)

- 随机样本中的成功次数
  - 泥植草（*Heteranthera multiflora*）采用一种简单的机制来避免自交
    - 一些个体的花器官中柱头向左偏斜，而另一些向右偏斜
    - p_{left} = 0.25 (success),　p_{right} = 0.75 (failure)
  - Q：从这个总体随机取样 n = 27，那么这个样本中恰好有X = 6个植株的柱头向左偏斜 (success) 的概率是多少?

© photo courtesy of Spencer C. H. Barrett

2.1 二项分布 (the binomial distribution)

- 随机样本中的成功次数
  - Q：从这个总体随机取样 n = 27，那么这个样本中恰好有X = 6个植株的柱头向左偏斜 (success) 的概率是多少?
    - p_{left}=0.25 (success)， p_{right}=0.75 (failure)
  - 从这一总体中进行随机抽样的过程完全符合二项分布的假设：
    - n个独立的随机试验，其中每次试验中成功的概率p在每次试验中都是相等的。

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$\Pr[6\ left-handed\ flowers] = \binom{27}{6} 0.25^6 (1-0.25)^{27-6}$$

2.2 比例的抽样分布 sampling distribution of the proportion

- 随机样本中的成功次数：完整的二项分布
  - 从这个总体随机取样 n = 27，那么这个样本中恰好有有 X = c(0:27) 个植株的柱头向左偏斜 (success) 的概率是多少?

- **p_{left} = 0.25 (success)**
- p_{right} = 0.75 (failure)

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$

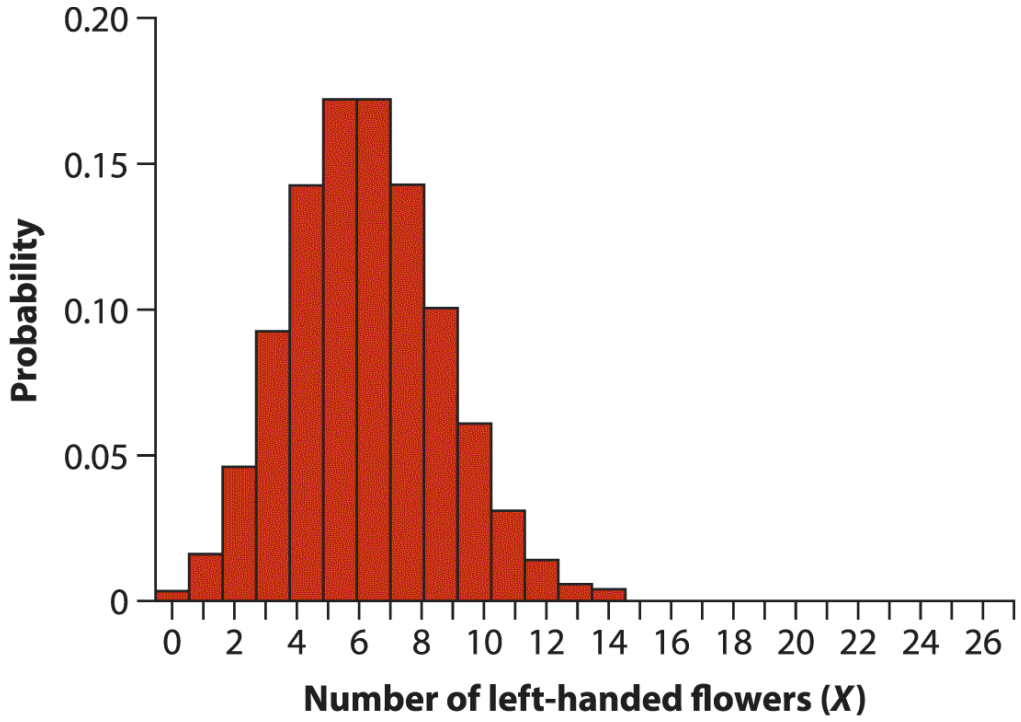
Number of left-handed flowers (X)	Pr[X]	Number of left-handed flowers (X)	Pr[X]
0	4.2×10^{-4}	14	0.0018
1	0.0038	15	5.1×10^{-4}
2	0.0165	16	1.3×10^{-4}
3	0.0459	17	2.8×10^{-5}
4	0.0917	18	5.1×10^{-6}
5	0.1406	19	8.1×10^{-7}
6	0.1719	20	1.1×10^{-7}
7	0.1719	21	1.2×10^{-8}
8	0.1432	22	1.1×10^{-9}
9	0.1008	23	7.9×10^{-11}
10	0.0605	24	4.4×10^{-12}
11	0.0312	25	1.8×10^{-13}
12	0.0138	26	4.5×10^{-15}
13	0.0053	27	5.5×10^{-17}

2.2 比例的抽样分布 sampling distribution of the proportion

- 随机样本中的成功次数：完整的二项分布
  - 从这个总体随机取样 n = 27，那么这个样本中恰好有有 X = c(0:27) 个植株的柱头向左偏斜 (success) 的概率是多少?

- **p_{left} = 0.25 (success)**
- p_{right} = 0.75 (failure)

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$



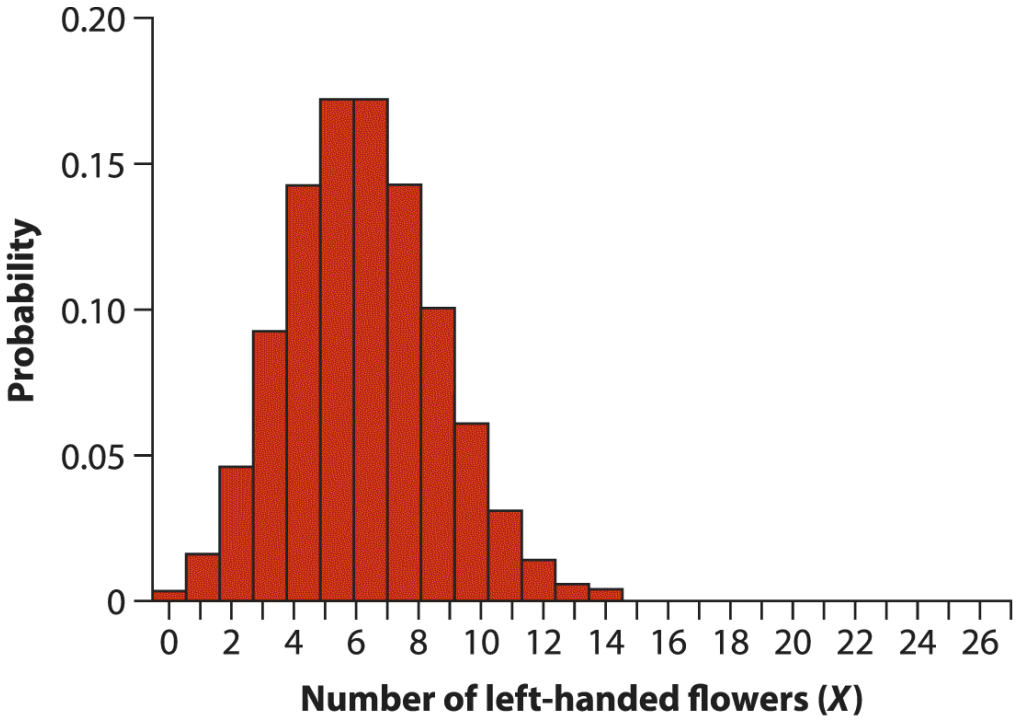
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.2 比例的抽样分布 sampling distribution of the proportion

- 随机样本中的成功次数：完整的二项分布
  - 从总体中随机取样，样本中恰好有X个个体为success 的概率是多少?
    - p_{left}=0.25 (success); p_{right}=0.75 (failure)
  - 比例的估算 (p-hat)

$$\hat{p} = \frac{X}{n}$$

    - p : 总体中某一类的比例
    - $\hat{p}$: 样本中某一类的比例
  - 大数定律 (the law of large numbers)
    - 估计值随样本量增大而更精确

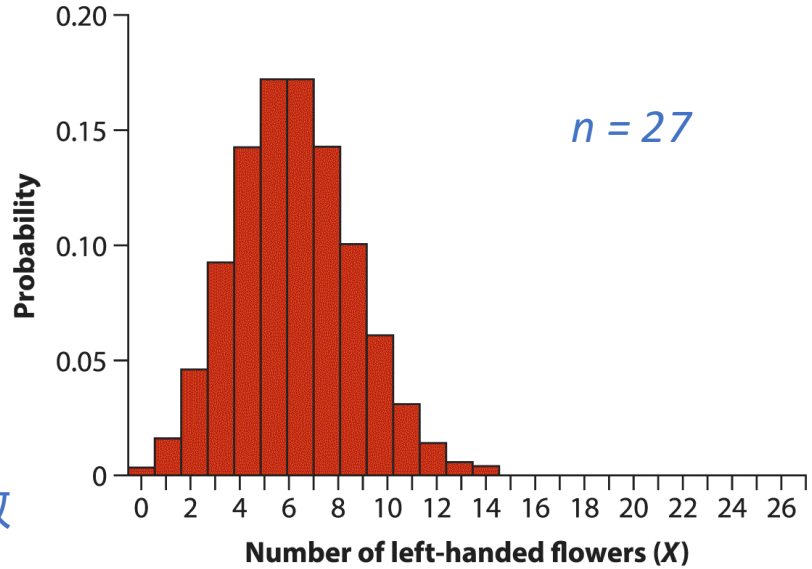


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

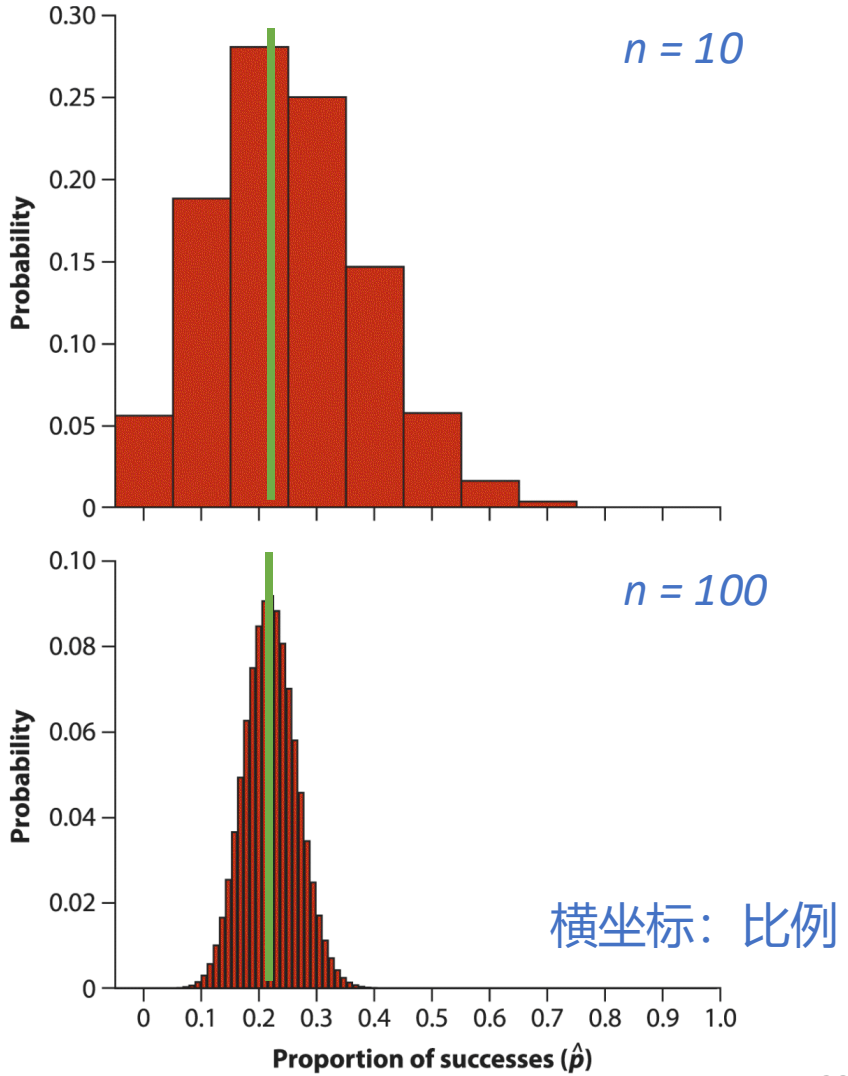
2.2 比例的抽样分布 sampling distribution of the proportion

- 比例的抽样分布 ~ 样本大小
  - 从总体中随机取样，样本中恰好有x个个体为success 的概率是多少?
    - p_{left}=0.25 (success)

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$



横坐标：频数

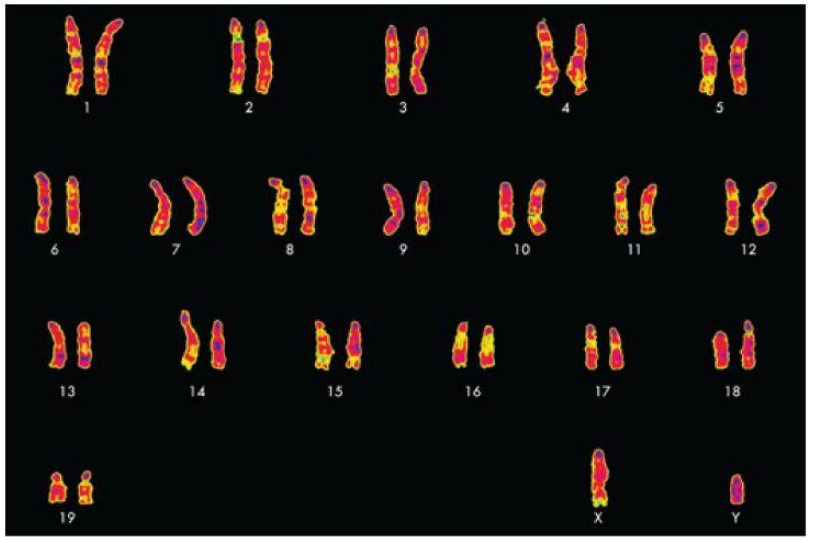


横坐标：比例

2.3 检验比例：二项检验 the binomial test

- 将二项抽样分布应用于比例的假设检验
  - 当总体中的一个变量具有两种可能的状态（即“成功”和“失败”），并且检验总体中成功的相对频率 (p) 是否与零假设 (p_0) 相匹配。

  - 例子：小鼠基因 ~ X染色体
    - 该研究目的是检验进化理论的一个预测，即与生殖细胞发育相关的基因在X染色体上出现的频率显著更高（相比常染色体）。
    - 观测数据：有25个相关基因，其中有10个位于X染色体上（40%）。

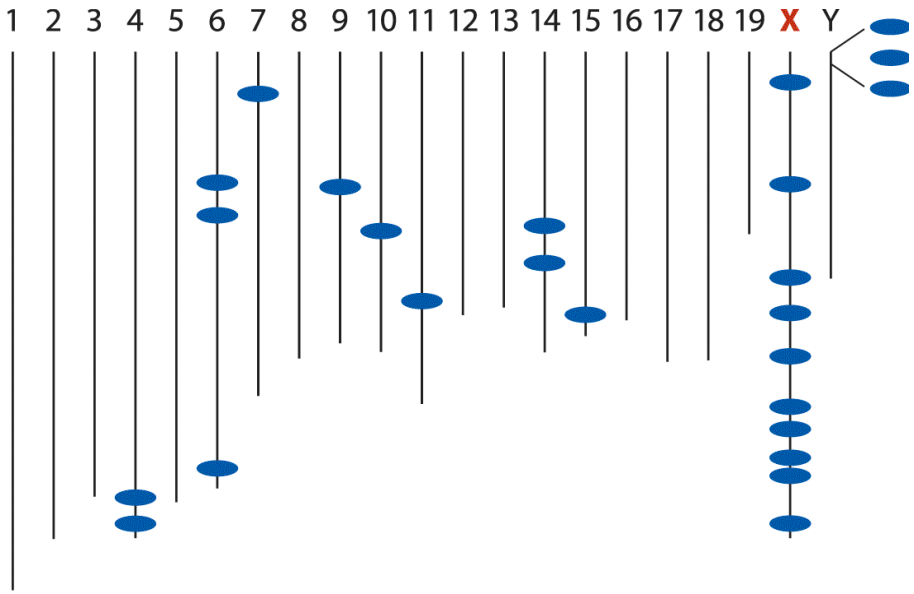


Look at Sciences/Science Source

2.3 检验比例：二项检验 the binomial test

- 将二项抽样分布应用于比例的假设检验
  - 当总体中的一个变量具有两种可能的状态（即“成功”和“失败”），并且检验总体中成功的相对频率 (p) 是否与零假设 (p_0) 相匹配。

- 例子：小鼠基因 ~ X染色体
  - 已知：X染色体包含基因组中6.1%的基因（基于染色体的长度）；
  - 观测数据：25个相关基因中，有10个位于X染色体上（40%）；
  - 如果这些特定基金在整个基因组中是“随机”分布的，那么我们预期只有6.1% 的基因会出现在X染色体上；

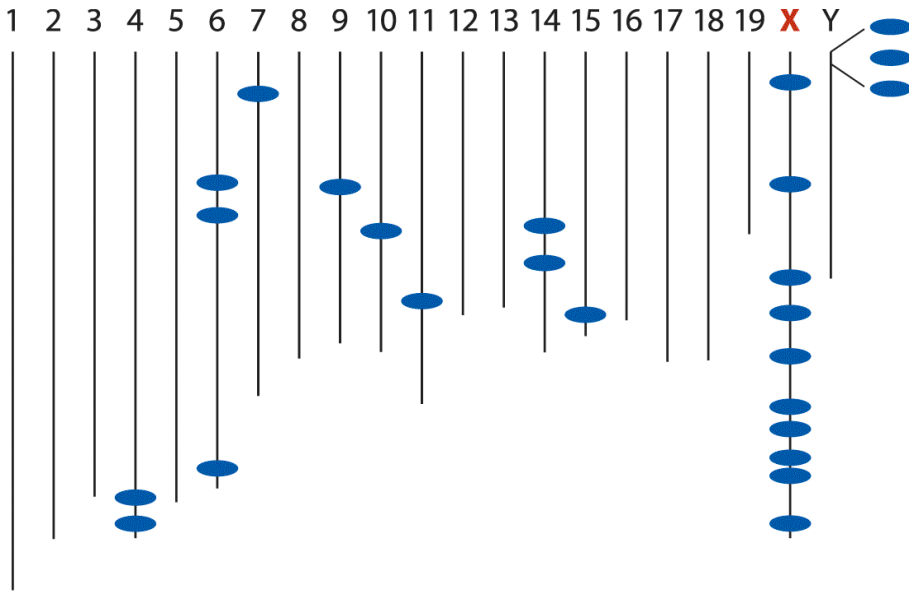


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 检验比例：二项检验 the binomial test

- $p = p_0$?
- 例子：小鼠基因 ~ X染色体
  - 提出假设
    - H_0?
    - H_A?
    - 单侧检验还是双侧检验?
  - 如果随机分布，那么该类基因在X染色体上的比例应该和X染色体上的基因占总基因组的比例相同 (proportional)。

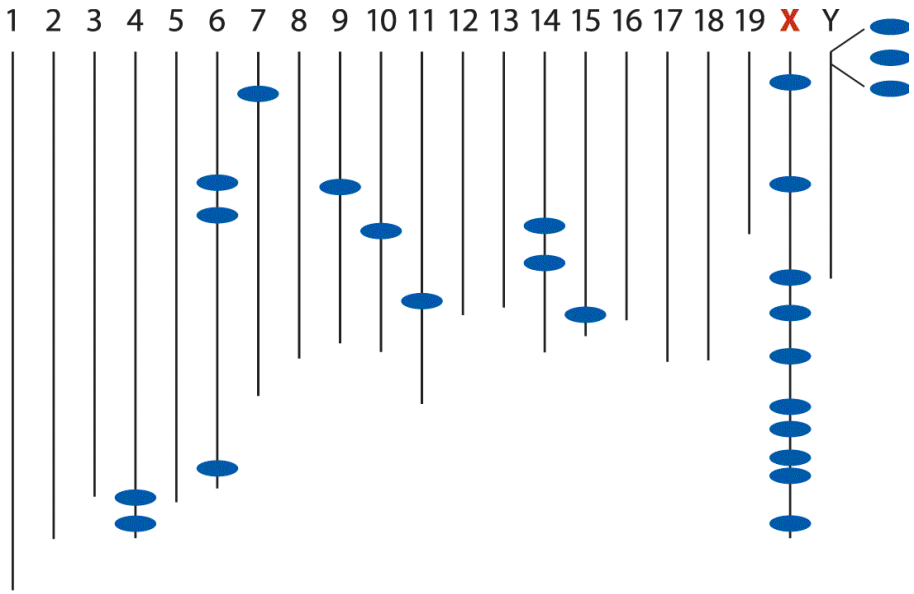
- 10 out of 25 genes on X chromosome (40%)
- If randomly, 6.1% on X chromosome (by chr length)



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 检验比例：二项检验 the binomial test

- 例子：小鼠基因 ~ X染色体
  - 提出假设
    - H_0：该类基因位于X染色体上的概率是 p=0.061
    - H_A：该类基因位于X染色体上的概率不是 p=0.061。
    - 双侧检验
  - 检验统计量
    - 观测值（？）
    - H_0下的预期值（$0.061 \times 25 = 1.525$）



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 检验比例：二项检验 the binomial test

- 例子：小鼠基因 ~ X染色体
  - P值: 出现与观测结果一致或更极端的概率
    - 单侧：$[X \geq 10]$
    - 双侧：$[X \geq 10] \times 2$

$$P = 2 \times \Pr[\text{number of successes} \geq 10]$$
$$= 2 \times \Pr[10]+\Pr[11]+\Pr[12]+\cdots+\Pr[25]$$
$$= 2 \times 9.9\times 10^{-7}$$
$$= 1.98\times 10^{-6}$$

Number of genes on X	$Pr - H_0$	Number of genes on X	$Pr - H_0$
10	9.1×10^{-7}	14	0.0018
11	8.0×10^{-8}	15	5.1×10^{-4}
12	6.1×10^{-9}	16	1.3×10^{-4}
13	4.0×10^{-10}	17	2.8×10^{-5}
14	2.2×10^{-11}	18	5.1×10^{-6}
15	1.0×10^{-12}	19	8.1×10^{-7}
16	4.3×10^{-14}	20	1.1×10^{-7}
17	1.5×10^{-15}	27	5.5×10^{-17}

$$\Pr[X\ success] = \binom{25}{X} 0.061^X (1-0.061)^{25-X}$$

2.3 检验比例：二项检验 the binomial test

- 例子：小鼠基因 ~ X染色体
  - 统计量（比例）：$\hat{p} = \frac{10}{25} = 0.40$
  - P值 = 1.98×10^{-6}
  - 显著性水平 0.05下的结论?

  - 拒绝零假设，并得出结论：
    - X染色体上存在某类与生育相关基因不成比例的数量。我们对位于小鼠X染色体上的该类基因的比例的最佳估计值为0.4，远大于零假设中的0.061。

  - “X染色体上某类与生育相关基因的比例明显偏大（0.40，SE=0.10；binomial test，n=25，P=1.98×10^{-6}）。”

2.3 检验比例：二项检验 the binomial test

- 比例估计值的标准误

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- 比例估计值的置信区间—the Agresti–Coull method

$$p' = (X+2)/(n+4)$$

$$p' - 1.96\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + 1.96\sqrt{\frac{p'(1-p')}{n+4}}$$

- 比例估计值的置信区间—the Wald method（**不推荐**）
  - 只在n较大且总体参数*p*不接近0或1时才准确

$$\hat{p} - 1.96\,\mathrm{SE}_{\hat{p}} < p < \hat{p} + 1.96\,\mathrm{SE}_{\hat{p}}$$

2.3 检验比例：二项检验 the binomial test

- 二项检验提供了精确的P值，并原则上可以应用于适用于有两个类别的任何数据。但是，在没有计算机的情况下，特别是当n较大时，计算二项检验的P值可能会很繁琐。

- 在适当的情况下，有一些快速计算出近似的P值的替代方法，可以节省大量时间：
  - 卡方拟合优度检验 χ^2 goodness-of-fit test (Chapter 8)
  - 二项检验的正态近似 the normal approximation (Chapter 10)

2.4 *Break – R coding*

- Binomial probability
  - > dbinom(6, size = 27, prob = 0.25)
  - > choose(27,6) * 0.25^6 * 0.75^(27-6) # choose(n,x)

- Binomial test
  - > binom.test(x = 6, n = 27, p = 0.25)

- 95% confidence interval
  - > binom.test(x = 6, n = 27, p = 0.25)$conf.int
  - > binom.confint(x = 6, n = 27, method = "ac") # library(binom)

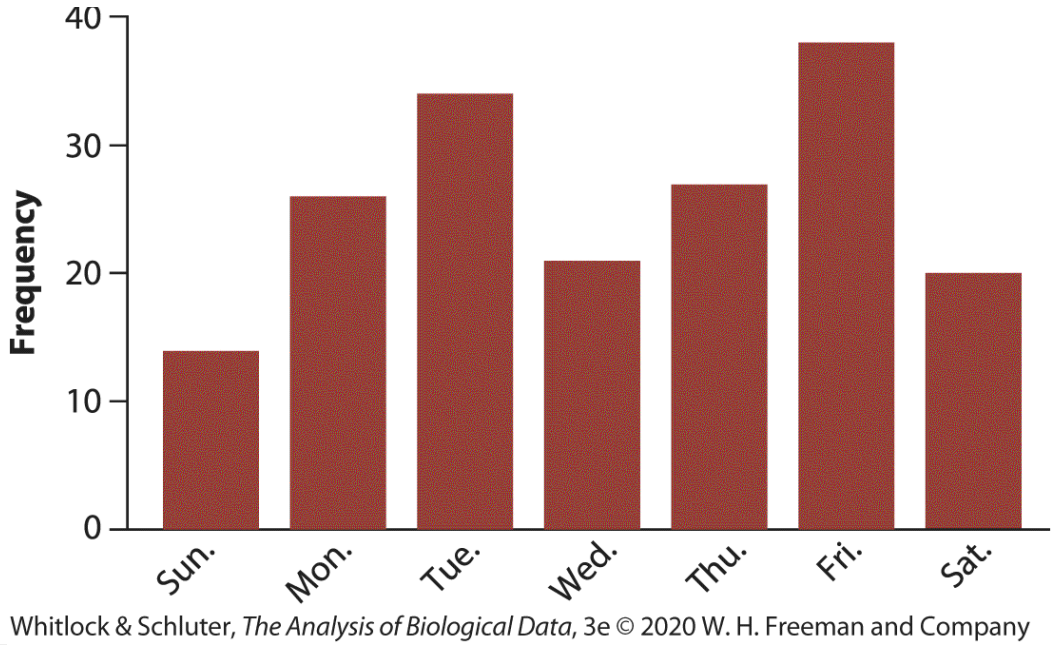
3. 卡方检验 the χ^2 goodness-of-fit test

- 二项检验的替代方法
  - 快速计算出近似的P值，可以节省大量时间
  - 卡方拟合优度检验 χ^2 goodness-of-fit test (Chapter 8)
  - 二项检验的正态近似 the normal approximation (Chapter 10)

- 卡方拟合优度检验（卡方检验） the χ^2 goodness-of-fit test
  - χ^2: Chi-square (Kye-square)

3.1 卡方检验 the χ^2 goodness-of-fit test — 检验比例/频数

- 例子：出生日期
  - 2016年新生儿日期的随机样本：n = 180
  - 2016年是闰年（53个周五和周六，其余日子是52个）



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Day	Number of births
Sunday	14
Monday	26
Tuesday	34
Wednesday	21
Thursday	27
Friday	38
Saturday	20
Total	180

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 例子：出生日期
  - 2016年新生儿日期的随机样本：n = 180
  - 2016年是闰年（53个周五和周六，其余日子是52个）

- 如何对出生日期的频数提出假设?
  - H_0? （期望的，随机的分布）
  - H_A?

Day	Number of births
Sunday	14
Monday	26
Tuesday	34
Wednesday	21
Thursday	27
Friday	38
Saturday	20
Total	180

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 如何对出生日期的频数提出假设?

  - H_0：每周的每一天的出生频数与该天在一年中出现的次数成比例。
    - The frequency of births on each day of the week is proportional to the number of times that day occurs in the year.

  - H_A：每周的每一天的出生频数与该天在一年中出现的次数不成比例。
    - The frequency of births each day of the week is not proportional to the number of times that day occurs in the year.

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 如何对出生日期的频数提出假设?
  - H_0：每周的每一天的出生频数与该天在一年中出现的次数成比例;
  - H_A：每周的每一天的出生频数与该天在一年中出现的次数不成比例;
- 数据： (Expected=180×52/366=25.574)

天/day	天数	期望 (expected) 频数
Sunday	52	25.574
Monday	52	25.574
Tuesday	52	25.574
Wednesday	52	25.574
Thursday	52	25.574
Friday	53	26.066
Saturday	53	26.066
Total	366	180

天/day	观察 (observed)频数
Sunday	14
Monday	26
Tuesday	34
Wednesday	21
Thursday	27
Friday	38
Saturday	20
Total	180

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 检验统计量是什么？
  - The χ^2 test statistic

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

  - χ^2 统计量测量了来自数据的观测频数与来自零假设的期望频数之间的不一致程度。
    - 如果实际数据与H_0的期望一致，那么 $\chi^2 = 0$。
    - 任何与零期望的偏离都会导致 $\chi^2 > 0$。
  - 重要的是要注意， χ^2的计算使用绝对频数（即：计数，count）。

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 检验统计量的计算
  - χ^2 = 15.795

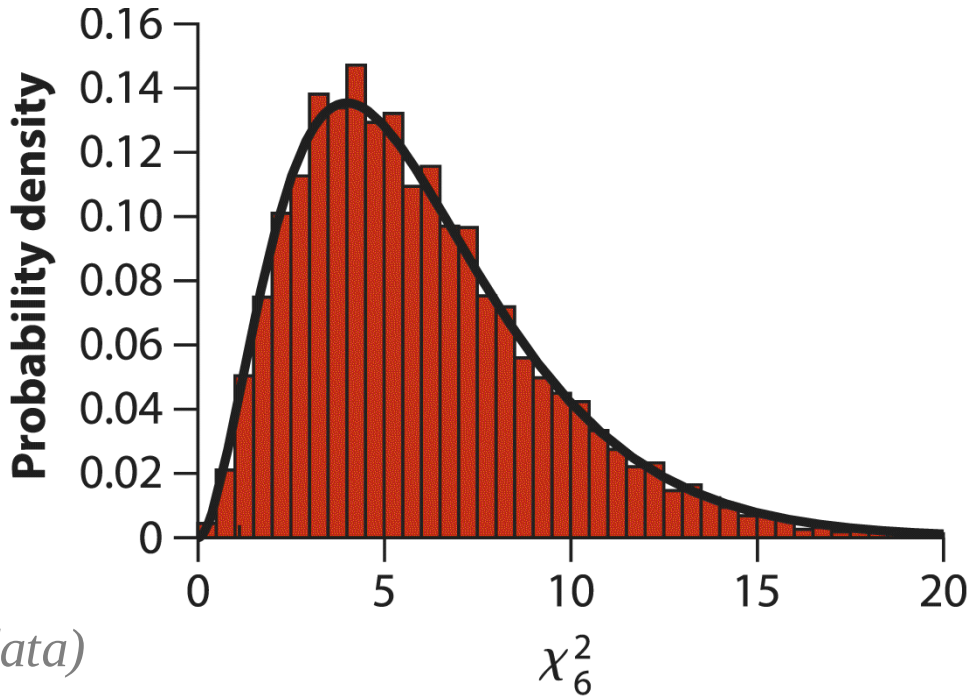
$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Day	Observed number of births	Expected number of births	$(O - E)^2/E$
Sunday	14	25.574	5.238
Monday	26	25.574	0.007
Tuesday	34	25.574	2.776
Wednesday	21	25.574	0.818
Thursday	27	25.574	0.08
Friday	38	26.066	5.464
Saturday	20	26.066	1.412
Total	180	180	15.795

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 检验统计量 χ^2 = 15.795
- 如何确定P值——首先需要获取零分布
  - 零假设下 χ^2 的抽样分布
    - 1. 计算机模拟抽样（大量重复）
      - 红色直方图
    - 2. 理论概率曲线（更为平滑）
      - 有数学形式
      - 基于自由度
        - the degrees of freedom (*df*)

$$df = (\text{Number of categories}) - 1 - (\text{Number of parameters estimated from the data})$$

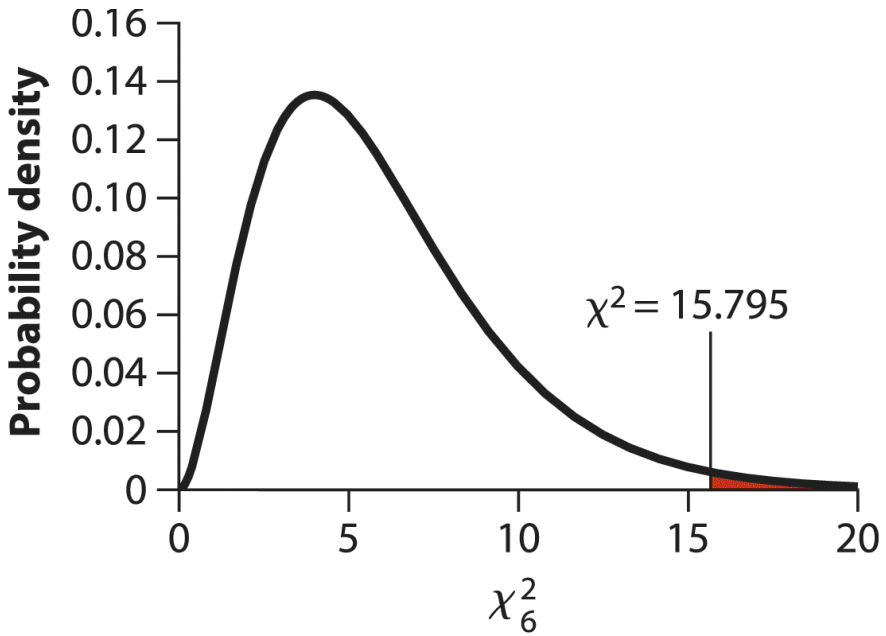


3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 基于零分布 χ^2_6 计算P值
  - 观察数据计算得出的到的 χ^2 = 15.795
  - P值 = 获得大于或等于观察数据χ^2值的概率
  - 只使用 χ^2 分布的右尾来计算P值
    - 如果偏离零假设 → $\chi^2 > 0$
  - χ^2 分布是一个连续概率分布，因此概率等于曲线下的面积
    - 而不是曲线的高度
    - 方法1：大多数计算机统计软件将直接计算出P值 → P=0.0149

```
chisq.test(birthDayTable, p = c(52,52,52,52,52,53,53)/366)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  birthDayTable
## X-squared = 15.795, df = 6, p-value = 0.0149
```

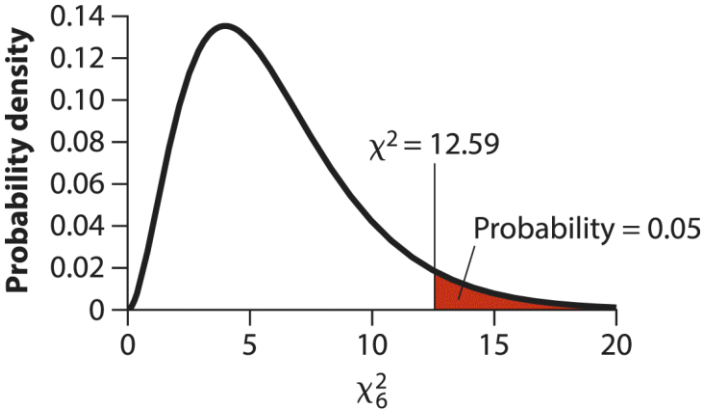


3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 基于零分布 χ^2_6 计算P值
  - 方法2：比较关键值 (critical values) 和观测数据的 χ^2 值 (15.795)
    - 关键值：标志着在零分布指定区域边界的检验统计量的数值。
    - 可通过卡方分布表查询关键值：

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- $\Pr[\chi^2_6 \geq 12.59] = 0.05$
- $\Pr[\chi^2_6 \geq 14.45] = 0.25$
- $\Pr[\chi^2_6 \geq 16.81] = 0.01$

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 基于零分布 χ^2_6 计算P值
  - 方法2：比较关键值和观测数据的 χ^2 (= 15.795)
    - 关键值：标志着在零分布指定区域边界的检验统计量的数值.

- $\chi^2 > 12.59$ → P值 < 0.05
- $\chi^2 > 14.45$ → P值 < 0.25
- $\chi^2 < 16.81$ → P值 > 0.01

关键值

- $\Pr[\chi^2_6 \geq 12.59] = 0.05$
- $\Pr[\chi^2_6 \geq 14.45] = 0.25$
- $\Pr[\chi^2_6 \geq 16.81] = 0.01$

→ 最后的P值估计：0.01 < P值 < 0.25

3.1 卡方检验 the χ^2 goodness-of-fit test ——出生日期

- 提出零假设和备择假设
  - H_0：出生频数与每周该天在一年中出现的次数成比例
  - H_A：出生频数与每周该天在一年中出现的次数不成比例
- 计算检验统计量 χ^2 = 15.795
- 基于零分布 χ^2_6 计算P值
  - 方法1：计算机统计软件直接给出 → P值 = 0.0149
  - 方法2：通过关键值对比 → 0.01 < P值 < 0.25
- 得出结论
  - 出生频数并非均匀地分布在一周中的每一天

3.2 卡方检验的前提/假设 (assumption)

- 数据是基于来自总体的一个随机样本
  - 这是所有检验的假设/前提之一

- 此外，检验统计量 χ^2 的抽样分布要满足（接近）理论的χ^2分布，这需要期望频数满足以下条件：
  - 任何类别的期望频数都不应低于1
  - 不应有超过20%的类别的期望频数低于5

- 否则，检验结果不可靠 (unreliable)

3.2 卡方检验的前提/假设 (assumption)

- 数据是基于来自总体的一个随机样本

- 如果期望频数不满足所需条件时，可以选择：
  - 1. 合并一些期望频率较低的类别，
    - 以得到更少的类别，其期望频率较高
    - 相应地更改自由度
  - 2. 其它检验方法

3.2 卡方检验和二项分布检验

- 当数据只有两个类别时，当n较小并且期望频率太低以至于不符合 χ^2 拟合优度检验的假设时，二项检验是最佳选择。
  - 即使n较大，当有计算机可用时，可优先选择二项检验
  - 因为它提供了精确的P值

3.3 *Break – R coding*

- Frequency table
  - > birthDayTable <- table(birthDay$day)
  - > # need to convert the variable into a factor

- χ^2 goodness-of-fit test
  - > chisq.test(birthDayTable, p = c(52,52,52,52,52,53,53)/366)

- Expected table
  - > exp_prob = c(52,52,52,52,52,53,53)/366
  - > chisq.test(birthDayTable, p = exp_prob)$expected

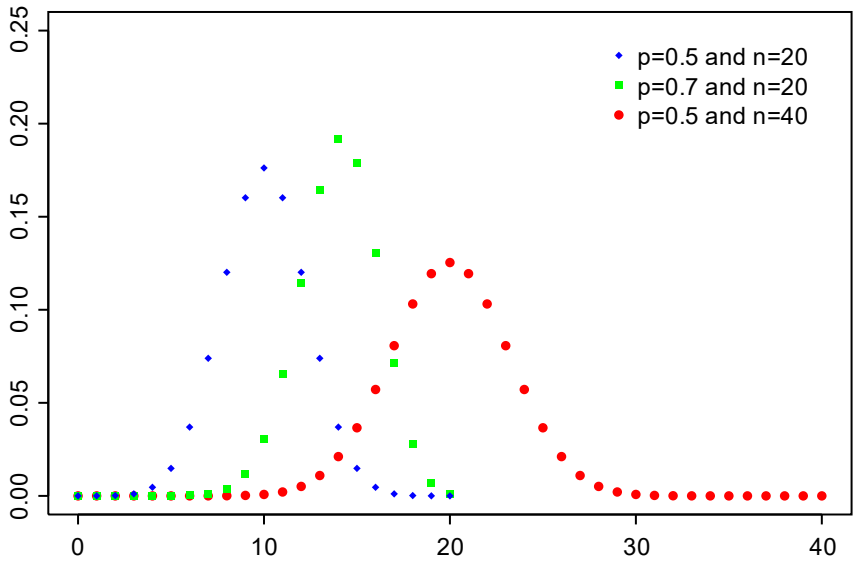
4. 概率分布

- 生物学家经常将他们的数据与某种概率分布拟合，这些分布也代表了自然现象的简单模型。
  - 在许多生物学研究中，概率分布常常被用作零假设。
  - 比例模型不是唯一使用适配度检验方法来检验的概率模型。

- 在这里，所谓的“模型”是指数学描述，模拟我们认为自然过程是如何工作的
  - 或者至少在没有复杂因素的情况下是如何工作的

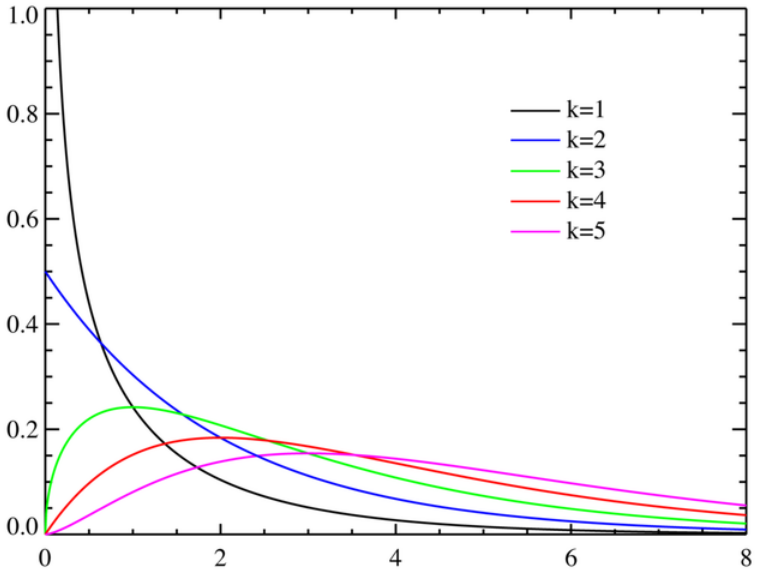
4. 概率分布 — 概率密度函数

二项分布



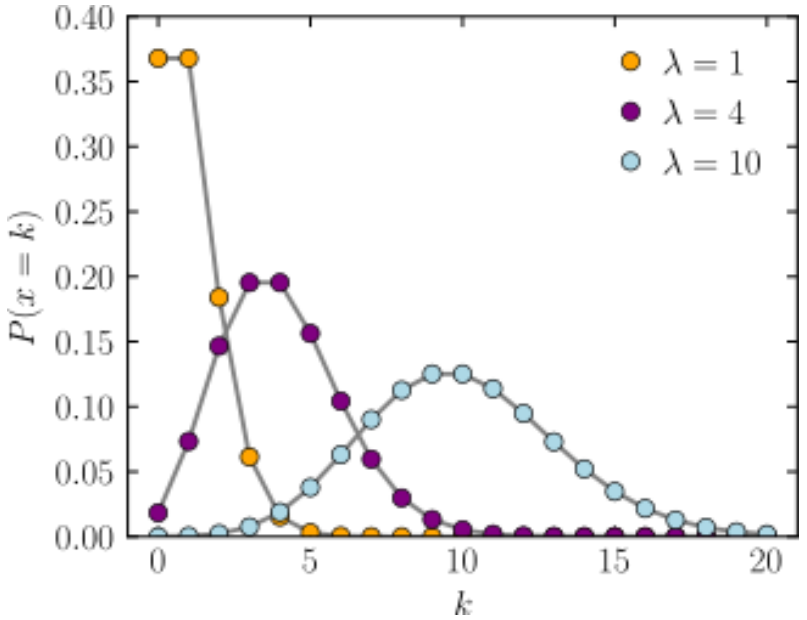
记号	$B(n,p)$
参数	$n \geq 0$ 试验次数 (整数) $0 \leq p \leq 1$ 成功概率 (实数)
值域	$k \in \{0,\ldots,n\}$
概率质量函数	$\binom{n}{k} p^k (1-p)^{n-k}$

卡方分布



参数	$k \in \mathbb{N}^\star$ 自由度
值域	$x \in [0;+\infty)$,
概率密度函数	$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$

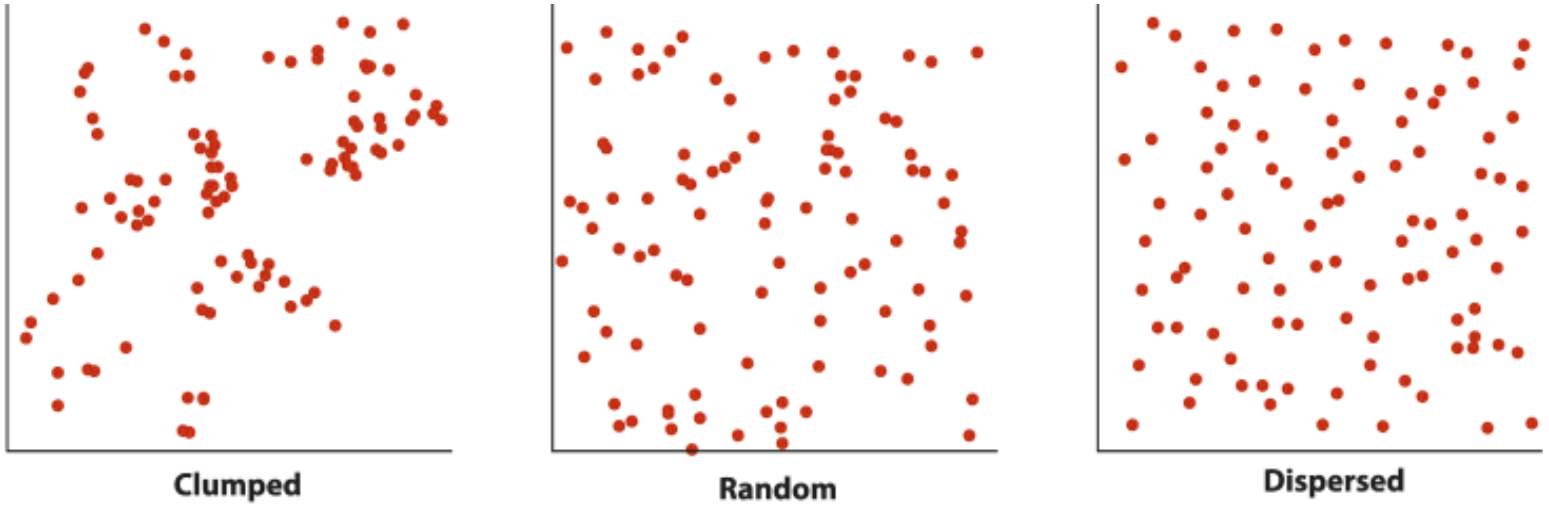
泊松分布



参数	$\lambda > 0$ (实数)
值域	$k \in \{0,1,2,3,\cdots\}$
概率质量函数	$\frac{\lambda^k}{k!} e^{-\lambda}$

4.1泊松分布 The Poisson distribution

- 泊松分布描述了在一段时间或空间内成功事件的数量
  - 当成功事件间相互独立
  - 事件在每个瞬间或空间点发生的概率相等



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.1泊松分布 The Poisson distribution

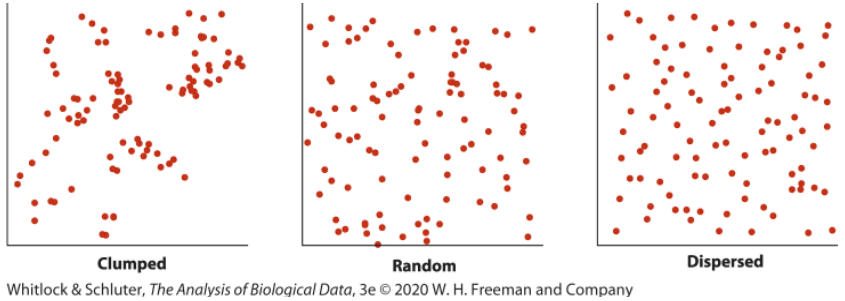
- 泊松分布是由法国数学物理学家西梅昂·丹尼·波瓦松（Siméon Denis Poisson）推导出的。他表明，在任何给定的时间或空间块中，成功事件发生X次的概率为：

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

  - μ：时间或空间中独立成功的平均数量
    - rate： 表示每单位时间或每单位空间的计数
  - e：自然对数的底
    - 一个约等于2.718的常数
  - X!： X的阶乘

4.1泊松分布 The Poisson distribution

- 泊松分布 = 随机分布



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 对于生物学家来说，泊松分布只是一种模型，用于描述自然中某类事件（成功）是否在时间和空间中随机分布。 其它可能:
  - 聚集分布
    - 例如，传染病的爆发通常会导致病例在空间上呈现聚集分布，因为个体会从更接近的人/邻居那里传染疾病。
  - 分散分布
    - 例如，领地性动物通常在空间上比随机分布更为分散，因为个体会互相驱赶。
- 与随机模式的偏差可以帮助我们识别导致这些模式的有趣生物过程。

4.2泊松分布 — 检验随机性 randomness

- 在生物学中，泊松分布的主要用途是提供一个零假设，用来测试成功是否在时间或空间中“随机”发生。

- 但在实践中，通常我们不知道成功发生的确切概率。

- 因此，要从泊松分布中对不同结果的概率进行预测，我们必须首先从数据中估计出成功事件的发生概率。

4.2泊松分布 — 检验随机性 randomness

- 例子：灭绝事件在事件尺度上的分布
  - 在76个长度类似的时间段/间隔，海洋无脊椎动物的灭绝数量/次数。
    - 地球历史上最好的灭绝记录来自化石化的海洋无脊椎动物，因为它们有坚硬的壳，因此往往保存得很好。
  - 如果家族的灭绝在化石记录中是“随机”的，那么每个时间段内的灭绝数量应该遵循泊松分布。

Number 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
Sum	76

4.2泊松分布 — 检验随机性 randomness

- 例子：灭绝事件在事件尺度上的分布
  - 在76个长度类似的时间段中，海洋无脊椎动物的灭绝数量/次数。
    - 是离散型的数值数据
  - 与泊松分布的偏离可能表明灭绝在时间上倾向于聚集并以爆发的方式发生（大规模灭绝）（clumped）。
  - 另一个可能性是，灭绝在时间上分布得比我们预期的更均匀（dispersed）。

Number 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
Sum	76

4.2泊松分布 — 检验随机性 randomness

- 例子：灭绝事件在事件尺度上的分布
  - 检验灭绝随机性最简单的方法是使用χ^2拟合度检验
    - 将灭绝的频数分布与从泊松分布预期的分布进行比较。
  - 假设
    - H_0: 时间间隔内的灭绝数量符合泊松分布。
    - H_A: 时间间隔内的灭绝数量不符合泊松分布。
  - 计算参数
    - 每个时间间隔内的平均灭绝数量：

$$\overline{X} = \frac{(0\times 0)+(13\times 1)+(15\times 2)+\cdots}{76} = 4.21$$

Number 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
Sum	76

4.2泊松分布 — 检验随机性 randomness

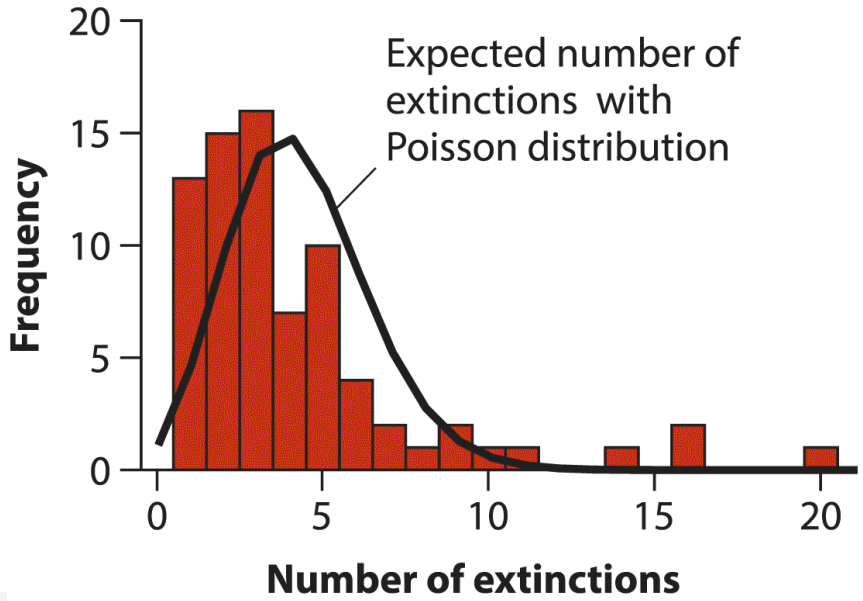
- 例子：灭绝事件在事件尺度上的分布
  - 每个时间间隔内的平均灭绝数量（对μ的估计）：

$$\overline{X} = \frac{(0\times0)+(13\times1)+(15\times2)+\cdots}{76} = 4.21$$

  - 灭绝数量的频率分布（直方图），与具有相同均值的泊松分布所期望的频数进行比较（曲线）。

$$\Pr[3\ extinctions] = \frac{e^{-\overline{X}}\,\overline{X}^3}{3!} = \frac{e^{-4.21}(4.21)^3}{3!} = 0.1846$$

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



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

```
expectedProportion <- dpois(0:20, lambda = meanExtinctions)
expectedFrequency <- expectedProportion * 76
```

https://whitlockschluter3e.zoology.ubc.ca/RExamples/Rcode_Chapter_8.html

4.2泊松分布 — 检验随机性 randomness

- 例子：灭绝事件在事件尺度上的分布
  - 基于泊松概率分布来计算期望频数
  - 我们将所有X ≥ 8 或X < 2的灭绝事件分为一个类别
    - 因为更大灭绝数量的预期频率已经非常小。

Number of extinctions (X)	Observed frequency	Expected probability	Expected frequency
0 or 1	13	0.077	5.88
2	15	0.132	10.00
3	16	**0.185**	14.03
4	7	0.194	14.77
5	10	0.164	12.44
6	4	0.115	8.73
7	2	0.069	5.25
8 or more	9	0.065	4.91
Sum	76	1	76

4.2泊松分布 — 检验随机性 randomness

- 例子：灭绝事件在事件尺度上的分布
  - 计算检验统计量 χ^2 = 23.93
  - *df* = ?
    - 8个类别
    - 估计了1个参数：均值
    - $df = (\text{Number of categories}) - 1 - (\text{Number of parameters estimated from the data})$

      $= 8 - 1 - 1 = 6$
  - 1. 关键值: $\chi^2_{\alpha=0.05,\, df=6}$ = 12.59; $\chi^2_{\alpha=0.001,\, df=6}$ = 22.46
    - P值 < 0.001
  - 2. 计算机软件可直接得出： P值 = 0.0005

```
saveChiTest <- chisq.test(obsFreqGroup, p = expFreqGroup/76)

pValue <- 1 - pchisq(saveChiTest$statistic, df = 6)
```

4.2泊松分布 — 检验随机性 randomness

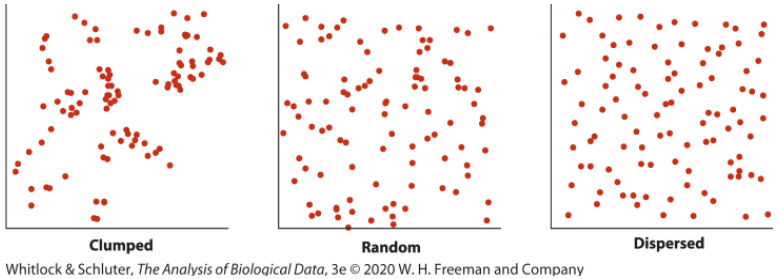
- 例子：灭绝事件在事件尺度上的分布
  - χ^2 = 23.93；P值 < 0.001 或 P值 = 0.0005
  - 我们如何描述一个偏离泊松分布的模式?
    - 泊松分布的一个特性是每个时间段内成功事件的次数的方差（variance, 标准差的平方）等于均值（μ）。因此，聚集或分散的模式通过方差与平均成功事件数量的比值来度量。
    - 如果方差大于均值，那么分布是聚集的 (clumped)。
    - 如果方差小于均值，那么分布是分散的 (dispersed)。
  - 结论：灭绝事件往往以爆发的方式发生（大规模灭绝, mass extinctions ）。

Variance:mean ratio

```
var(extinctData$numberOfExtinctions) / meanExtinctions
```

```
## [1] 3.257333
```

4.3 基于泊松分布来检验空间或时间上的随机性



- 步骤：
  - 提出零假设和备择假设
  - 估算参数：单位时间/空间内平均发生次数的均值
  - 基于估计的参数来计算泊松分布的期望概率和期望频数
    - 期望频数 = 期望概率*总的时间/空间单元数
  - 使用χ^2拟合度检验比较观察频数和期望频数
    - χ^2 = ?
    - P值？
  - 结论：描述分布模式：是否偏离泊松分布的随机模式?
    - 如果方差大于均值，那么分布是聚集的 (clumped)。
    - 如果方差小于均值，那么分布是分散的 (dispersed)。

MASS EXTINCTIONS

A mass extinction is a sharp spike in the rate of extinction of species caused by a catastrophic event or rapid environmental change. Scientists have been able to identify five mass extinctions in Earth's history, each of which led to a loss of more than 75 percent of animal species.

NATIONAL GEOGRAPHIC

1. ORDOVICIAN-SILURIAN EXTINCTION 440 MILLION YEARS AGO (MA)

86% LOSS

Scientists theorize that there were two main phases to this extinction: a glaciation event and a heating event. Abundant plant life removed carbon dioxide (CO_2) from the air, causing global cooling and glacier formation. This led to a drop in sea levels, reducing habitat. Later came global warming and sea level rising again. Creatures that had adapted to the cooler climate were unable to survive the increased temperature. Since most fauna was marine at the time, 86% of life was lost.

2. LATE DEVONIAN EXTINCTION—365 MA

75% LOSS

About 75% of life died off during this period. One theory suggests that land plants developed deep roots, releasing an abundance of nutrients into the oceans that fed algae. Because of this, algae blooms consumed vast amounts of oxygen (O_2) in the oceans, suffocating many species. Another theory suggests that another global cooling took place, resulting in glaciation and a fall in sea level, leading to habitat loss.

3. PERMIAN-TRIASSIC EXTINCTION—252 MA

96% LOSS

The Permian-Triassic was the deadliest extinction in history: 96% of all life perished. Scientist believe that volcanic activity in Siberia put massive amounts of carbon dioxide, a greenhouse gas, into the atmosphere. Bacteria that thrive on CO_2 began producing methane, another greenhouse gas. Large quantities of both gases warmed the planet and combined with Earth's water, making the ocean and rain acidic, creating a highly toxic environment for life.

4. TRIASSIC-JURASSIC EXTINCTION—201.3 MA

80% LOSS

Some scientists theorize that volcanic eruptions spewed tons of CO_2 into the atmosphere, which trapped heat and acidified the oceans, causing this mass extinction. Other scientists contend that an asteroid or comet impact triggered the extinction. About 80% of life was lost in this extinction, including most of the mammals, making way for the dinosaur's ancestors.

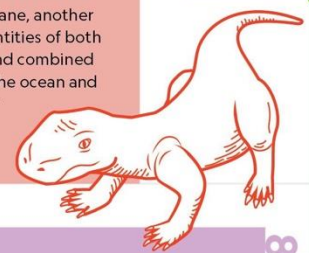
5. CRETACEOUS-PALEOGENE EXTINCTION—66 MA

60-76% LOSS

The Cretaceous-Paleogene extinction wiped out the dinosaurs, along with 60–76% of all life on Earth. A widely accepted theory is that an asteroid landed in the Yucatán Peninsula in Mexico and killed the dinosaurs. The impact would have ejected enormous amounts of debris into the atmosphere, causing global temperatures to drop. The impact may have also caused local fires, earthquakes, tsunamis, and acid rain.

6. HOLOCENE EXTINCTION 11,700 YEARS AGO TO PRESENT

The next mass extinction may already be happening. The current extinction rate is at least a thousand times greater than the "normal" extinction rate. A "normal" or background extinction rate is the average rate of extinction based on the longevity of species through time without human influence, determined by the fossil record. Scientists believe that human impact on the environment is the leading cause of extinctions today.



4.4 *Break – R coding*

- Mean rate

```
> meanExtinctions <- mean(extinctData$numberOfExtinctions)
```

- Expected frequency

```
> expectedProportion <- dpois(0:20, lambda = meanExtinctions)
```

- Expected frequency

```
> # after combining groups
> saveChiTest <- chisq.test(obsFreqGroup, p = expFreqGroup/76)
> pValue <- 1 - pchisq(saveChiTest$statistic, df = 6)
```

5. 小结 – 比例模型

- 二项分布表示在 n 次试验中获得 X 次成功的概率
  - 假设每次试验都是独立的，成功的概率（p）相同。
  - 样本比例是对总体比例的最佳估计。
  - 根据大数定律，非 常大的样本的成功比例将接近总体的真实比例。

$$\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$$

- 二项检验将数据集中观察到的成功次数与零假设下的预期次数进行比较。
  - 在 H_0 条件下，成功次数的零分布是二项分布。
  - 因此可用二项公式计算检验的 P 值。
  - 比例的置信区间可以使用 Agresti-Coull 方法计算。

5. 小结 – 比例模型

- χ^2 拟合度检验将离散变量或分类变量的频数分布与概率模型的预期频数进行比较。
  - χ^2 拟合度检验比二项检验更通用，因为它可以处理两个以上的类别。即使只有两个类别，它也更容易计算。
  - χ^2 检验统计量有一个近似于理论 χ^2 分布的零分布。前提是所有预期频数不小于 1，且小于 5 的预期频数的类别不超过 20%（可能有必要合并某些低频数类别来满足条件）。

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

- 理论上的 χ^2 分布是一种连续分布。
  - 概率用曲线下的面积来衡量。
  - 在比例模型下，事件按机会数的比例归入不同类别。
  - 拒绝 H_0 意味着事件发生的概率不与和预期频数那样成正比（proportional）。

5. 小结 – 比例模型

- 泊松分布描述了在时间或空间间隔中成功事件的频数分布
  - 假设成功事件在时间或空间上以相等的概率独立发生。
  - 拒绝成功次数泊松分布的零假设，意味着成功事件之间不是独立的，或者成功次数在时间或空间上的概率不是固定的。

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

- 时间或空间上偏离随机性的方向
  - 通过将每个时间或空间间隔的成功事件次数的方差与平均成功次数进行比较来衡量。
  - 如果方差大于平均值，则说明成功事件是聚集的（clumped）；
  - 如果方差小于平均值，则说明成功事件的分布比泊松分布预期的分散（dispersed）。

6. 课堂练习

- Ch7 - Q1. 计算练习：二项式概率。肠球菌是人类正常肠道菌群的一部分，但某些菌株可能引起疾病。在一个医院里，30%的致病分离物对抗生素万古霉素具有抗药性（Wenzel 2004）。假设已从患者体内提取了七个独立的致病分离物，并对其抗药性进行了测试。使用以下步骤，计算其中五个或更多分离物对万古霉素具有抗药性的概率：
  - a. 二项分布的假设是什么？这个例子是否符合这些假设?
  - b. 使用二项分布，对于这个例子，什么是成功的概率？n 是多少?
  - c. 使用二项分布，计算恰好有五个对万古霉素具有抗药性的分离物的概率。
  - d. 计算恰好有六个对万古霉素具有抗药性的分离物的概率，然后计算恰好有七个对万古霉素具有抗药性的分离物的概率。
  - e. 使用加法原理，结合前面的答案，计算七个分离物中有五个或更多对万古霉素具有抗药性的概率。

6. 课堂练习

- Ch8 - Q2. 寄生线虫 *Camallanus oxycephalus* 会感染许多淡水鱼，包括鲥鱼(shad)。下表列出了每条鱼的线虫数量（Shaw 等，1998 年）：
  - a. 绘制数据图表。哪种图表最合适?
  - b. 计算如果线虫 "随机"（即独立且概率相等）感染鱼类的预期频率。
  - c. 将预期频率叠加到您的图表上。有什么明显不同?
  - d. 是否有证据表明线虫不是随机感染鱼类的? 请在此处并始终显示假设检验的所有四个步骤。

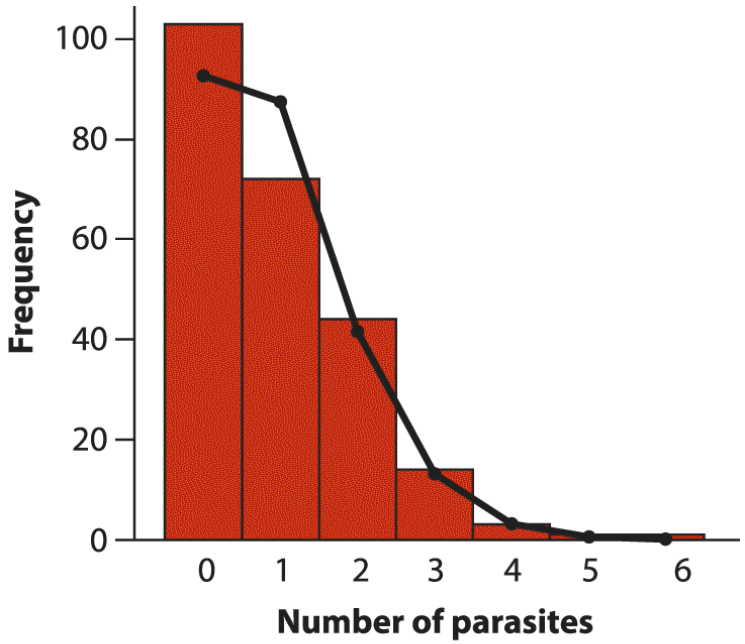
线虫数量	鱼的频数
0	103
1	72
2	44
3	14
4	3
5	1
6	1

6. 课堂练习

- Q1 – 答案
- a. 二项分布的假设包括有固定数量的独立试验（n），每次试验有两种可能的结果（成功或失败），且成功的概率在每次试验中保持不变。在这个例子中，我们有七个独立的试验，每次试验代表从患者中提取一个致病分离物，结果是对万古霉素具有抗药性（成功）或不具有抗药性（失败）。因此，这个例子符合二项分布的假设。
- b. 这个例子中的成功是分离物对万古霉素具有抗药性，成功的概率是给定的，即30%或0.3。试验的总数（n）是7，因为有7个独立的试验。
- c. 使用二项分布计算恰好有五个对万古霉素具有抗药性的分离物的概率，可以使用二项概率公式：Pr(X = 5) = (n choose X) * p^X * (1-p)^(n-X)
  - 其中，X 是成功的数量，这里是5，n 是试验的总数，p 是成功的概率0.3，(n choose X) 表示二项系数，可以计算为：(7 choose 5) = 7! / (5!(7-5)!) = 21；
  - 所以：Pr(X = 5) = 21 * (0.3)^5 * (0.7)^(7-5) = 0.0250。
- d. 类似地， Pr(X = 6) = 0.0036； Pr(X = 7) =0.0002。
- e. 使用加法原理，将 Pr(X = 5) + Pr(X = 6) + Pr(X = 7) = 0.0288。

6. 课堂练习

- Ch8 - Q2. 寄生线虫 *Camallanus oxycephalus* 会感染许多淡水鱼，包括鲥鱼(shad)。下表列出了每条鱼的线虫数量（Shaw 等，1998 年）：
  - a. 直方图
  - b. 首先，我们需要估算每条鱼的平均线虫数量：103(0)+72(1)+44(2)+14(3)+3(4)+1(5)+1(6)=0.945.
  - 使用泊松公式得到 0 到 6 条线虫的鱼的预期概率和预期频率：92.58, 87.35, 41.41, 13.09, 3.09, 0.48, 0.00。
  - c. （见图）观察到的鱼类感染 0 条线虫的频率略高于预期，而感染 1 条线虫的频率略低于预期。在有 2 至 6 条线虫的鱼类中，观察到的频率与预期频率相似。
  - d. 假设检验
    - H0：每条鱼的线虫数量呈泊松分布。
    - HA：每条鱼的线虫数量不呈泊松分布。
    - 我们需要将 4、5 和 6 条寄生虫对应的频率合并，以满足 χ^2 线虫频率标准，得出 χ^2 = 4.67, *df*=5-1-1=3；而关键值 $\chi^2_{\alpha=0.05,\,df=3}$ =7.81。
    - 由于 χ^2 小于临界值，P>0.05；在计算机上，P=0.20。
    - 我们不拒绝 H_0，即没有证据表明线虫不是随机进入鱼体内的。



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company