СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

p^*	выборочная вероятность (частость) случайного события
φ	величина центрального угла при преобразовании Фишера
$M[X]$	математическое ожидание случайной величины X
$\tilde{m}_x$	выборочная средняя, оценка математического ожидания случайной величины X
$\sigma[X]$	среднее квадратическое отклонение случайной величины X
ν_k	начальный момент случайной величины k-го порядка
μ_k	центральный момент случайной величины k-го порядка
A_s	коэффициент асимметрии
E_k	эксцесс
$\Phi(x)$	функция Лапласа
$\Phi^{-1}(x)$	обратная функция Лапласа
$\varphi(\bar{X}_1\bar{X}_2)$	функция расстояния или метрика
D_E	расстояние по Евклиду
D_M^2	расстояние по Махалонобису
D_χ^2	расстояние по Сангви
D_{St}^2	расстояние Стейнберга
D_{OH}	расстояние Оливье — Хауэллса
$\overline{p}$	выборочная вероятность случайного события
ν_k	начальный момент случайной величины k-го порядка

μ_k	центральный момент случайной величины k-го порядка
τ	коэффициент ранговой корреляции Кендалла
$\sigma_{\bar{p}_1}$	средняя квадратическая ошибка выборочной вероятности случайного события
$\overline{k'}$	средний темп прироста (снижения)
$\Delta\overline{Y}$	средний прирост
A, B, C	факторы в дисперсионном анализе
a_0, a_1	параметры линейной функции
a_i	параметры (коэффициенты) уравнения регрессии, $i = 1, n$
Ass	асимметрия
$D[X]$	дисперсия случайной величины X
D_Q	расстояние по Евклиду по величине
D_r	мера расстояния в кластерном анализе
$D_r(X)$	каноническая линейная дискриминантная функция
D_S	расстояние Спулера
D_Z	расстояние по Евклиду по форме
F	критерий Фишера
$F(x)$	функция распределения случайной величины X
H_0	нулевая гипотеза
k	число степеней свободы
k	темп роста (прироста, снижения)
$L_j(X)$	линейная классификационная функция
m_r	средняя квадратическая ошибка коэффициента корреляции
n	объём выборки

$N(m_x, \sigma_x)$ — двумерное нормальное распределение с параметрами m_x, σ_x

Q — коэффициент ассоциации (коэффициент Юла)

r — коэффициент корреляции

$R_M = R^2$ — коэффициент детерминации

S_o — среднее квадратическое отклонение наблюдавшихся значений случайной величины от рассчитанных по уравнению регрессии

V — критерий Крамера

V_x — коэффициент вариации распределения случайной величины X

$W(\lambda, k)$ — двухпараметрическое распределение Вейбулла — Гнеденко с параметрами λ, k

$X, Y, Z,$ — учитываемые признаки в дисперсионном анализе

$\hat{y}$ — прогнозируемое значение функции

β — доверительная вероятность

ΔY — абсолютный прирост

ε — ширина доверительного интервала

η — корреляционное отношение

ρ — коэффициент ранговой корреляции Спирмена

$\sigma^2{}_0$ — дисперсия, обусловленная неучтёнными факторами

$\sigma^2{}_r$ — дисперсия, обусловленная регрессией

χ^2 — критерий *Хи*-квадрат (по Пирсону)

ДСВ — дискретная случайная величина

$Me[X]$ — медиана случайной величины X

$Mo[X]$ — мода случайной величины X

НСВ	непрерывная случайная величина
ОВЧ	относительная величина частоты
ППП	пакет прикладных программ
$P(A)$	вероятность случайного события A
ОРВИ	острая респираторная вирусная инфекция
РВ	радиоактивные вещества
C	коэффициент сопряженности признаков
СВ Y	случайная величина Y
СКО	среднее квадратическое отклонение СВ X
Φ	коэффициент контингенции
$x_1, x_2, \dots, x_n$	возможные значения случайной величины X

«Современная наука нашла как будто более надёжное средство против предубеждения в практической медицине, — это медицинская статистика, основанная на цифрах».

Н. И. Пирогов[1]

ВВЕДЕНИЕ

Математическая статистика — раздел математики, посвящённый математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов.

Методы математической статистики (мета-анализ) дают возможность объективно оценивать количественные результаты исследований, позволяют обрабатывать данные небольшого числа наблюдений, увеличивают достоверность заключений по материалам исследований. Математическая обработка результатов позволяет также в сжатой и четкой форме публиковать цифровые данные исследований вместо громоздких таблиц, занимающих много места и трудных для восприятия.

Обработка числовой информации в наши дни немыслима без применения средств вычислительной техники. Современный исследователь (врач, провизор) обязан обладать навыками компьютерной обработки данных и иметь представление о программном обеспечении, с помощью которого ее можно выполнять. Сегодня существует большое количество специализированных приложений для статистического анализа. Среди лидеров таких продуктов признана программа STATISTICA фирмы Tibco США. Помимо мощного набора процедур статистического и графического анализа, эта программа обладает весьма дружественным интерфейсом, что делает ее достаточно легкой для освоения и удобной в работе.

В последнее время особое внимание уделяется разработке вопросов доказательной медицины, основанной на применении современных информационных технологий. Приведенные в настоящем

[1] Пирогов Н. Н. Вопросы жизни. Дневник старого врача / Н. И. Пирогов. — СПб.: ВМедА, 2008. — С. 343.

издании методы статистической оценки результатов клинических испытаний соответствуют требованиям GCP (*Good Clinical Practice*). Предлагаемые подходы позволяют решать практически все задачи статистической оценки клинических экспериментальных исследований. Кроме того, методы статистического анализа используются в процессе производства лекарственных препаратов, что позволяет проектировать создание по ряду оптимальных критериев технологических процессов, контролировать качество исходных ингредиентов и многое другое.

Важную роль статистические методы выполняют и в исследованиях маркетингового характера, включая определение потребности в тех или иных медицинских изделиях, тенденции замены одних поколений лекарственных препаратов другими, оптимальную тактику их продвижения на рынок.

В связи с ограниченным объемом пособие содержит примеры и пошаговые описания решений лишь типовых задач, возникающих в медицинских исследованиях (описательная статистика, сравнение двух и более выборок, корреляционный и регрессионный, дисперсионный и дискриминантный, факторный и кластерный анализы …).

Авторы будут благодарны всем, пожелавшим изложить свои оценки, пожелания и советы по данной книге по адресу: *stat-pharm@yandex.ru*.

1. СОВРЕМЕННАЯ СТАТИСТИКА В БИОМЕДИЦИНЕ, ФАРМАЦИИ И ФАРМАЦЕВТИКЕ

Современный руководитель фармацевтического бизнеса любого ранга вынужден принимать решения в условиях постоянного прессинга, при этом, как правило, испытывая дефицит достоверной информации. Однако существуют методы, позволяющие максимально полно использовать любую доступную информацию. Речь идет о бизнес-статистике (от *лат. status* — состояние дел), науке, сочетающей учет и анализ, фиксирующей, систематизирующей и изучающей показатели наиболее типичных, массовых экономических процессов и их изменение во времени.

Благодаря статистическим методам руководитель фармацевтического бизнеса имеет возможность извлекать ценную информацию из совокупности данных и на ее основе определить возможные риски и ожидаемые результаты. Такой подход является полной противоположностью той категории топ-менеджмента, которая считает, что принятие решения в бизнесе должно быть основано исключительно на интуиции и практическом опыте, что исключает использование какой либо количественной информации. Но даже и такие руководители не могут действовать в информационном вакууме и вынуждены принимать во внимание всю имеющуюся ценную информацию.

Таким образом, статистические методы следует рассматривать как важную часть процесса принятия решений, позволяющую выработать обоснованные стратегические решения, сочетающие интуицию специалиста с тщательным анализом имеющейся информации. Примерами такой информации полезной как для фармацевтического ритейла, так и для других отраслей фармацевтического бизнеса могут быть: динамика товарооборота; количество посещений аптеки, количество заказов у дистрибьютора, показатель стоимости одного чека и др.

Практическая ценность бизнес статистики заключается, прежде всего, в том, что она позволяет получать в концентрированном виде объективную характеристику определенной сферы бизнеса. Предположим, что в некоторой аптечной сети трудится несколько сотен работников первого стола. Соответственно каждый из них трудится с различной интенсивностью, однако информация о том, что величина средней выручки в расчете на одного сотрудника за смену составляет 82350 руб. дает определенное, хотя далеко не полное представление об интенсивности труда фармацевтического персонала данной аптечной сети.

Фармацевтический бизнес во всех его областях отличается исключительной сложностью, изменчивостью, индивидуальным многообразием явлений и процессов. В большинстве случаев фармацевтические бизнес-процессы протекают неоднозначно. По этой причине попытки применить в каждом конкретном случае ранее наработанные эмпирические подходы, технологии или методы могут дать далеко не самые лучшие результаты.

Причиной этого является наличие и воздействие субъективных факторов, обстоятельств, которые нельзя контролировать, но которые, тем не менее, влияют на результативность конкретного бизнес-процесса. Следовательно, неоднозначность протекания того или иного бизнес-процесса порождается наличием присущих ему случайных факторов. Однако это вовсе не означает отсутствие общих закономерностей, присущих бизнес-процессам. Например, приобретение препарата Актовегин в аптеке является случайным событием, но для пациентов с сосудистыми нарушениями головного мозга такие покупки являются уже не случайностью, а закономерностью. Выявлением и описанием случайных событий и явлений в контексте бизнес-процесса и занимается статистика.

Возможности бизнес-статистики достаточно широки, а ее результаты могут быть использованы как для информационных целей (количественная характеристика), так и для целей прогнозирования (кратко-, средне- и долгосрочные прогнозы), а также для решения целого комплекса аналитических задач.

С информационной точки зрения бизнес-статистика интересна тем, что представляет широкие возможности для сбора, обобщения и представления объективной информации о конкретном бизнес-процессе. В связи с тем, что для получения объективных результатов в ходе статистического исследования необходимо изучить тысячи событий (явлений), то в этом случае необходимым является переход от сплошного исследования к выборочному по нескольким показателям. В этом случае значение приобретает технология сбора, обработки и анализа данных, которая позволяет использовать информационные возможности частных исходных данных для обобщенной информации об изучаемом бизнес-процессе.

Что касается прогнозирования, то здесь роль бизнес-статистики заключается в оценке вероятностей тех или иных случайных событий, присущих бизнес-процессу. Кроме того, методами бизнес-статистики можно прогнозировать количественные значения показателей тех самых случайных величин. Такой прогноз, основанный на результатах статистического анализа, является фундаментом принятия управленческого решения. Немаловажен и тот факт, что на основе прогноза можно предусмотреть развитие тех или иных рисков, которые могут негативно повлиять на результативность рассматриваемого бизнес-процесса.

2. ОСНОВНЫЕ ПОЛОЖЕНИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ. СЛУЧАЙНЫЕ СОБЫТИЯ И СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

При изучении различных дисциплин, как в средних, так и в высших учебных заведениях предполагается определенная детерминированность событий и функций, когда каждое событие является следствием другого. Примером тому служат физические законы, представляющие собой строгие математические закономерности и описывающие зависимости одних величин от других.

Однако повседневная деятельность постоянно опровергает это положение. Например, при проверке любых физических законов (на лабораторных занятиях по физике) обнаруживается, что каждое новое экспериментальное измерение одного и того же параметра даёт различные результаты. Есть также величины, которые являются ещё более неопределенными. Невозможно, например, точно предсказать количество заболевших ОРВИ в производственном коллективе на определенный момент времени, длительность пребывания пациента в медицинской организации и др.

Величины, точное значение которых заранее неизвестно для каждой реализации, называются случайными. При этом случайность не означает, что невозможно получить и использовать на практике какую-либо информацию о процессе. Так при анализе заболеваемости можно отслеживать тенденции изменения этого показателя на 1000 человек и делать выводы о проведении необходимых мероприятий, планировании требуемого количества лечебных учреждений, медицинского персонала или лекарственных препаратов. Несмотря на случайный характер исследуемых величин, возможно нахождение определенных закономерностей, которые и используются далее в практической деятельности специалистов. Теоретической базой для этого выступает теория вероятностей и ее важнейшее понятие — *случайное явление*. *Теория вероятностей* — математическая наука, изучающая закономерности в массовых случайных явлениях независимо от их природы и дающая методы количественной оценки влияния случайных факторов на различные явления.

Большинство из окружающих нас явлений повседневной жизни являются случайными. Следует отметить, что теория вероятностей, как правило, абстрагируется от физического смысла изучаемого явления, а интересуется лишь ответом на вопрос: произошло или нет в данном эксперименте ожидаемое событие, каковы в этот раз были количественные показатели рассматриваемого случайного явления. Можно сказать, что случайное явление имеет две стороны: качественную и количественную. Качественной стороной случайного явления является случайное событие, а количественной — случайная величина.

2.1. Понятие вероятности случайного события

Событие — это факт, который в результате опыта (эксперимента, испытания) может произойти или не произойти. *Случайные события* принято обозначать заглавными буквами латинского алфавита, начиная с первых — *A, B, C, ...*

В повседневной жизни люди иногда пытаются охарактеризовать события по степени возможности их появления в той или иной ситуации, по частоте проявления их в более или менее одинаковых условиях. При этом не математики говорят об относительной частоте появления случайных событий, рано или поздно приходя к термину «*вероятность случайного события*».

В большинстве отраслей науки определение фундаментальных понятий — это весьма трудная, а чаще всего просто невыполнимая задача. Поэтому придётся дать несколько, может быть, не самых удачных на чей-то взгляд, подходов к понятию «вероятность случайного события». Отметим, что это не математические определения, не элементы аксиоматического построения теории вероятностей, сделанного в 1929–1933 годах А. Н. Колмогоровым [2], позволившее описать уже существовавшие к тому времени классические разделы теории вероятностей, дать толчок развитию её новых разделов, например теории случайных процессов, это лишь описание, иллюстрация научного термина.

---

[2] Колмогоров Андрей Николаевич (1903–1987) — сов. математик, академик АН СССР и Академии пед. наук СССР, ин. член Польской АН, Нидерландской королевской АН, Лондонского королевского общества, Румынской АН, Венгерской АН, Национальной АН США, Парижской АН, АН ГДР, Междунар. академии истории науки и пр.

Вероятность случайного события может рассматриваться как мера «объективной возможности» этого события, как количественная мера «степени уверенности» познающего субъекта. Это так называемое «*философское*» *определение вероятности случайного события*.

Вероятность случайного события принято обозначать $P(A)$ от латинского слова «*probabilitas*» — возможность. Одним из важнейших свойств вероятности случайного события является то, что она лежит в пределах от 0 (нуля) до 1 (единицы):

$$0 \le P(A) \le 1.$$

Кстати, это упрощенная запись первой из трёх аксиом в аксиоматическом построении теории вероятностей, сделанном А. Н. Колмогоровым. К сожалению, данное определение не отвечает на вопрос: «Откуда берутся объективные закономерности, о которых упоминалось в определении теории вероятностей как математической науки?».

Можно попытаться установить их из практики, основываясь на результатах наблюдений. С древних времен подмечено, что частость появления многих событий демонстрирует тенденцию к стабилизации около некоторого значения при неограниченном увеличении количества испытаний. Таким образом, приходим к изложенному ниже следующему определению.

Вероятность случайного события может рассматриваться как предел частости (частоты) появления события в большом количестве испытаний. Это можно назвать «*статистическим*» *определением вероятности случайного события*:

$$P^*(A) = \frac{m}{n} \to P(A) = \lim_{n \to \infty} \frac{m}{n}.$$

Подобный подход является, пожалуй, самым распространенным в трактовке термина «вероятность случайного события», им охватывается огромное множество случайных событий. Он не требует от исследователя математической строгости, требуется лишь одно — сколь возможно большое число экспериментов.

Ближе всего аксиоматическому подходу к определению понятия вероятности случайного события отвечает формулировка, приведенная в следующем пункте.

Вероятность случайного события рассматривается как мера равновероятности событий, «равновозможности» исходов опытов. Иногда это называют «*классическим*» *определением*. При этом вероятность события оценивается даже до проведения испытаний, основываясь только на структуре изучаемых явлений. Множество исходов опыта разбивается на группы равновозможных исходов, тогда вероятность случайного события определяется как отношение числа равновозможных исходов опыта, благоприятствующих появлению события A, к общему числу равновозможных исходов:

$$P(A) = m/n \ .$$

Уместно еще раз повторить, что отдельные исходы опыта должны иметь одинаковую возможность появления и общее их количество должно быть конечно.

Из классического определения вероятности вытекают её свойства. Вероятность достоверного события, т. е. такого, которое происходит неизбежно в результате каждого испытания, равна единице.

Действительно, для достоверного события U количество появления события $m = n$, следовательно, $P(U) = m/n = 1$. Вероятность невозможного события, т. е. такого, которое в результате каждого испытания вовсе не может произойти, равна нулю. Для невозможного события V количество появления события $m = 0$, откуда $P(V) = 0/n = 0$.

Вероятность любого события A удовлетворяет двойному неравенству:

$$0 \le P(A) \le 1 .$$

Действительно, любому событию благоприятствует число элементарных исходов опыта m, удовлетворяющее неравенству $0 \le m \le n$. Согласно классической формуле вероятности, $0 \le P(A) \le 1$.

В качестве оценки вероятности появления случайного события принимают его частость:

$$p^* = \frac{m}{n} .$$

Доказано, что оценка вероятности появления случайного события, вычисленная по приведенной выше формуле является несмещенной.

Доверительный интервал для оценки вероятности случайного события

Ширина доверительного интервала ε и требуемое количество экспериментов n для оценки вероятности случайного события выражаются формулами:

$$\varepsilon = \sqrt{\frac{p(1-p)}{n}} \cdot \Phi^{-1}(\beta);$$

$$n \ge \frac{p(1-p)}{\varepsilon^2} \cdot \left[\Phi^{-1}(\beta)\right]^2,$$

где в качестве вероятности p берётся частость случайного события p^*,

β — *доверительная вероятность* (*надёжность*) того, что истинное значение вероятности изучаемого случайного события будет лежать в интервале $(p^* - \varepsilon, p^* + \varepsilon)$,

$\Phi^{-1}(\beta)$ — *обратная функция Лапласа* (*прил. Б*).

Иными словами, доверительным интервалом для вероятности случайного события будет $(p^* - \varepsilon, p^* + \varepsilon)$.

Пример 2.1

Из общего числа 520 пациентов с врожденными пороками сердца у 260 был установлен диагноз «незаращенный артериальный проток». Оценить вероятность выявления данного заболевания у пациентов и определить её доверительный интервал:

$$p^* = m/n = 260/520 = 0{,}5.$$

Точность оценки вероятности данного случайного события по результатам наблюдения характеризуется величиной:

$$Sp = \frac{\sqrt{p\ (1-p)}}{n\ \Phi^{-1}(0{,}95)} = \frac{\sqrt{0{,}5\ (1-0{,}5)}}{520 \cdot 2} = 0{,}04.$$

Таким образом, вероятность наличия у пациента незаращенного артериального протока может находиться в диапазоне (0,5 ± 0,04), то есть, от 0,46 до 0,54.

2.2. Проверка гипотез о вероятности случайного события

Определение значения t-критерия Стьюдента в тех случаях, когда величина вероятности случайного события (относительного показателя) находится в интервале от 0,1 до 0,9

В случае, если значения вероятностей случайных событий, вычисленные по двум выборкам, находятся в интервале от 0,1 до 0,9, рекомендуется для вычисления t-критерия Стьюдента применять следующую формулу:

$$t = \frac{|\,\overline{p}_1 - \overline{p}_2\,|}{\sqrt{\sigma^2_{\overline{p}_1} + \sigma^2_{\overline{p}_2}}},$$

где $\overline{p}_1$ и $\overline{p}_2$ — сравниваемые выборочные вероятности случайного события;

$\sigma_{\overline{p}_1}$ и $\sigma_{\overline{p}_2}$ — их средние квадратические ошибки.

Число степеней свободы определяют так:

$$n' = n_1 + n_2 - 2\,,$$

где n_1 и n_2 — количество наблюдений в первой и второй выборках соответственно.

Определение значения t-критерия Стьюдента в тех случаях, когда величина вероятности случайного события приближается к 0 или 1

Прежде всего, чтобы «отступить» от нуля или единицы применяют поправку Йетса, тогда вместо частости, равной 0 (это же не невозможное событие), принимают $\overline{p} = 0 + 1/2n$; а вместо 1 (ведь здесь не достоверное событие) $\overline{p} = 0 - 1/2n$.

В случае, если значения вероятности случайного события, вычисленные по двум выборкам, меньше 0,1 или больше 0,9, рекомендуется для вычисления t-критерия Стьюдента применять метод, называемый «φ-преобразование Фишера», заменяя значение $\overline{p}$ на φ по формуле $\varphi = 2 \cdot arcsin\sqrt{\frac{\overline{p}}{1}}$.

Суть углового преобразования Фишера состоит в переводе процентных долей в величины центрального угла, который измеряется в радианах. Большей процентной доле будет соответствовать больший угол φ, а меньшей доле — меньший угол, но соотношения здесь не линейные: $\varphi = 2 \cdot arcsin\left(\sqrt{p}\right)$, где p — процентная доля, выраженная в долях единицы.

При увеличении расхождения между углами φ_1 и φ_2 и увеличения численности выборок значение критерия возрастает. Чем больше величина φ*, тем более вероятно, что различия достоверны.

Среднее квадратическое отклонение вероятности случайного события (относительного показателя), выраженное через φ, определяется по формуле:

$$\sigma_{\varphi} = \frac{1}{\sqrt{n}}.$$

Оценка значимости различия выборочных показателей, выраженных через φ производится по формуле[3]:

$$t = \left|\varphi_1 - \varphi_2\right| \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \; .$$

2.3. Определение частостей случайных событий. Оценка значимости различия частостей случайных событий в независимых выборках с помощью пакетов прикладных программ (ППП) MS Excel и Statistica StatSoft. Inc

Постановка задачи

Исследуется уровень летальности при различных формах внутрибольничной инфекции у больных с сочетанной травмой. В хирургической клинике сформированы данные о количестве

[3] Критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннесса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

наблюдений и случаев летальности для четырех форм острых гнойно-септической инфекции (*табл. 2.1*).

Таблица 2.1

Число случаев летальных исходов
при различных формах внутрибольничной инфекции

Номер группы	Форма гнойно-септической инфекции	Число больных	Число летальных исходов
1	Трахеобронхит	26	3
2	Пневмония	18	7
3	Уроинфекция	5	3
4	Сепсис	4	3

Определить частости (относительные величины частоты) летальных исходов; оценить их точность и надежность.

Построить график частости летальных исходов с указанием 95 % доверительных интервалов.

Определить уровни значимости различия частостей летальных исходов для различных форм заболевания.

Решение задачи

Ввести исходные данные в MS Excel.

На этом же листе повторить (скопировать) заголовок таблицы (№ группы, Форма заболевания) чуть ниже исходных данных. То есть, *табл. 2.1* может быть скопирована относительно расположения её в постановке задачи.

Рассчитать частости летальных исходов *P* с использованием ввода формулы: $= \mathrm{D3}/\mathrm{C3}$; скопировать эту формулу для групп № 2 и № 3, № 4. Для группы № 1 следует учесть поправку Йетса: $= \dfrac{\mathrm{D6}}{\mathrm{C6}} + \dfrac{1/2}{\mathrm{C6}}$.

Для групп №№ 2, 3, 4 рассчитать среднюю квадратическую ошибку относительной величины частоты (ОВЧ) летальных исходов: =КОРЕНЬ(C11*(1-C11)/C4) и т. п.

Рассчитать значения переменной Фишера для группы № 1: = 2·ASIN(КОРЕНЬ(D3/C3)).

Вычислить среднюю квадратическую ошибку переменной Фишера: =1/КОРЕНЬ(C3) для этой группы.

Найти границы 95%-го доверительного интервала для переменной Фишера: = E10 ± 1,96·F10 для группы № 1.