Биостатистика

В ноябре 1948 года состоялись очередные выборы президента США. 2 кандидата конкурировали между собой: Томас Дьюи от партии республиканцев и Гарри Труман от партии демократов. За 2 дня до выборов газета Chicago Daily Tribune провела телефонный социологический опрос и опубликовала его результаты. Из них следовало, что Дьюи побеждает Трумана с разгромным счетом. Тем неожиданнее было то, что победу одержал Гарри Труман.

На фото: Гарри Труман, победитель выборов 1948 года

Как же так получилось, что при в целом правильно поставленном опросе, его результаты оказались абсолютно неверны? Этот вопрос совсем не праздный. В любом эксперименте, в том числе и в биологии, исследователи выполняют статистический анализ своих результатов. Неправильный подход к их анализу в биологии и медицине может привести к куда более значительным последствиям, чем просто неверному предсказанию исхода выборов. В данном курсе мы изучим основы прикладной статистики, разберем типичные ошибки и заблуждения при анализе данных, и попробуем сами проанализировать клинические данные.

Введение в матстатистику

Для начала, дадим несколько определений. Представим себе, что мы хотим измерить что-либо у всех людей, пусть это будет рост, и узнать, чему равен средний рост всех людей на Земле. Люди в данном примере являются **объектами** исследования. Обратим внимание, что у каждого человека есть какой-то рост, но мы не знаем, чему он равен, пока не измерим его. Мы будем называть **случайными величинами** те величины наших *объектов*, значений которых мы не знаем до их измерения (это не совсем точное определение, но оно вполне годится для нашей дальнейшей работы). Для нашего эксперимента мы бы хотели измерить рост всех людей на свете, однако мы отдаем себе отчет в том, что это просто физически невозможно – слишком много потребуется времени. Поэтому, мы выберем случайно несколько людей, и измерим их рост, а потом скажем, что он примерно равен среднему росту всех людей. *Все объекты* измерения – в нашем случае все люди Земли – называются **генеральной совокупностью**. *Выбранная* нами случайно *часть* этих *объектов* (группа людей) – называется **выборкой**.

Случайно выбираем несколько людей из всех людей на свете

Очень важно выбирать людей случайным образом, поскольку если прийти в детский садик и измерить там рост всех, то можно получить средний рост всех людей в этом садике, однако он будет далек от среднего

роста всех людей, потому что в мире есть не только дети. Если же мы будем выбирать людей произвольно – в идеале, просто занумеруем всех людей на свете от 1 до 7 000 000 000, и будем случайно генерировать число, например, на компьютере, то также как и среди всех людей, среди выбранных нами будут попадаться и старики, и младенцы, и люди среднего возраста; и поэтому, измерив их рост, мы сможем примерно узнать, чему будет равен рост всех людей на Земле. Если же взять выборку для исследования не случайно, то можно получить совершенно неожиданные результаты.

Именно это и произошло в истории с выборами США 1948 года. Газета CDT совершенно честно выбирала людей случайно, но только из числа тех, у кого были телефоны. Сейчас это кажется нормальным решением – у всех же есть телефоны! Но в 1948 году примитивный по сегодняшним меркам домашний телефон себе могли позволить только богатые граждане, среди которых, действительно, очень многие поддерживают республиканскую партию. Однако подавляющее большинство населения США жило в то время намного беднее, телефонов не имело, и поддерживало в основном демократов. Именно это и обусловило победу Гарри Трумана, кандидата от демократов, на выборах 1948 года. Как видим, хоть газета и старалась сделать выборку случайно, это оказалось не так просто.

Мораль: По данным интернет-опроса 100% людей пользуется интернетом.

Средние значения

Итак, мы выбрали случайно группу людей, у которых будем измерять среднее значение роста. Но можно делать это по-разному. Мы рассмотрим два способа: среднее арифметическое и медиану.

Пусть у нас есть n человек, их роста равны: k_1, k_2, k_3, …, k_n.

Средним арифметическим значением случайной величины в выборке называется число = $\frac{k_1+k_2+\cdots+k_n}{n}$. Это наиболее часто используемая из всех средних величина. У неё есть один недостаток – если вы случайно измерили очень редкий объект, который сильно отличается от всех других, то он сильно повлияет на значение среднего арифметического. Чтобы решить эту проблему, используют *медиану*.

Медианой называется значение, такое, что если мы выстроим все объекты в порядке возрастания случайной величины (всех людей в порядке их роста), то значение среднего объекта и будет медианой. Например, у вас есть 5 людей, их роста равны: 160, 170, 175, 180, 210 сантиметров. Среднее арифметическое равно 179 см. Медиана равна росту третьего человека, т.е. 175 см. Еще пример: роста 4 человек, 160, 170, 180, 200 сантиметров. Т.к. среднего человека тут нет, медиана будет равна среднему арифметическому двух средних людей: (170+180)/2 = 175 см.

Случайные величины

Теперь разберемся со случайными величинами. Мы уже дали определение случайной величины, как величины у объекта, чьего значения мы не знаем до того, как измерим её. Типичным примером СВ является число, выпадающее на игральном кубике. Можно записать все возможные результаты выпадения числа на кубике – это 1, 2, 3, 4, 5 и 6. Если кубик – честный, то любое число выпадает одинаково часто. Для измерения частоты выпадения того или иного значения случайной величины в математике используются *вероятности*. Математики договорились, что **сумма вероятностей всех возможных исходов** эксперимента (в нашем случае эксперимент – это один бросок кубика) всегда **равна 1** – это означает, что что-нибудь из перечисленного произойдет. Если частоты выпадения всех чисел на кубике одинаковы, то их вероятности равны, а значит каждая из них равна 1/6. Можно записать все возможные результаты эксперимента по броску кубика в виде таблицы:

Результат броска кубика, **x**	Вероятность результата, **P(x)**
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Такая таблица называется **таблицей распределения вероятностей**. Случайные величины, для которых можно написать ТРП, называются **дискретными**. Кратко упомянем, что не все случайные величины можно записать как таблицу. Те СВ, которые содержат бесконечно много возможных исходов, причем настолько бесконечно много, что их нельзя никак сосчитать, нельзя и записать таблицей. Их называют **непрерывными**, и для них используют формулы, описывающие вероятность того, что значение СВ будет *примерно* равно какому-то значению. Примером такого закона для случайной величины является так называемое **нормальное распределение**, часто встречающееся в биологии:

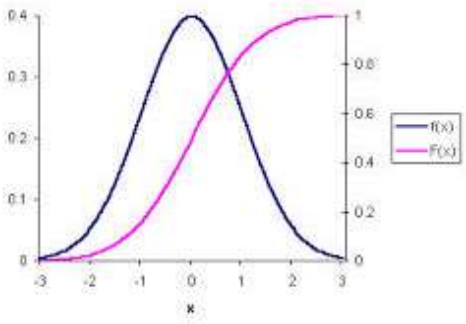

График *нормального распределения*.

Синий график – вероятность того, что у случайно взятого объекта величина x будет равна какому-то значению, или *примерно* ему (ось Y подписана слева). Чем он выше в какой-то точке, тем больше шансов случайно выбрать объект, со значением x, равному значению в этой точке.

Розовый график – вероятность того, что значение x у случайно взятого объекта *меньше или равно*, чем заданное значение x (ось Y справа).

Анализ гипотез

В естественных науках действует *принцип фальсифицируемости*. Он гласит, что научной является не та теория, которую мы можем *проверить*, а та, которую мы можем **опровергнуть**, или что то же самое – фальсифицировать (отсюда и название). Поэтому ученые во всем мире следуют следующему алгоритму: они принимают какую-либо научную теорию за верную, и работают с ней, пока достаточно многие исследователи не начнут замечать, что теория не всегда работает так, как надо. Обнаруживается это потому, что при постановке своих экспериментов, исследователи замечают, что наблюдаемые ими результаты эксперимента не совпадают с тем, что им должна предсказывать теория.

Для анализа гипотез существует четкий алгоритм, который в том или ином виде действует в любом исследовании:

0. Формулируем H0 <–> Ha
1. Выбираем α (часто – 0.05, или 5%)
2. Ставим эксперимент
3. Рассчитаем вероятность того, что то, что мы видим, и даже более необычное – могло получиться случайно (при верной H0) – p_v
4. Сравниваем:
    a. Если pv > α – не отвергаем H0 , принимаем H0;
    b. Если pv <= α – отвергаем H0, принимаем Ha.

H0 и Ha – это так называемые **нулевая гипотеза** и **альтернативная гипотеза**. Нулевая гипотеза – это та гипотеза, которая принята сейчас, и которую мы будем опровергать; альтернативная – та, которую мы предлагаем ей взамен. α (альфа) – **пороговое значение**, к его смыслу мы вернемся немного позже.

p_v (читается «**пи-вэлью**») – это вероятность того, что при верной нулевой гипотезе мы увидим настолько же неожиданные, сильные

результаты, как видим сейчас, или еще более сильные. Альфа служит для сравнения с ним – если пи-вэлью больше порогового значения, то событие достаточно ординарное, а значит нулевая гипотеза может спать спокойно. Если же пи-вэлью равно или меньше порогового значения, то это значит, что увиденное нами – очень маловероятно, а значит нулевая гипотеза уже не справляется с объяснением окружающего мира, и требуется другая, альтернативная ей.

План поиска новых лекарств

На первой стадий ученые используют клетки, которые моделируют изучаемое заболевание, и ищут вещество, которое воздействует на них желаемым образом. На второй стадии выбранное на первой стадии вещество изучается на животных, с целью показать два свойства: его биобезопасность – т.е. доказать, что найденное вещество не убивает и не травмирует животных; и его эффективность – т.е. доказать, что новое лекарство работает, и делает это лучше имеющихся аналогов. Если лекарство проходит испытания второй стадии, наступает третья стадия – ограниченные клинические испытания – исследования, в которых больные люди получают данное лекарство. В ходе этих исследований также проверяется биобезопасность и эффективность вещества, которое планируется использовать для лечения.

В ходе клинических испытаний необходимо решить как минимум три проблемы. Первое: мы не знаем, как много людей выздоравливает без лекарства – возможно лекарство бесполезно, и люди лечатся сами? Чтобы решить эту проблему, нужно поделить больных на две группы, одной давать лекарство, а другой – нет; и сравнить между собой результаты двух групп. Здесь появляется вторая проблема – для некоторых заболеваний известен эффект **плацебо** – люди, которые считают, что их лечат, выздоравливают лучше, чем те, кто так не считает. Чтобы он не вносил своего влияния на результат, необходимо второй группе пациентов – тем, кто не получает лекарство, давать «пустышку» - таблетку или сироп, полностью такой же, как и настоящее лекарство, но не содержащий само лекарство. Такие испытания называются «**слепыми**» - пациенты не знают, лечат их или нет на самом деле. Третья проблема связана с человеческим фактором – компании, разрабатывающие лекарства, заинтересованы в хорошем исходе эксперимента и могут пытаться подкупить врачей. Чтобы избежать этого, проводят «**двойные слепые**» испытания, в которых ни врач, ни пациент не

знает, что получает пациент; а только какой-то третий источник знает это (например, компьютерная программа, которая случайно выбирает, кому какую пробирку дать). При таком подходе решаются все три проблемы, которые возникают при испытаниях лекарств.

Основы молекулярной биологии

Клетка – это основная единица всего живого. Все живые организмы (кроме вирусов, чье отношение к живому оспаривается) состоят из клеток. Живые организм делятся на **прокариотов** – одноклеточные живые организмы, у которых нет ядра («про» - перед, «карион» - ядро), и **эукариотов** – одноклеточные и многоклеточные живые организмы, к летках которых есть ядро («эу» - настоящий, «карион» - ядро).

В ядре у эукариотов, и в цитоплазме у прокариотов находится **ДНК** – молекула, отвечающая за хранение и передачу потомкам наследственной (генетической) информации. ДНК расшифровывается как **дезоксирибонуклеиновая кислота**. Она представляет собой химическую молекулу, состоящую из двух «нитей», закрученных друг вокруг друга. На эти «нити» прикреплены один за другим отдельные элементы ДНК, называемые **нуклеотидами**. Всего существует 4 нуклеотида в ДНК – A, C, G, T. Каждый нуклеотид имеет определенную химическую формулу, которая отличается у всех четырех.

Последовательность ДНК может быть скопирована, например, при делении клетки на две – этот процесс копирования ДНК называется **репликацией**. С ДНК также может быть считана **мРНК – рибонуклеиновая кислота**. Этот процесс называет **транскрипцией**. РНК также является нитью, но всего одной нитью и состоит из 4 нуклеотидов, но немного других – A, C, G, U. С РНК может быть считан **белок**. Белок – это также одна нить, но состоит он из **аминокислот**, их имеется 20 разных. Каждая аминокислота кодируется тремя нуклеотидами, процесс перевода РНК в белок называется **трансляцией**.

Участки ДНК, с которых идет транскрипция мРНК (из которой потом получится белок), называются **генами**. У прокариот гены целиком превращаются в мРНК, а та, в свою очередь, целиком становится белком. У

эукариот с ДНК будет прочитана только **пре-мРНК**. Также как и мРНК, пре-мРНК – это нить с последовательностью нуклеотидов на ней. Ее отличие в том, что она состоит из элементов двух типов – **интронов** и **экзонов**. Интроны вырезаются из пре-мРНК, а оставшиеся экзоны – склеиваются между собой, и только после этого у эукариот образуется мРНК. Таким образом, у эукариотов гены превращаются в белок не целиком; только экзоны из них становятся в конце белком.