

Случайность в пространстве или времени

Основы биостатистики, осень 2022

Марина Варфоломеева

- Пропорциональная модель. Хи-квадрат статистика
- Распределение Пуассона. Проверка соответствия распределению Пуассона
- Расположение в пространстве. Сравнение дисперсии и среднего

Вероятностные модели

Вероятностные модели (probabilistic models) описывают вероятности событий.

Тесты адекватности модели (goodness-of-fit tests) — описывают, насколько наблюдаемые значения соответствуют теоретическому распределению.

Биномиальный тест — тест адекватности только для бинарных величин.

Сегодня другие тесты.

Пропорциональная модель Хи-квадрат статистика

Пропорциональная модель

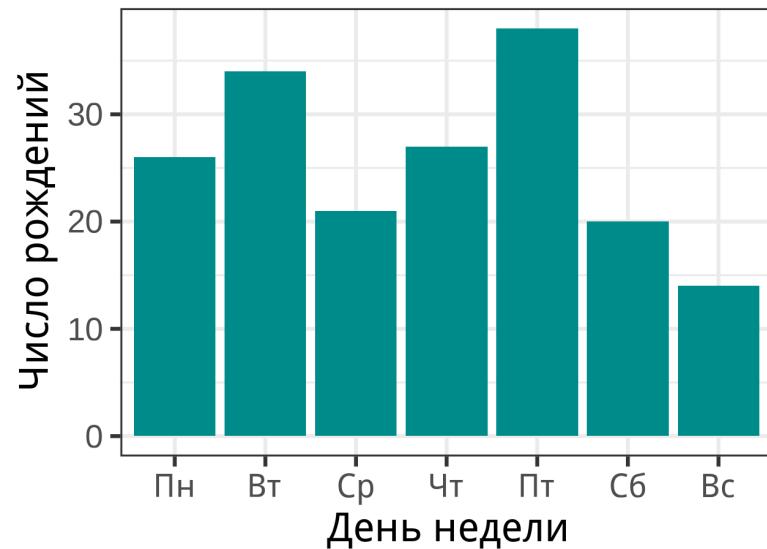
Пропорциональная модель (proportional model) — вероятностная модель, в которой вероятность события пропорциональна числу возможностей его возникновения.

- гены сперматогенеза на X хромосоме

Пример: дни рождения

Дни недели, на которые пришлось рождение, в случайной выборке из 180 младенцев 2016 году (данные the U.S. National Center for Health Statistics; Martin et al. 2018).

День недели	Число рождений
Пн	26
Вт	34
Ср	21
Чт	27
Пт	38
Сб	20
Вс	14



Согласно пропорциональной модели вероятность рождения в разные дни недели д.б. пропорциональна их числу в году.

Так ли это?

χ^2 -тест

Гипотезы в общем виде:

χ^2 -тест оценивает соответствие наблюдаемого частотного распределения теоретическому (нулевой вероятностной модели).

H_0 : — вероятности (или частоты) в генеральной совокупности соответствуют нулевой модели

H_A : — вероятности (или частоты) в генеральной совокупности какие-то другие.

Ожидаемые частоты

День недели	Число дней в 2016	Доля в 2016	Ожидаемая частота рождений
Пн	52	52 / 366	25.6
Вт	52	52 / 366	25.6
Ср	52	52 / 366	25.6
Чт	52	52 / 366	25.6
Пт	53	53 / 366	26.1
Сб	53	53 / 366	26.1
Вс	52	52 / 366	25.6
Сумма	366	1	180.0

Ожидаемая частота

$$Expected = N \cdot p_{expected}$$

Сумма ожидаемых частот д.б. равна объему выборки N.

χ^2 -статистика

χ^2 измеряет, насколько наблюдаемые частоты соответствуют ожидаемым

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$\chi^2 = 0$, когда данные соответствуют ожиданиям при H_0

$\chi^2 > 0$, если данные отклоняются от ожиданий при H_0

χ^2 -статистика

χ^2 измеряет, насколько наблюдаемые частоты соответствуют ожидаемым

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$\chi^2 = 0$, когда данные соответствуют ожиданиям при H_0

$\chi^2 > 0$, если данные отклоняются от ожиданий при H_0

Гипотезы в χ^2 -тесте:

H_0 : — вероятности (или частоты) в генеральной совокупности соответствуют нулевой модели

H_A : — вероятности (или частоты) в генеральной совокупности какие-то другие.

χ^2 -статистика

χ^2 измеряет, насколько наблюдаемые частоты соответствуют ожидаемым

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$\chi^2 = 0$, когда данные соответствуют ожиданиям при H_0

$\chi^2 > 0$, если данные отклоняются от ожиданий при H_0

Гипотезы в χ^2 -тесте:

H_0 : — вероятности (или частоты) в генеральной совокупности соответствуют нулевой модели

H_A : — вероятности (или частоты) в генеральной совокупности какие-то другие.

$$H_0 : \chi^2 = 0$$

$$H_A : \chi^2 > 0$$

Односторонний тест.

Считаем χ^2

День недели	Число рождений	Ожидаемое число рождений	Хи-квадрат
Пн	26	25.6	0.007
Вт	34	25.6	2.776
Ср	21	25.6	0.818
Чт	27	25.6	0.080
Пт	38	26.1	5.464
Сб	20	26.1	1.411
Вс	14	25.6	5.238
Сумма	180	180.0	15.795

Считаем χ^2

День недели	Число рождений	Ожидаемое число рождений	Хи-квадрат
Пн	26	25.6	0.007
Вт	34	25.6	2.776
Ср	21	25.6	0.818
Чт	27	25.6	0.080
Пт	38	26.1	5.464
Сб	20	26.1	1.411
Вс	14	25.6	5.238
Сумма	180	180.0	15.795

$$\chi^2 = 15.795$$

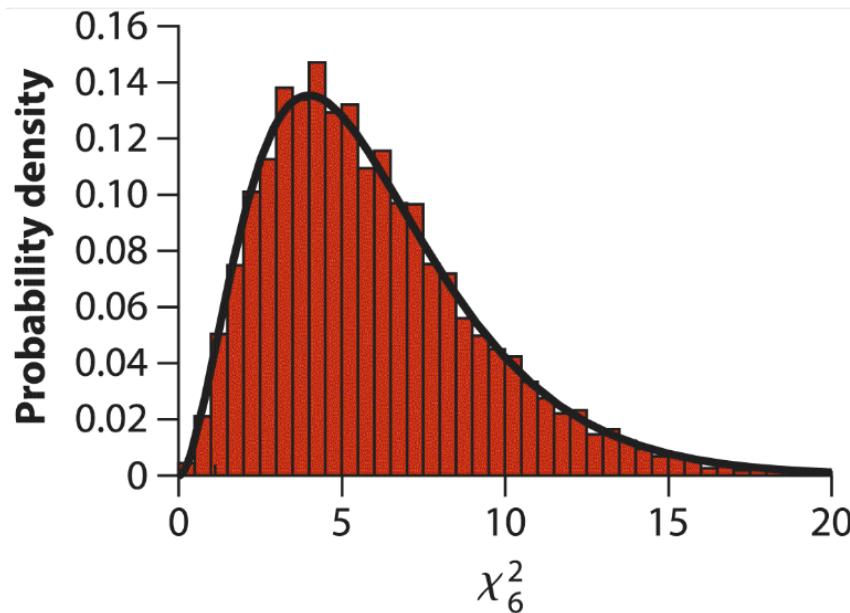
Но с чем нам сравнивать эту величину?

Выборочное распределение χ^2 -статистики при H_0

Во множестве повторных выборок значение χ^2 будет подчиняться распределению χ^2 с числом степеней свободы df

$$df = m - 1 - p$$

- m число категорий
- p число параметров, оцененных по данным (здесь 0)



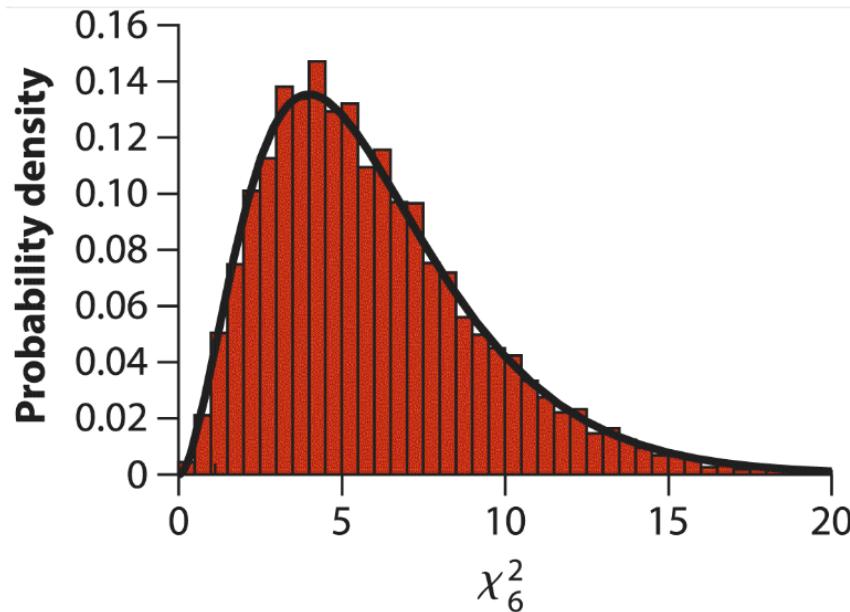
Whitlock, Schluter, 2015

Выборочное распределение χ^2 -статистики при H_0

Во множестве повторных выборок значение χ^2 будет подчиняться распределению χ^2 с числом степеней свободы df

$$df = m - 1 - p$$

- m число категорий
- p число параметров, оцененных по данным (здесь 0)



Whitlock, Schlüter, 2015

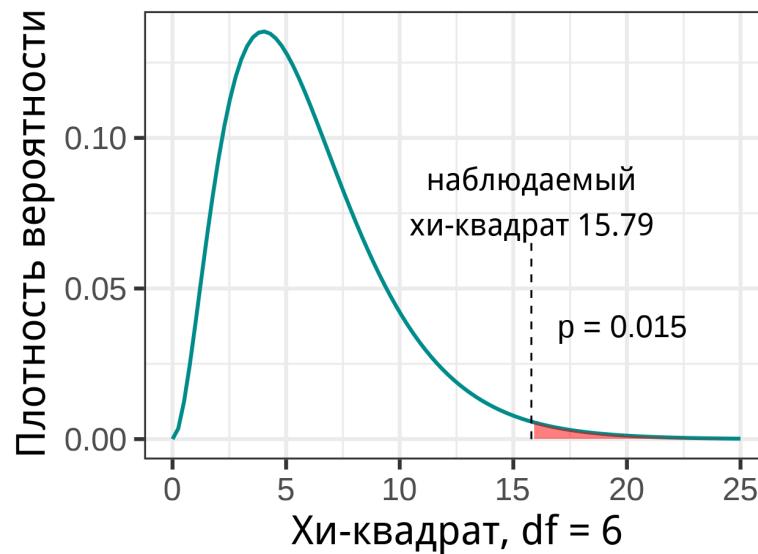
В примере $df = 7 - 1 - 0 = 6$

χ^2 -тест

$H_0 : \chi^2 = 0$ — вероятности (или частоты) в генеральной совокупности соответствуют нулевой модели

$H_A : \chi^2 > 0$ — вероятности (или частоты) в генеральной совокупности какие-то другие.

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$



Условия применимости χ^2 -теста

- наблюдения независимы друг от друга

χ^2 -статистика приблизительно следует χ^2 -распределению, если:

- нет ожидаемых частот < 1
- ≤ 20 ожидаемых частот < 5

Условия применимости χ^2 -теста

- наблюдения независимы друг от друга

χ^2 -статистика приблизительно следует χ^2 -распределению, если:

- нет ожидаемых частот < 1
- ≤ 20 ожидаемых частот < 5

Если условия нарушены:

- можно объединить редкие категории, если они имеют биологический смысл
- использовать непараметрическую статистику или компьютерную симуляцию

Тесты для двух категорий

χ^2 -тест работает для двух категорий (при тех же условиях).

Т.е. χ^2 -тест может заменять биномиальный тест:

- χ^2 быстрее считать
- он менее точен
- более требователен к данным

Распределение Пуассона

Случайное распределение во времени и пространстве

Распределение Пуассона

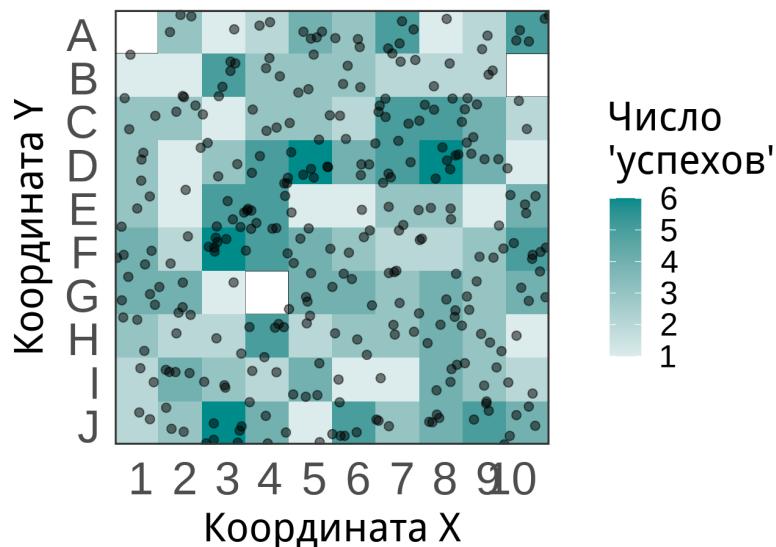
“Счетные” данные:

- Число орлов, выпадающих за 1 минуту
- Число левшей в выборках из 100 человек
- Число заболевших в день (с оговорками)

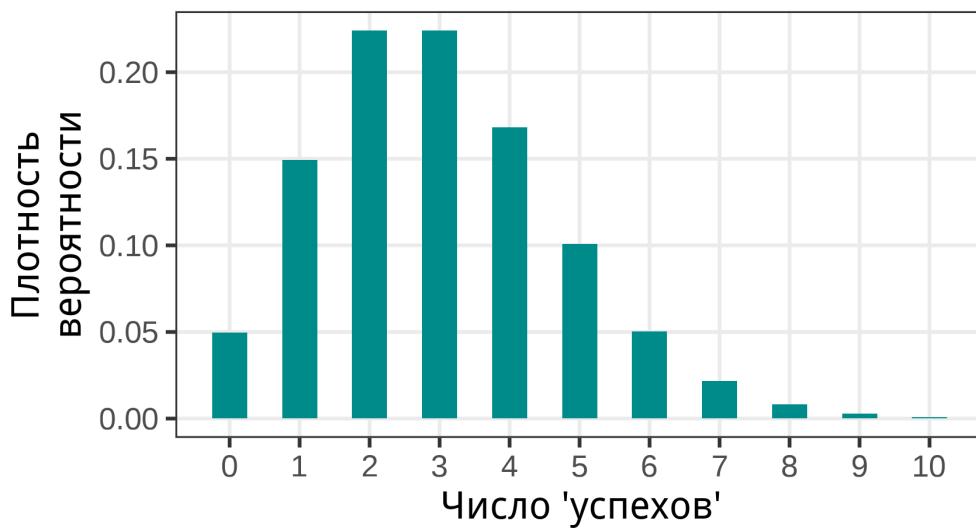
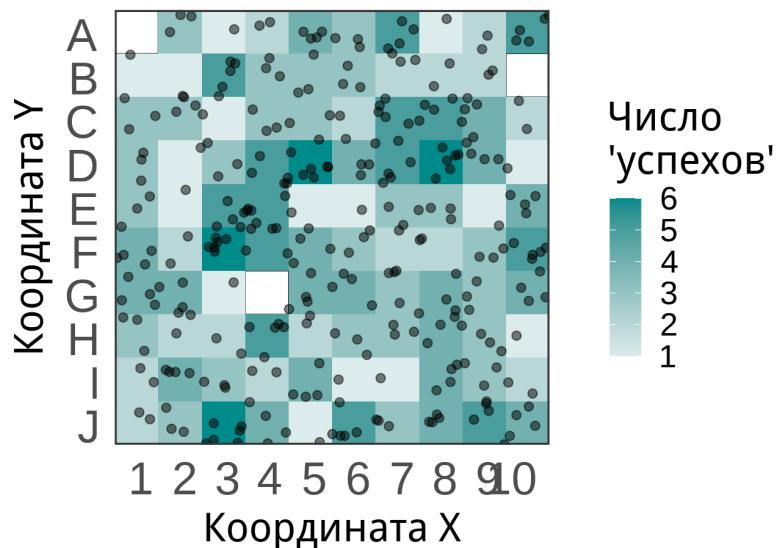
Распределение Пуассона описывает вероятность определенного числа “успехов” за единицу времени или на единицу пространства, если

- (1) испытания независимы
- (2) вероятность успеха постоянна

Распределение Пуассона в пространстве

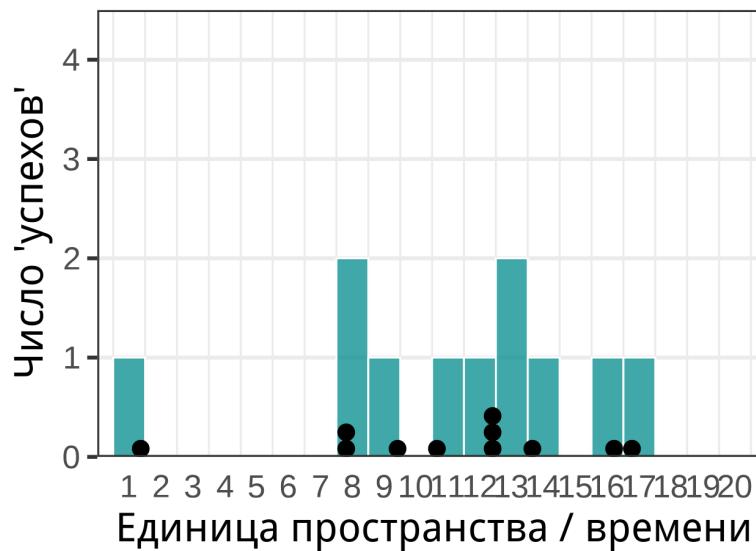


Распределение Пуассона в пространстве

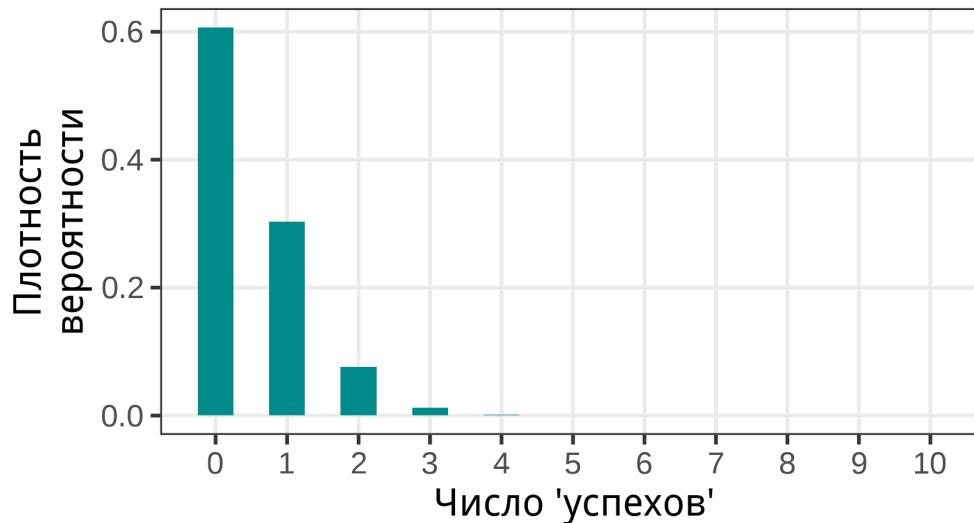
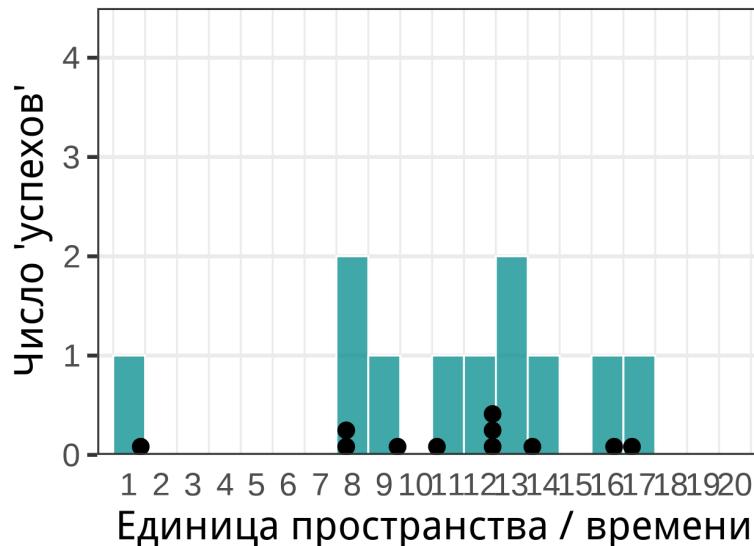


- число деревьев на единицу площади
- число пыльцевых зерен на единицу площади
- число моллюсков на единицу площади

Распределение Пуассона в пространстве / времени

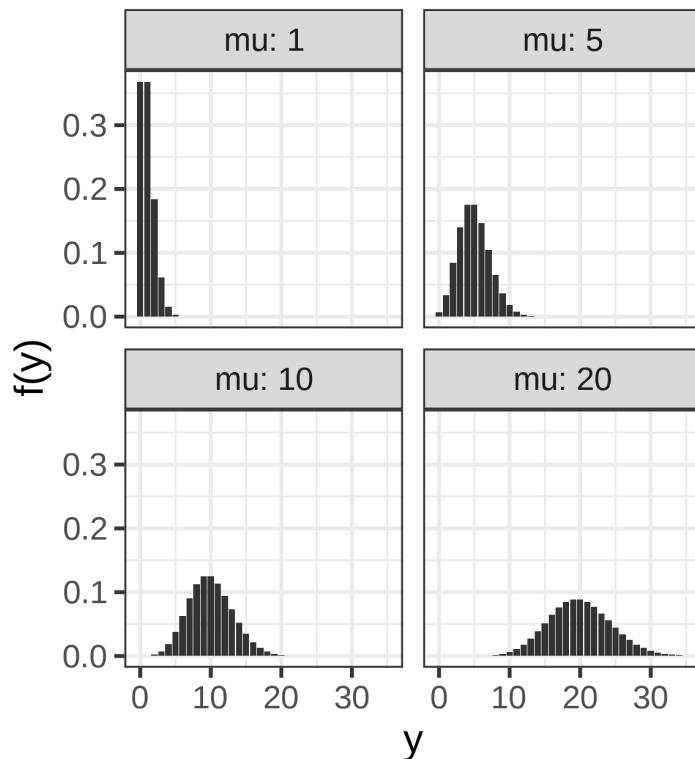


Распределение Пуассона в пространстве / времени



- число пчел, посещающих цветок за 15 минут
- число рождений / госпитализаций / смертей в день
- число сиквенсов, пришедшихся на один ген

Распределение Пуассона



$$P(X \text{ 'успехов'}) = \frac{e^{-\mu} \mu^X}{X!}$$

Параметр μ — определяет и среднее, и дисперсию числа “успехов”

Возможные значения:

$$0 \leq X \leq +\infty, X \in \mathbb{N}$$

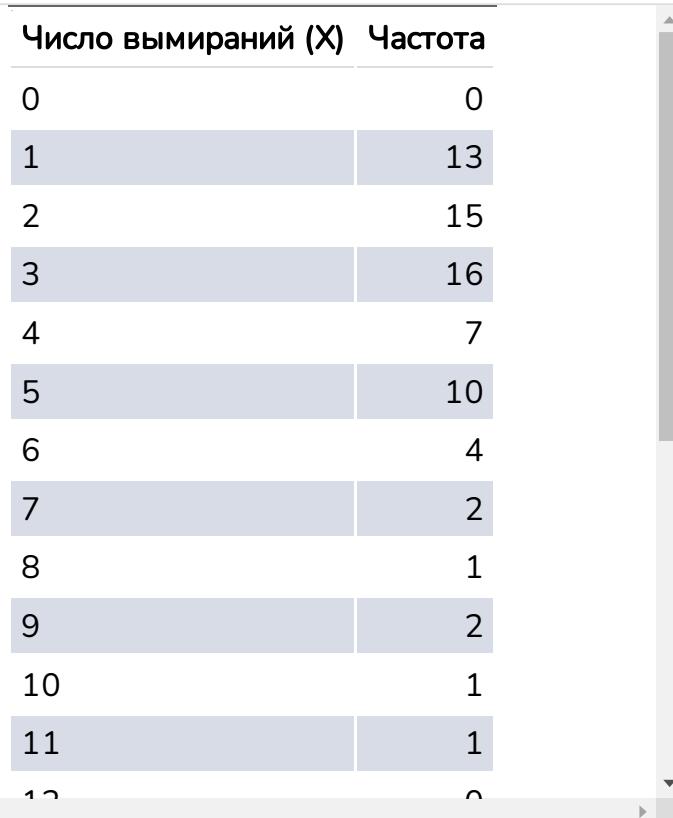
Чем больше среднее, тем больше дисперсия.

Тестируем случайность при помощи
распределения Пуассона

Пример: Вымирания видов в истории Земли

Данные о вымирании семейств морских беспозвоночных за 76 отрезков времени (Raup, Sepkoski, 1982).

Случаются ли вымирания “равномерно” или бывают периоды массовых вымираний?

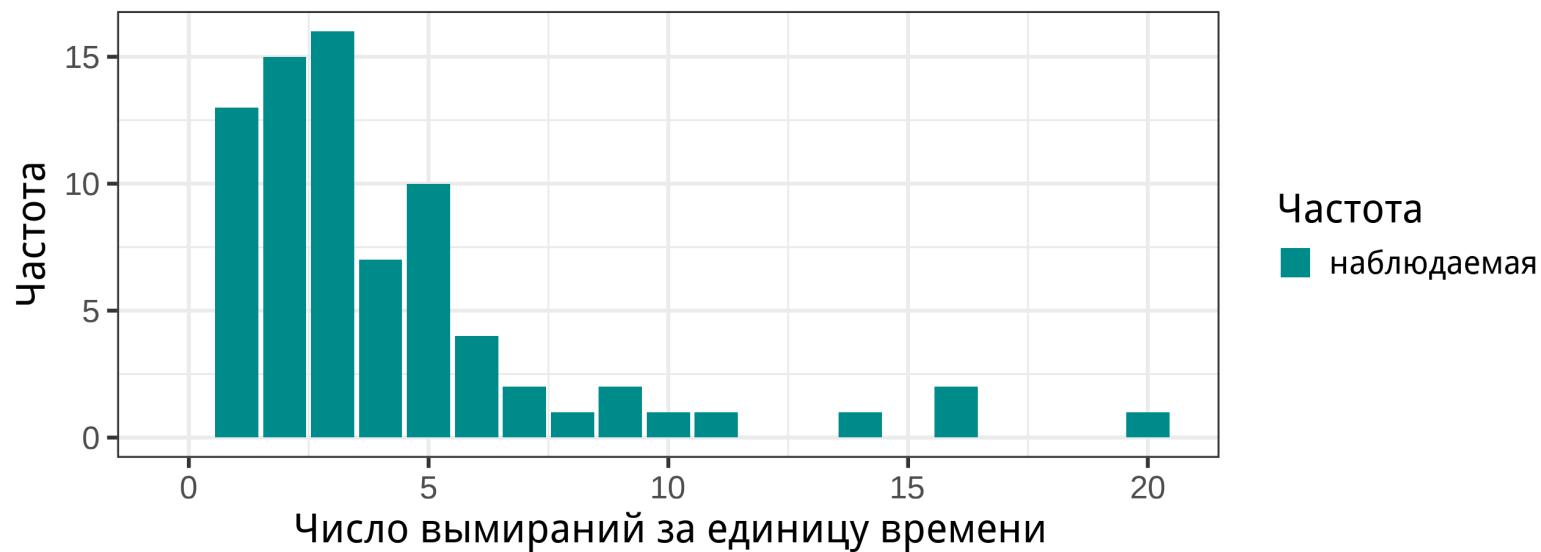


Если вымирания случайно распределены — то они подчиняются распределению Пуассона.

Если нет, то

- массовые вымирания
- равномерное распределение

Наблюдаемые частоты



Проверяем соответствие распределению Пуассона при помощи χ^2

$H_0 : \chi^2 = 0$ число вымираний подчиняется распределению Пуассона

$H_A : \chi^2 > 0$ число вымираний не соответствует распределению Пуассона

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Проверяем соответствие распределению Пуассона при помощи χ^2

$H_0 : \chi^2 = 0$ число вымираний подчиняется распределению Пуассона

$H_A : \chi^2 > 0$ число вымираний не соответствует распределению Пуассона

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Чтобы посчитать ожидаемые нужно знать...

Проверяем соответствие распределению Пуассона при помощи χ^2

$H_0 : \chi^2 = 0$ число вымираний подчиняется распределению Пуассона

$H_A : \chi^2 > 0$ число вымираний не соответствует распределению Пуассона

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Чтобы посчитать ожидаемые нужно знать...

$$\text{среднее число вымираний } \bar{X} = \frac{(0 \times 0) + (13 \times 1) + (15 \times 2) + \dots}{76} = 4.211$$

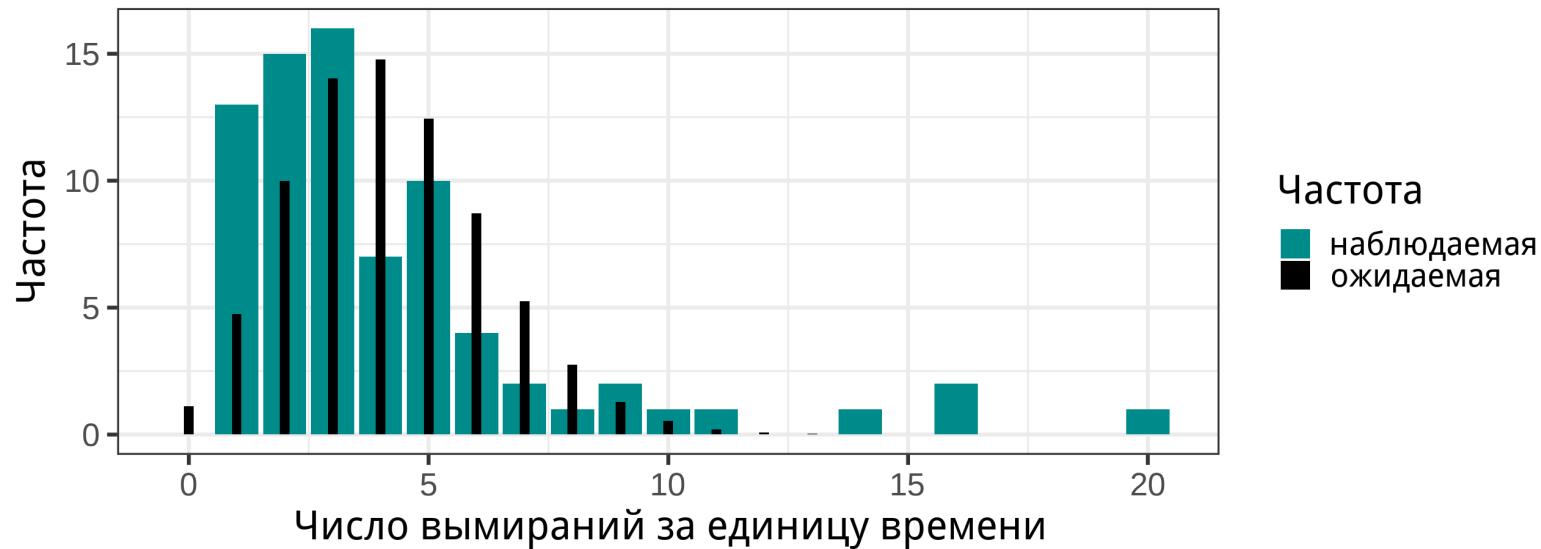
Поэтому ожидаемая частота

$$P(X \text{ вымираний}) = \frac{e^{-4.211} 4.211^X}{X!}$$

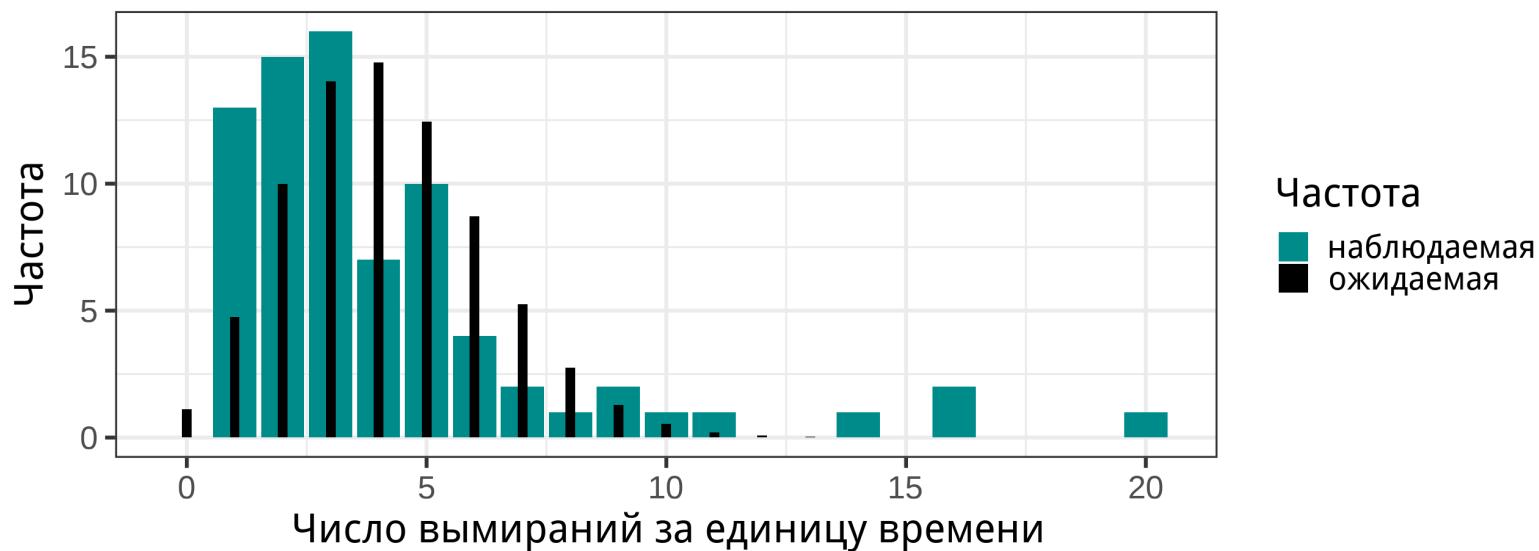
Наблюдаемые и ожидаемые частоты

Число вымираний (X)	Наблюдаемая частота	Ожидаемая доля	Ожидаемая частота
0	0	0.015	1.128
1	13	0.062	4.748
2	15	0.132	9.997
3	16	0.185	14.030
4	7	0.194	14.769
5	10	0.164	12.437
6	4	0.115	8.728
7	2	0.069	5.250
8	1	0.036	2.763
9	2	0.017	1.293
10	1	0.007	0.544
11	1	0.003	0.208
12	0	0.001	0.073
13	0	0.000	0.024
14	1	0.000	0.007

Проверяем условия применимости χ^2



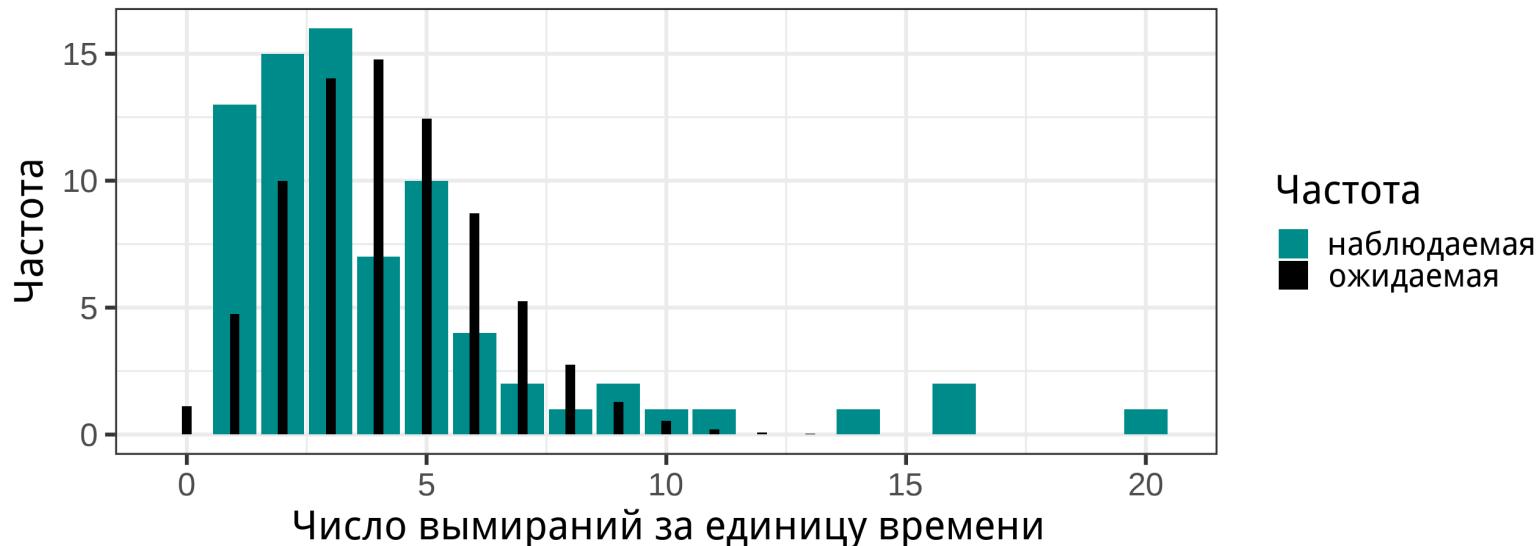
Проверяем условия применимости χ^2



Данные не соответствуют условиям применимости критерия χ^2 :

- одна ожидаемая частота < 1 ,
- более 20% ожидаемых частот < 5

Проверяем условия применимости χ^2



Данные не соответствуют условиям применимости критерия χ^2 :

- одна ожидаемая частота < 1 ,
- более 20% ожидаемых частот < 5

Выход - объединить категории:

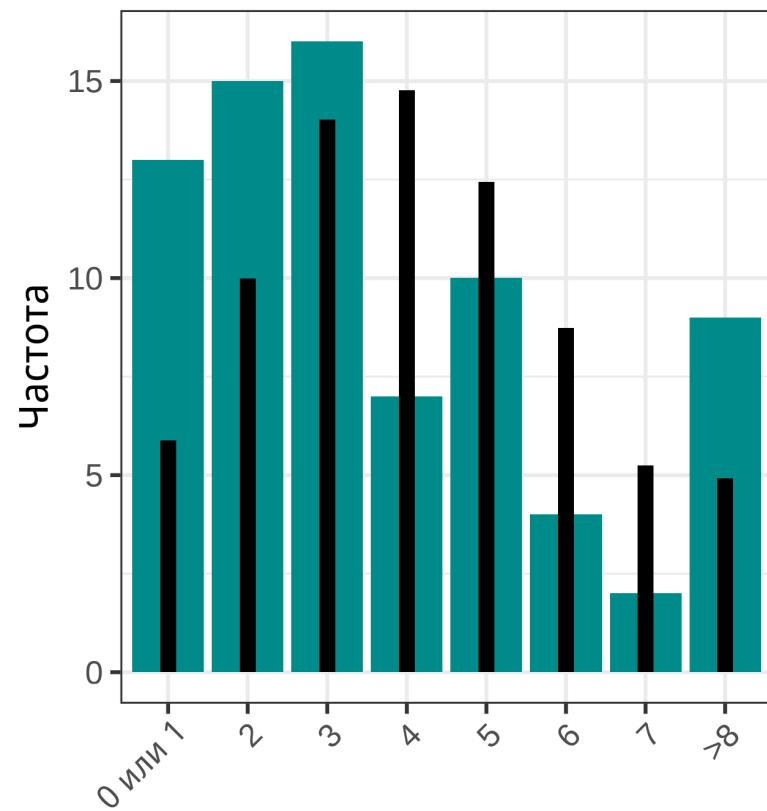
- $X = 0$ и $X = 1$
- $X \geq 8$

Данные после объединения категорий

Число вымираний (X)	Наблюдаемая частота	Ожидаемая частота
0 или 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.73
7	2	5.25
>8	9	4.92
Сумма	76	76.00

Данные после объединения категорий

Число вымираний (X)	Наблюдаемая частота	Ожидаемая частота
0 или 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.73
7	2	5.25
>8	9	4.92
Сумма	76	76.00



Частота ■ наблюдаемая ■ ожидаемая

Считаем хи-квадрат

Число вымираний (X)	Наблюдаемая частота	Ожидаемая частота	Хи-квадрат
0 или 1	13	5.88	8.637
2	15	10.00	2.504
3	16	14.03	0.277
4	7	14.77	4.086
5	10	12.44	0.477
6	4	8.73	2.561
7	2	5.25	2.012
>8	9	4.92	3.396
Сумма	76	76.00	23.950

Считаем хи-квадрат

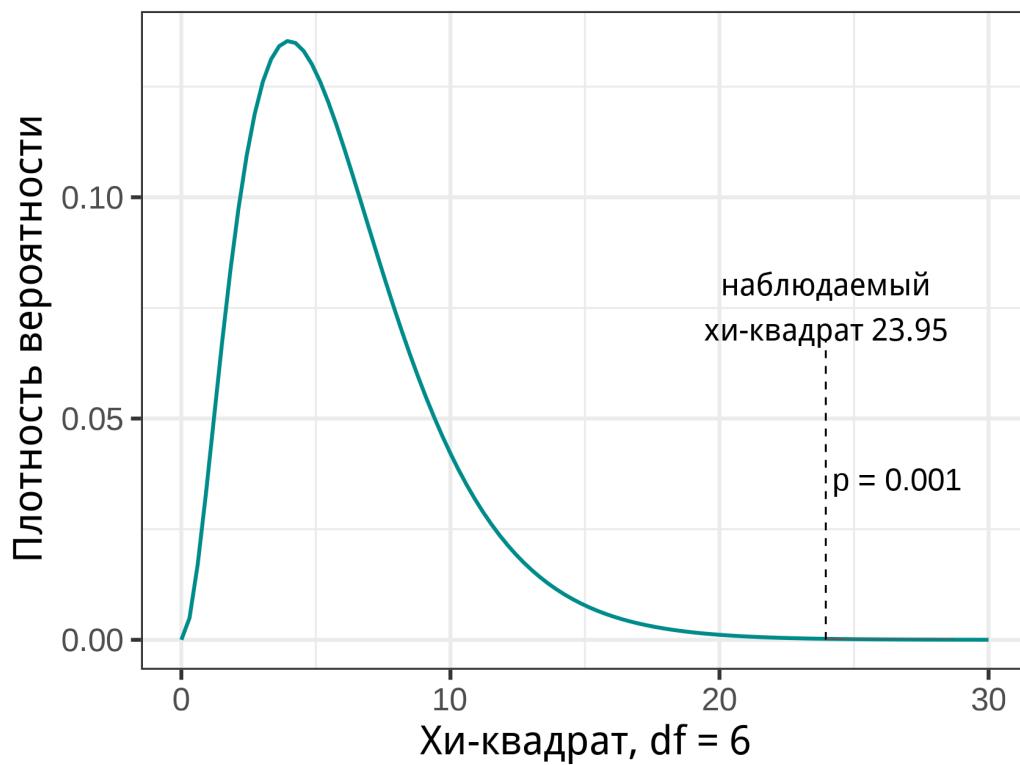
Число вымираний (X)	Наблюдаемая частота	Ожидаемая частота	Хи-квадрат
0 или 1	13	5.88	8.637
2	15	10.00	2.504
3	16	14.03	0.277
4	7	14.77	4.086
5	10	12.44	0.477
6	4	8.73	2.561
7	2	5.25	2.012
>8	9	4.92	3.396
Сумма	76	76.00	23.950
$\chi^2 = 23.95$			

χ^2 -тест

$H_0 : \chi^2 = 0$ — частоты в генеральной совокупности соответствуют распределению Пуассона

$H_A : \chi^2 > 0$ — частоты в генеральной совокупности не подчиняются распределению Пуассона.

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

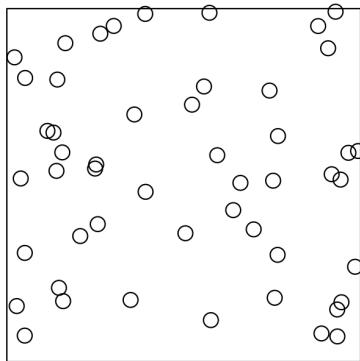


Расположение в пространстве

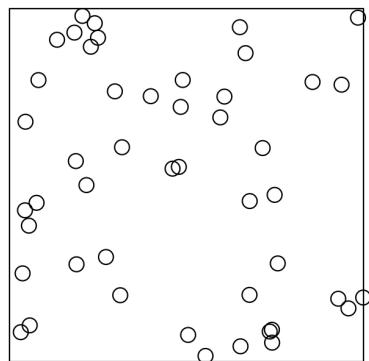
Сравнение дисперсии и среднего

Расположение в пространстве

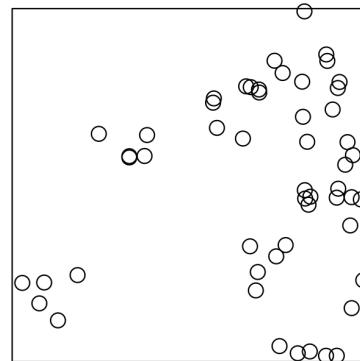
Регулярное
Dispersed



Случайное
Random

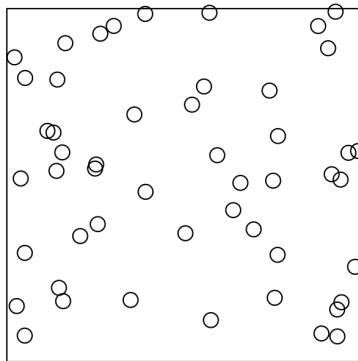


Агрегированное
Clumped

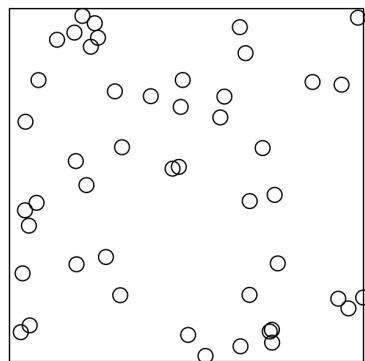


Расположение в пространстве

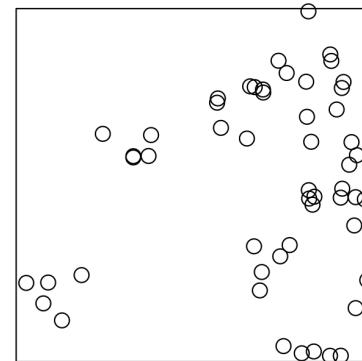
Регулярное
Dispersed



Случайное
Random



Агрегированное
Clumped

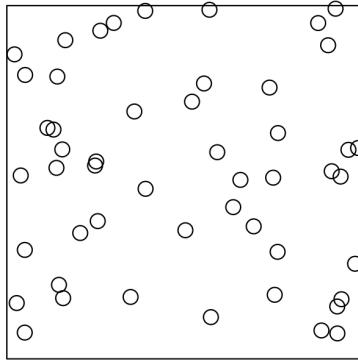


Liam Quinn from Canada, CC BY-SA 2.0

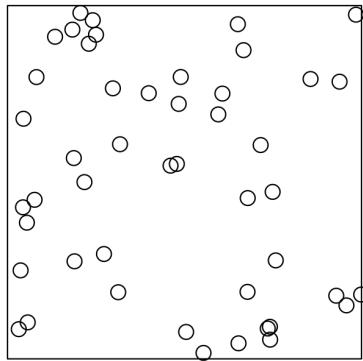
<https://creativecommons.org/licenses/by-sa/2.0/>, via Wikimedia

Расположение в пространстве

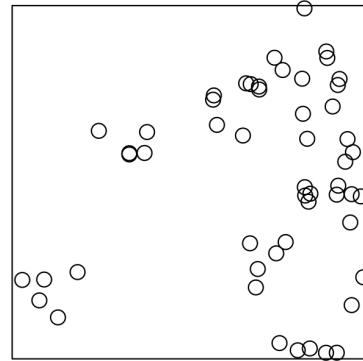
Регулярное
Dispersed



Случайное
Random



Агрегированное
Clumped



Liam Quinn from Canada, CC BY-SA 2.0

<https://creativecommons.org/licenses/by/3.0/>, via Wikimedia

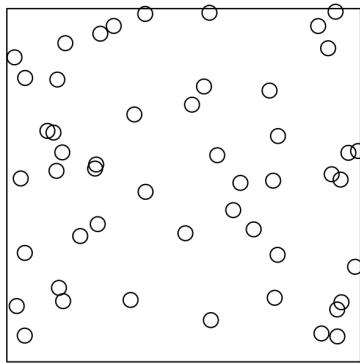
Commons



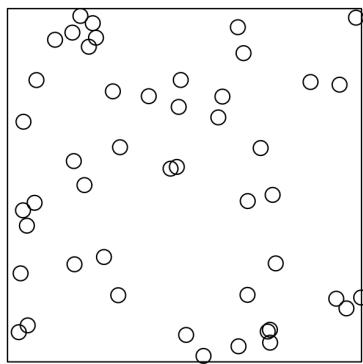
Dwight Burdette, CC BY 3.0

Расположение в пространстве

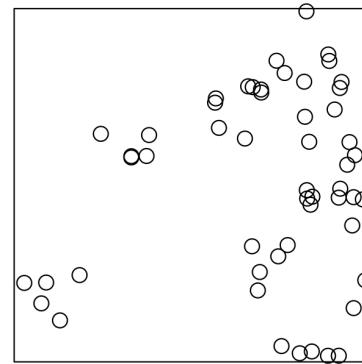
Регулярное
Dispersed



Случайное
Random



Агрегированное
Clumped



Liam Quinn from Canada, CC BY-SA 2.0

<https://creativecommons.org/licenses/by/3.0/>, via Wikimedia

Commons



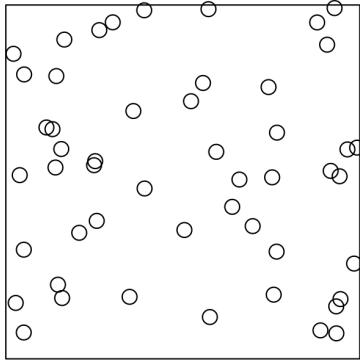
Dwight Burdette, CC BY 3.0



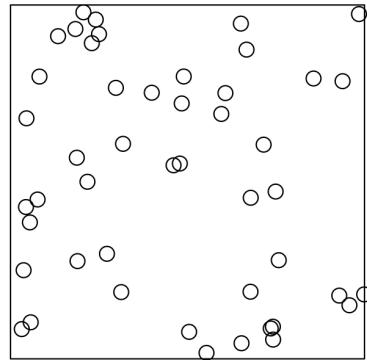
Benh Lieu Song on Flickr

Расположение в пространстве

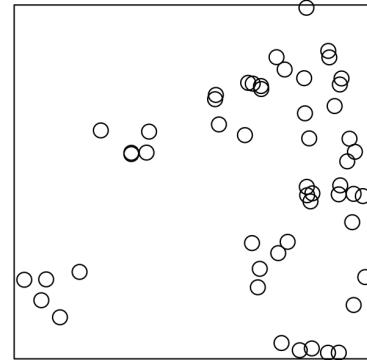
Регулярное
Dispersed



Случайное
Random



Агрегированное
Clumped



Индекс дисперсии (отношение дисперсии к среднему) — показывает, насколько расположение наблюдений в пространстве (или во времени) соответствует Пуассоновскому.

$$I_s = \frac{s^2}{\bar{x}}$$

- $I_s < 1$ — регулярное, равномерное
- $I_s = 1$ — случайное (распределение Пуассона)
- $I_s > 1$ — агрегированное

Сравнение дисперсии и среднего

$$I_s = \frac{s^2}{\bar{x}}$$

Для тестирования значимости индекса дисперсии используют хи-квадрат распределение:

$$I_s \cdot (n - 1) \sim \chi^2,$$

$$df = n - 1$$

Сравнение дисперсии и среднего

$$I_s = \frac{s^2}{\bar{x}}$$

Для тестирования значимости индекса дисперсии используют хи-квадрат распределение:

$$I_s \cdot (n - 1) \sim \chi^2, \\ df = n - 1$$

$H_0 : I_s = 1$ — распределение в пространстве случайно
(дисперсия равна среднему)

$H_0 : I_s \neq 1$ — распределение в пространстве неслучайно
(дисперсия не равна среднему)

Сравнение дисперсии и среднего

$$I_s = \frac{s^2}{\bar{x}}$$

Для тестирования значимости индекса дисперсии используют хи-квадрат распределение:

$$I_s \cdot (n - 1) \sim \chi^2, \\ df = n - 1$$

$H_0 : I_s = 1$ — распределение в пространстве случайно
(дисперсия равна среднему)

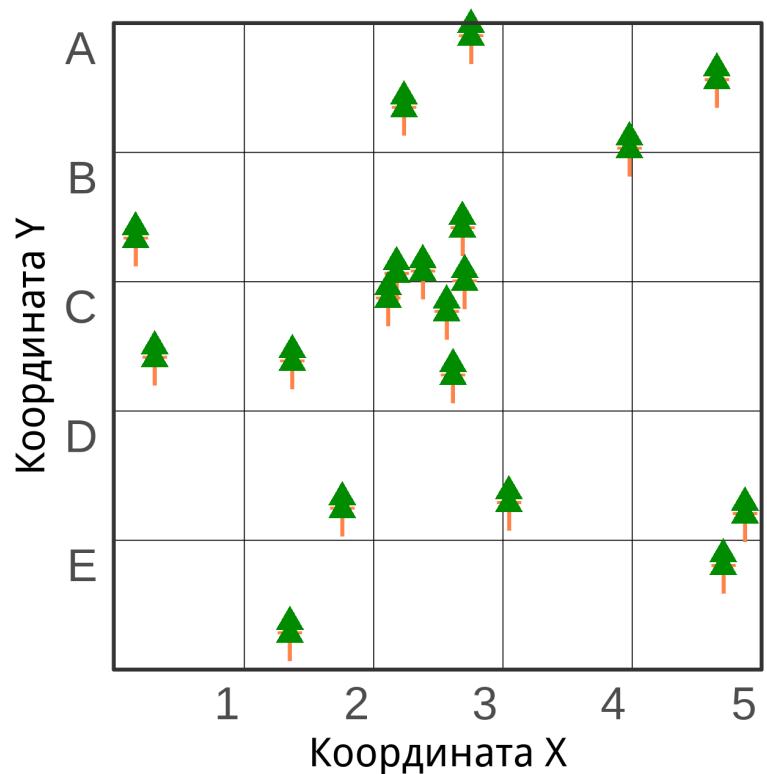
$H_0 : I_s \neq 1$ — распределение в пространстве неслучайно
(дисперсия не равна среднему)

Двусторонний тест (!)

Определяем пространственное расположение

Лес разделен на 25 квадратов. Генератор случайных чисел предлагает посчитать деревья в квадратах D-1, C-2, A-4, E-3, B-1, D-4, E-5, C-3

Посчитайте индекс дисперсии, чтобы определить, как располагаются деревья в лесу.



Определяем пространственное расположение

Лес разделен на 25 квадратов. Генератор случайных чисел предлагает посчитать деревья в квадратах D-1, C-2, A-4, E-3, B-1, D-4, E-5, C-3

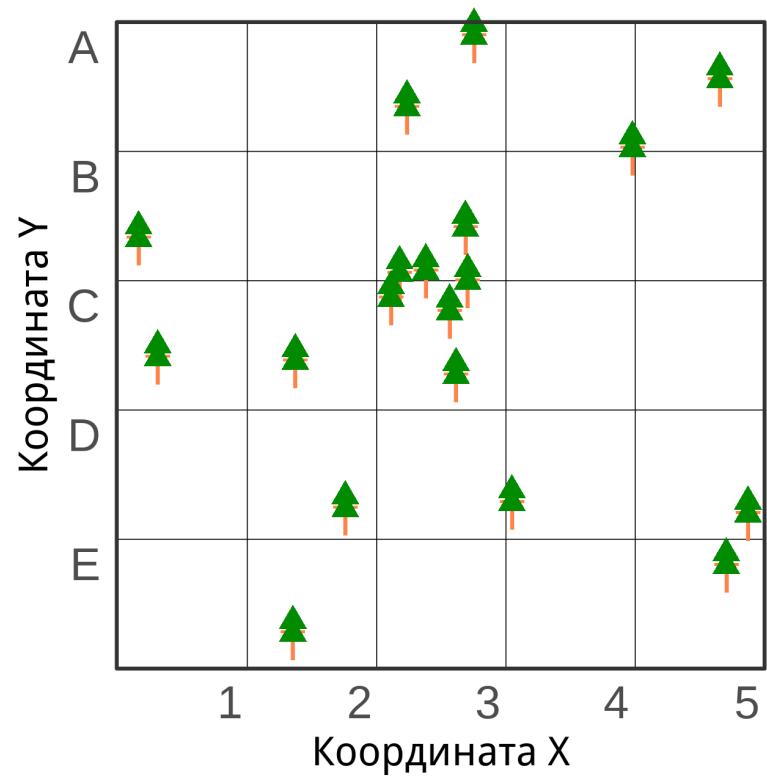
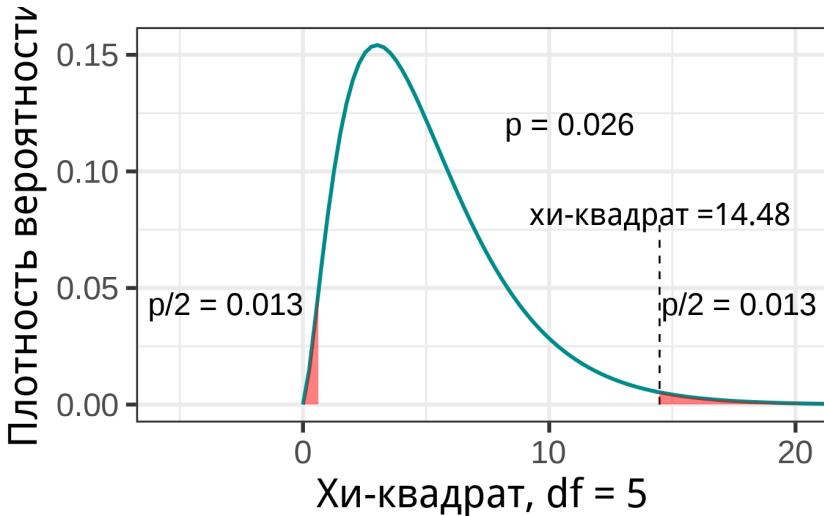
Посчитайте индекс дисперсии, чтобы определить, как располагаются деревья в лесу.

Количество деревьев:
0, 1, 0, 0, 1, 1, 2, 6

Среднее 1.4, дисперсия 4

$$\text{Индекс структурности } I_s = \frac{4}{1.4} = 2.9$$

$$\chi^2 = 2.9 \cdot (6 - 1) = 14.48, df = 6 - 1 = 5$$



Summary

Summary

- Вероятностные модели (probabilistic models) описывают вероятности событий.
- Тесты адекватности модели (goodness-of-fit tests) — описывают, насколько наблюдаемые значения соответствуют теоретическому распределению (вероятностной модели).
- Пропорциональная модель (proportional model) — вероятностная модель, в которой вероятность события пропорциональна числу возможностей его возникновения.

Summary

- χ^2 -тест оценивает соответствие наблюдаемого частотного распределения теоретическому (нулевой вероятностной модели).
- Для применения требуется, чтобы
 - наблюдения независимы друг от друга
 - не было ожидаемых частот < 1
 - ≤ 20 ожидаемых частот < 5

Summary

- Распределение Пуассона описывает вероятность определенного числа “успехов” за единицу времени или на единицу пространства, если (1) испытания независимы и (2) вероятность успеха постоянна.
- Дисперсия Пуассоновской случайной величины равна среднему.
- Индекс дисперсии (отношение дисперсии к среднему) показывает, насколько расположение наблюдений в пространстве или во времени соответствует Пуассоновскому и позволяет отличить регулярное, случайное и агрегированное распределения

ЧТО ПОЧИТАТЬ

Agresti, A., Franklin, C. A., & Klingenberg, B. (2017). Statistics: The art and science of learning from data (Fourth edition). Pearson. — глава **6.3 Probabilities When Each Observation Has Two Possible Outcomes**

Whitlock, M., & Schluter, D. (2015). The analysis of biological data (Second edition). Roberts and Company Publishers.