**Комментарии при проверке отчётов.**

**Как проверять подтверждается ли гипотеза о случайном распределении генов по цепочкам (14 дек. 2019)**

**Гипотеза о случайности распределения в данном случае значит**, что (1) каждый ген с вероятностью 0.5 либо лежит на «+» цепочке, либо на «–» цепочке; (2) выбор цепочки *независим* для генов.

(2) требует уточнения. Пример *зависимости*: если 127-й ген лежит на «+» цепочке, то вероятность того, что 128-й ген лежит на «+» цепочке больше чем 0.5. Такая зависимость может быть в геноме, в котором большое число генов объединены в *опероны*, состоящие из более, чем одного гена. По определению оперона, гены в опероне идут подряд и закодированы на одной и той же цепочке.

Про опероны мы обсуждали на занятии в пятницу 13го декабря. Спросите кого-нибудь, кто слушал и понял.

Определение независимости: вероятность того, что 1й ген на цепочке «+», 27й ген на «+», 127й ген на «-» равна произведению вероятностей, т.е. (0.5)3 . И так для любых наборов (ген, цепочка).

ПРИМЕР данных.

4106 генов. Из них 2024 гена на «+» цепочке.

Ответ 0.19, получаемый формальным применением функции =БИНОМ.РАСПР(2024,4106, 0.5, ИСТИНА) - НЕПРАВИЛЬНЫЙ!

Кстати, **а вопрос то какой?**

**Вопрос:** наблюдаемое распределение генов по цепям (2024 «+», 2082 «-») противоречит случайному распределению генов или нет?

При случайном распределении *ожидаемое* распределение по цепям (2053 «+», 2053 «-»). Значит, мы наблюдаем отклонение от ожидаемого 2053 – 2024 = 29 на «+» и 2082 – 2053 = 29 на «-».

**Уточнение вопроса:** Какая вероятность **p** наблюдать отклонение на 29 или более от ожидаемого при случайном распределении по цепочкам?

**Ответ 1:** симуляция случайного распределения. Тысячу раз разыграем случайный выбор цепочки для 4106 генов. Для каждого розыгрыша вычислим отклонение от ожидаемого. Подсчитаем сколько раз из 1000 отклонение было ≥29.
 Предположим 2 раза. Тогда искомая вероятность **p** = 0.002. Если так, наблюдаемое распределение генов не соответствует гипотезе о случайности распределения. Надо искать почему.

 Предположим отклонение ≥29 наблюдалось 374 раза из 1000. Тогда искомая вероятность равна 0,374 ≈ 0.4 Наблюдаемое отклонение соответствует гипотезе о случайном распределении генов.
 Получение вероятности **p** таким способом средствами Excel см. в задании 4.2.2. в практикуме 14.

**Ответ 2:** найдем **p** с помощью функции =БИНОМ.РАСПР().

Объясняю, что делает эта функция. Вот какие вероятности вычисляются по формулам.

p0 – вероятность того, что на «+» нет ни одного гена ( =БИНОМ.РАСПР(**0**, 4106, 0.5, ЛОЖЬ) )
p1 – вероятность того, что на «+» ровно 1 ген ( =БИНОМ.РАСПР(1, 4106, 0.5, ЛОЖЬ) )
p2 – вероятность того, что на «+» ровно 2 гена ( =БИНОМ.РАСПР(2, 4106, 0.5, ЛОЖЬ) )
……………………………………………………………………………………………………………………………………………………..
p2024 – вероятность того, что на «+» ровно 2024 генов ( =БИНОМ.РАСПР(2024, 4106, 0.5, ЛОЖЬ) )
……………………………………………………………………………………………………………………………………………………..
p2053 – вероятность того, что на «+» ровно 2053 гена ( =БИНОМ.РАСПР(2053, 4106, 0.5, ЛОЖЬ) )
……………………………………………………………………………………………………………………………………………………..
p2082 – вероятность того, что на «+» ровно 2082 генов ( =БИНОМ.РАСПР(2082, 4106, 0.5, ЛОЖЬ) )

……………………………………………………………………………………………………………………………………………………..
p4106 – вероятность того, что на «+» ровно 4106 генов (т.е все!) ( =БИНОМ.РАСПР(4106, 4106, 0.5, ЛОЖЬ) )

Что же вычисляет функция =БИНОМ.РАСПР(2024, 4106, 0.5, **ИСТИНА**)?
а вот что: **p** = p0 + p1 + p2 + ……. + p2023 + p2024 и это значение приблизительно равно 0.19

Ожидаю, что дочитав до сюда, вы сообразите какой правильный ответ при том, что отклонение на 29 от ожидаемого получается и при 2082 генах на «+» цепи.

И сообразите почему, если на «+» цепи обнаружилось 2082 гена, то функция
 =БИНОМ.РАСПР(2082, 4106, 0.5, **ИСТИНА**)
даст совершенно неправильный ответ.

В Excel есть функция =БИНОМ.РАСП.ДИАП(4106, 0.5, ….., …… ) в которой указывается диапозон суммирования вероятностей **p\_i**
Например, =БИНОМ.РАСП.ДИАП(4106, 0.5, 0, 2024) равно p0 + p1 + ……. + p2024
 =БИНОМ.РАСП.ДИАП(4106, 0.5, 2082, 4106) равно p2082 + p2083 +……. + p4106

Какой вывод можно сделать, вычислив **p**

Если **p < 0.05** то можно утверждать, что гипотеза о случайности распределения генов по цепям не верна, с уровнем значимости 5% Это самый слабый уровень значимости, принятый в биологии. Следовательно, есть какая-то причина, известная или неизвестная, не случайности распределения. Можно высказывать гипотезы!

Если **p < 0.001** то можно утверждать, что гипотеза о случайности распределения генов по цепям не верна, с высоким уровнем значимости.

Если **p > 0.05** то наблюдаемое распределение генов по цепям согласуется с гипотезой о случайном распределении.

**Что делать если вы уже сообщили о готовности отчёта (чтобы не опоздать), но хотите что-то исправить (11 дек. 2019)**

Исправьте текст отчёта.

Запишитесь в очередь и укажите что перепроверить.

Дата сдачи останется прежней.

**Оформление ссылок на источники в списке. (11 дек. 2019)**

**Как сослаться на источник хромосомной таблицы.**

-------- пример 1

(4) NCBI Genomes, [страница генома Chlamydophila pneumoniae CWL029](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/008/745/GCA_000008745.1_ASM874v1/)

-------- комментарий
В Word гиперссылка делается так. Выделяете текст, правая кнопка мыши, гиперссылка и в окошко вставляете адрес гиперссылки. В .pdf гиперссылка сохранится.

**Как сослаться на информацию из базы данных.**

--------- мне попался такой пример 2

(1) база данных KEGG, страница GENOME: [Chlamydophila pneumoniae CWL029](https://www.genome.jp/dbget-bin/www_bget?gn:cpn)

-------- комментарий

Такая ссылка приемлема.

Однако ссылка на публикацию более приоритетна, чем ссылка на БД (базу данных). Исключение – ссылка на данные – материалы – как в примере 1.

**Как сослаться на Википедию.**

 --------- мне попался такой пример 3

(3) Википедия, русская, [страница Chlamydophila pneumoniae](https://ru.wikipedia.org/wiki/Chlamydophila_pneumoniae)

-------- комментарий

Такая ссылка приемлема.

Однако ссылка на публикацию более приоритетна, чем ссылка на Википедию. Исключение – ссылка в Википедии или еще где-нибудь на рисунок, фотографию и т.п. имеет высокий (но не высший) приоритет.

**Нужно ли изучить содержание статьи если вы ссылаетесь на неё?**

Нет, не обязательно.

Ссылаются на каое-то

**Почему публикации в научных журналах и книгах имеют высший приоритет по сравнению с Википедией и другими сайтами?**

 Потому, что публикации в научных журналах проходят независимую научную экспертизу. Обычно каждая рукопись, поданная в журнал, посылается двум рецензентам из числа специалистов в данной области. Редколлегия контролирует, что у рецензентов нет заинтересованности или конфликта с авторами (нет общих публикаций, не работают в одном и том же месте и т.п.)

**Зачем учёные рецензируют статьи? (можно не читать, этот раздел – крик души рецензента) )**

- им платят за это? **НЕТ**

- их рукописи принимают к печати без рецензий? **НЕТ**

- им публично объявляют благодарность за хорошую рецензию? **Почти ДА**, но своеобразно. В журнале запоминают их фамилию и чаще посылают рукописи на рецензию.

- они считают, что добротное рецензирование поднимает уровень журнала и науки в целом; т.е. своеобразное волонтерство? **ДА**

- трудно ли написать рецензию? **ДА,** трудно. На рецензию дается две – три недели. Несколько рабочих дней у меня обычно уходит на изучение других публикаций по теме, рукописи, обдумывание и написание рецензии.

ААл