УДК 622.232

ОБОСНОВАНИЕ ОБЪЕМА ВЫБОРКИ И КРИТИЧЕСКОГО ЗНАЧЕНИЯ КРИТЕРИЯ СОГЛАСИЯ ПРИ ПРОВЕРКЕ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

д-р техн. наук, проф. С.Г. ЕХИЛЕВСКИЙ,
канд. физ.-мат. наук, доц. О.В. ГОЛУБЕВА, канд. физ.-мат. наук, доц. Н.А. ГУРЬЕВА
(Полоцкий государственный университет)

В работе показано, что истинной независимой переменной, определяющей объем выборки, вид критерия согласия и его критическое значение, является приемлемый риск. И нет соображений, позволяющих минимизировать риск, если для альтернативной гипотезы не конкретизирован закон распределения критерия. При таком подходе уровень значимости и мощность критерия возникают на стадии промежуточных вычислений в качестве побочного продукта и как самостоятельные понятия не требуются для решения задачи о проверке статистических гипотез.

Как известно, статистической называют гипотезу о виде закона распределения случайной величины и его параметрах. Принять или отвергнуть гипотезу нужно на основе данных выборки, что вносит элемент случайности и может привести к ошибкам (принятию неверной гипотезы или неприятию верной), поэтому можно говорить лишь о некоторой вероятности того, что гипотеза имеет (или не имеет) место.

Для принятия решения из данных выборки составляют некоторую случайную величину (критерий), возможные значения которой в зависимости от требуемой вероятности делят на две части. Внутри одной из них гипотезу принимают, во второй – отвергают.

Пример. Пусть гипотеза состоит в том, что матожидание некоторого признака генеральной совокупности равно m. Критерием в этом случае является выборочное среднее, областью принятия – некоторая δ-окрестность *m*, отвечающая доверительной вероятности γ.

Следует подчеркнуть, что при этом, во-первых, остается конечная вероятность $\alpha = 1 - \gamma$ отклонить правильную гипотезу, называемую уровнем значимости; во-вторых, само принятое решение может измениться при увеличении объема выборки (δ-окрестность уменьшится, и критерий перестанет в нее попадать).

Таким образом, сразу возникает три вопроса:

1. Из каких соображений выбирать γ (а значит и границы окрестности, являющиеся критическими значениями критерия)?
2. Каким объемом следует ограничить выборку?
3. Какой из критериев предпочесть остальным?

С ответами на эти вопросы непосредственно связан метод минимума риска, излагаемый, однако, в литературе вне данного контекста [1].

Метод минимума риска

Пусть ущерб, связанный с ошибкой первого рода (отклонение правильной гипотезы), равен C_1, а второго рода (принятие неправильной) – C_2[1]. Тогда матожидание ущерба *C*, называемое риском, равно

$$r = M(C) = P_1 \cdot C_1 + P_2 \cdot C_2, \tag{1}$$

где P_1 – вероятность ошибки первого рода:

$$P_1 = P(H) \cdot P_H(\bar{A}). \tag{2}$$

Здесь $P(H)$ – априорная (найденная до обследования выборки) вероятность проверяемой гипотезы, а $P_H(\bar{A})$ – вероятность эту гипотезу отклонить при условии, что она имеет место.

Аналогично вероятность совершения ошибки второго рода представим в виде

$$P_2 = P(\bar{H}) \cdot P_{\bar{H}}(A), \tag{3}$$

где $\bar{H}$ – альтернативная гипотеза[2]; A – ее непринятие.

[1] Определение C_1 и C_2 – задача специалистов той отрасли, которая прибегает к услугам математической статистики.
[2] Заметим, что гипотез может быть много, однако все они, кроме *H*, могут быть объединены в альтернативную.

Фигурирующие в (2), (3) условные вероятности зависят от того, какой критерий согласия используется, чему равно его критическое значение и каков объем выборки. Естественно при данном объеме выборки критическое значение критерия выбирать так, чтобы риск был минимален. Если даже в этом случае он оказывается неприемлемо велик, следует увеличивать выборку, принимая, однако, во внимание растущие затраты на сбор и обработку статистических данных. Последнее обстоятельство может не позволить снизить обусловленные ошибками издержки до приемлемых. Поэтому из нескольких критериев предпочтительнее критерий, обеспечивающий данный риск минимальным объемом выборки.

Пример. Изделия одинаковыми партиями поступают с двух заводов. Вероятности выпуска бракованной продукции первым и вторым заводами равны соответственно p_1 и p_2. Известны доли всех партий, поступающих с первого и второго заводов. Требуется на основании данных выборочного контроля принять решение о том, на каком заводе изготовлена выставленная на продажу партия.

Решение. Пусть n – объем выборки. Значит, число бракованных изделий m – случайная величина с возможными значениями $0, 1, 2, ..., n.$ Очевидно m можно использовать в качестве критерия согласия. Положим для определенности $p_1 > p_2$. Тогда, если $m > m^*$, считаем, что партия изготовлена на первом заводе. Здесь m^* – критическое значение критерия, которое предстоит определить.

Нетривиальность задачи состоит в том, что из плохой партии (изготовленной первым заводом) можно случайно осуществить выборку качественных изделий[3], а из хорошей – бракованных. Поэтому партию первого завода можно считать изготовленной на втором и наоборот. Коротко эти ошибки будем называть первой и второй соответственно. Найдем их вероятности.

По схеме Бернулли имеем следующие законы распределения m для первого $(i = 1)$ и второго $(i = 2)$ заводов:

$$P_n^i(m) = C_n^m p_i^m q_i^{n-m}, \tag{4}$$

где $q_i = 1 - p_i$ – вероятность выпуска качественной продукции i-м заводом. Пусть гипотеза H заключается в том, что выставленная на продажу партия изготовлена на первом заводе ($\bar{H}$ – на втором). В соответствии с ранее изложенным вероятность отклонить H при условии, что она верна, равна сумме вероятностей хороших выборок ($m \le m^*$) из плохой партии ($i = 1$):

$$P_H(\bar{A}) = \sum_{m \le m^*} P_n^1(m). \tag{5}$$

Случайное событие $\bar{A}$ заключается в том, что гипотеза H отвергается.

Аналогично для альтернативной гипотезы имеем

$$P_{\bar{H}}(A) = \sum_{m > m^*} P_n^2(m) = 1 - \sum_{m \le m^*} P_n^2(m), \tag{6}$$

где A означает отклонение $\bar{H}$ (принятие H). В (6) учтено, что полная вероятность равна единице.

По условию априорные вероятности гипотез известны (это доли партий, поступающих с первого и второго заводов), поэтому, подставив (5) – (6) в (1), найдем риск как функцию m^*

$$r(m^*) = C_1 P(H) \sum_{m \le m^*} P_n^1(m) + C_2 P(\bar{H}) \sum_{m > m^*} P_n^2(m) = C_1 P(H) \sum_{m \le m^*} P_n^1(m) + C_2 P(\bar{H}) \left(1 - \sum_{m \le m^*} P_n^2(m) \right) = C_2 P(\bar{H}) + \sum_{m \le m^*} f(m), \tag{7}$$

где

$$f(m) = C_1 P(H) P_n^1(m) - C_2 P(\bar{H}) P_n^2(m). \tag{8}$$

Если n достаточно велико, то $f(0) < 0$, а $f(n) > 0$, в чем легко убедиться с помощью (4), (8), приняв во внимание, что $p_1 > p_2$:

$$\left.\begin{aligned} f(0) &= C_1 P(H) q_1^n - C_2 P(\bar{H}) q_2^n \xrightarrow[n \to \infty]{} - C_2 P(\bar{H}) q_2^n \\ f(n) &= C_1 P(H) {p_1}^n - C_2 P(\bar{H}) {p_2}^n \xrightarrow[n \to \infty]{} C_1 P(H) {p_1}^n \end{aligned}\right\} \tag{9}$$

[3] Более того, первый завод может случайно выпустить даже хорошую партию (${p_1}^* \le p_2$, где ${p_1}^*$ – фактическая доля брака). Однако это крайне маловероятно (партия намного больше выборки) и никак не учитывается.

Чтобы при данном n риск был минимален, суммировать в (7) нужно до тех пор, пока $f(m)$ отрицательно. Из условия $f(m)=0$ определим критическое значение критерия согласия

$$m^* = \frac{\ln \frac{\eta q_1^n}{q_2^n}}{\ln \frac{p_2 q_1}{p_1 q_2}} \equiv m^*(n), \tag{10}$$

где

$$\eta = \frac{C_1 P(H)}{C_2 P(\bar{H})}. \tag{11}$$

На этом решение задачи о проверке статистической гипотезы «выставленная на продажу партия изготовлена первым заводом» можно считать завершенным. Если по результатам выборочного контроля $m > m^*$ гипотезу следует принимать, а при $m < m^*$ отвергать. Точку $m = m^*$ [4] можно отнести как к критической области, так и к области принятия гипотезы. На величину риска это не влияет, ибо $f(m^*)=0$. Подчеркнем, что мы никак не задавали уровень значимости и мощность критерия (и даже не вводили эти понятия).

Проинтерпретируем полученные результаты.

Вначале выясним, каким должен быть объем выборки. Для малых n может оказаться, что m^* отрицательно или больше n, если очень велик ущерб, связанный соответственно с первой и второй ошибками или велики соответствующие априорные вероятности. В этом легко убедиться, рассмотрев в (10) предельные ситуации с η, стремящимся к бесконечности или к нулю. В первом случае, чтобы меньше рисковать, партию нужно считать худшей (принимать *H*), даже если оказалось, что $m=0$. А во втором – качественной даже при $m=n$. Очевидно, что такие выборки неинформативны и делать их бессмысленно. В связи с этим определим минимальное n, при котором результаты выборочного контроля влияют на принимаемое решение. Заметим для этого, что логарифм в знаменателе (10) отрицателен. Поэтому первая ситуация ($m^* < 0$) не возникнет, если

$$\frac{\eta q_1^n}{q_2^n} \le 1,$$

откуда

$$n \ge \frac{\ln \eta}{\ln \frac{q_2}{q_1}}. \tag{12}$$

А чтобы не возникла вторая ситуация ($m^* > n$), требуется (см. (10)) выполнение неравенства:

$$\frac{\ln \frac{\eta q_1^n}{q_2^n}}{\ln \frac{p_2 q_1}{p_1 q_2}} \le n,$$

из которого следует

$$n \ge \frac{\ln \eta}{\ln \frac{p_2}{p_1}}. \tag{13}$$

Видно, что при $p_1 \to p_2$, когда различие между партиями разных заводов нивелируется, объем информативной выборки неограниченно возрастает. В случае строгого равенства $p_1 = p_2$ ни одно из условий (12), (13) не может быть выполнено, то есть выборочный контроль не позволяет установить происхождение партии.

[4] Такое равенство возможно не всегда (только если m^* окажется целым числом).

В зависимости от конкретных C_1, C_2, p_1, p_2, $P(H)$ и $P(\bar{H})$ при определении минимального объема информативной выборки пользоваться нужно тем из условий (12), (13), которое нарушено для малых n (напомним, что по смыслу $n>0$). Знаменатели в (12), (13) разных знаков, поэтому оба условия сразу нарушены быть не могут.

Выясним теперь, как риск зависит от объема выборки. Для малых n (неинформативных выборок) он определяется априорными вероятностями гипотез. В случае равенства (12) имеем $m^*=0$, то есть неравенство $m<m^*$ невозможно, и гипотеза H не будет отклонена ни при каких результатах выборочного контроля, что исключает ошибку первого рода. Следовательно $r=C_2P(\bar{H})$, в чем можно убедиться и непосредственно, приняв во внимание, что сумма в правой части (7) состоит из одного слагаемого $m=m^*=0$, к тому же равного нулю $f(m^*)=0$.

Аналогично в случае равенства (13) имеем $m^*=n$, то есть приниматься всегда будет $\bar{H}$, что делает невозможной ошибку второго рода, поэтому $r=C_1P(H)$. В последнем нетрудно убедиться непосредственно, заметив в (7), что $\sum_{m=0}^{n}P_n^i(m)=1$.

Для информативных n при оптимальном выборе m^* следует пользоваться соотношением (7) с $m^*=m^*(n)$, определяемым формулой (10):

$$r(m^*(n))=C_2P(\bar{H})+\sum_{m=0}^{m^*(n)}f(m). \tag{14}$$

Задавшись приемлемым значением риска, с помощью (14) можно найти достаточный для него объем выборки. Естественно, объем будет тем меньше, чем больше отличие между p_1 и p_2. В предельной ситуации, когда первый завод выпускает сплошной брак, а второй – только качественную продукцию ($p_1-p_2=1$), происхождение партии достоверно определяется контролем единственного изделия (то есть $r=0$ уже для $n=1$). Последнее совершенно очевидно, однако с целью проверки развитого формализма убедимся в этом аналитически. Подставив в (10) p_1 и p_2, равные соответственно $1-\varepsilon$ и ε, для бесконечно малых ε получим

$$m^*(n)=\frac{\ln\eta+n\ln\dfrac{\varepsilon}{1-\varepsilon}}{\ln\dfrac{\varepsilon^2}{(1-\varepsilon)^2}}\xrightarrow[\varepsilon\to0]{}\frac{n}{2}.$$

Таким образом, для $n=1$ целочисленное по смыслу m может принять в сумме (7) только нулевое значение:

$$r(0)=C_2\cdot P(\bar{H})+C_1\cdot P(H)\cdot P_1^1(0)-C_2\cdot P(\bar{H})\cdot P_1^2(0).$$

В результате с учетом (4) получим

$$r(0)=C_2\cdot P(\bar{H})+C_1\cdot P(H)\cdot\varepsilon-C_2\cdot P(\bar{H})\cdot(1-\varepsilon)=\varepsilon\left(C_1\cdot P(H)+C_2\cdot P(\bar{H})\right)\xrightarrow[\varepsilon\to0]{}0,$$

что и требовалось доказать.

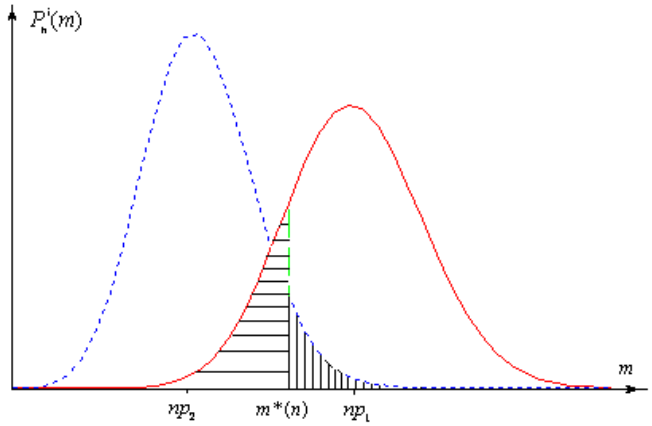
Для очень больших n риск сводится к нулю при любом конечном различии между p_1 и p_2. Доказательство проведем с помощью следующей из (10) асимптотической формулы:

$$m^*(n)\underset{n\to\infty}{\to}n\frac{\ln\dfrac{q_1}{q_2}}{\ln\dfrac{p_2q_1}{p_1q_2}}<n. \tag{15}$$

При этом имеет место двойное неравенство:

$$np_2<m^*(n)<np_1. \tag{16}$$

В справедливости неравенства (16) нетрудно убедиться непосредственно (методами дифференциального исчисления)[5], или приняв во внимание, что качественная выборка из плохой партии и некачественная из хорошей – события маловероятные. Но при $m^* < np_2$ не будет малой $P_{\bar{H}}(A)$ (рисунок), а при $m^* > np_1$ соответственно $P_H(\bar{A})$.



Площади горизонтально и вертикально заштрихованных областей равны вероятностям $P_H(\bar{A})$ и $P_{\bar{H}}(A)$ соответственно

Оценка (16) означает, что разность между критическим значением критерия и матожиданием числа бракованных деталей (для обоих заводов) пропорциональна n. Значит

$$\left.\begin{aligned} x_1(m^*(n)) &= \frac{m^*(n) - np_1}{\sigma_1} \underset{n\to\infty}{\to} -\infty; \\ x_2(m^*(n)) &= \frac{m^*(n) - np_2}{\sigma_2} \underset{n\to\infty}{\to} +\infty, \end{aligned}\right\} \qquad (17)$$

где $\sigma_i = \sqrt{np_iq_i}$ – среднеквадратические отклонения m для первого $(i=1)$ и второго $(i=2)$ заводов.

Асимптотика (17) означает, что в соответствии с интегральной теоремой Лапласа при больших n бесконечно малыми становятся $P_H(\bar{A})$ и $P_{\bar{H}}(A)$:

$$P_H(\bar{A}) \underset{n\to\infty}{\to} 0{,}5 - \big(\Phi(0) - \Phi(x_1)\big) = 0{,}5 + \Phi(x_1) \underset{x\to-\infty}{\to} 0;$$

$$P_H(\bar{A}) \underset{n\to\infty}{\to} 0{,}5 - \big(\Phi(x_2) - \Phi(0)\big) = 0{,}5 - \Phi(x_2) \underset{x\to+\infty}{\to} 0,$$

а с ними и риск (см. (1) – (3)). Что и требовалось доказать.

На практике затраты на осуществление выборочного контроля не позволяют неограниченно увеличивать n. Будем считать затраты пропорциональными объему выборки, добавим их к риску и таким образом определим средние издержки, связанные с однократным принятием решения об истинности (ложности) проверяемой статистической гипотезы

$$I(n) = r(m^*(n)) + z\,n, \qquad (18)$$

где z – затраты на обследование одного объекта выборки.

---

[5] Подставив (15) в (16) и приняв во внимание отрицательность логарифма в знаменателе, перепишем правое неравенство в эквивалентном виде $\ln q_1 + p_1 \ln\frac{p_1}{q_1} > \ln(1-p_2) + p_1 \ln\frac{p_2}{1-p_2} \equiv y(p_2)$. Равенство достигается на правой границе области определения $y(p_2)$ (напомним, что $p_2 \in (0, p_1)$). С учетом последнего обстоятельства для доказательства неравенства достаточно убедиться в монотонном возрастании его правой части $y(p_2)' = \frac{p_1 - p_2}{p_2 q_2} > 0$.

Аналогично доказывается левое неравенство (16): $\ln q_2 + p_2 \ln\frac{p_2}{q_2} > \ln(1-p_1) + p_2 \ln\frac{p_1}{1-p_1} \equiv y(p_1)$, $y(p_1)' = \frac{p_2 - p_1}{p_1 q_1} < 0$ и этого достаточно, так как равенство теперь достигается на левой границе ($p_1 \in (p_2, 1)$).

Первое слагаемое в (18), как показано выше, с ростом n монотонно стремится к нулю, а второе – неограниченно увеличивается. Значит, существует оптимальное n, обеспечивающее минимум $I(n)$, которое можно найти из условия

$$r(m^*(n))' = -z\,. \tag{19}$$

Если даже найденное из (19) n не обеспечивает приемлемое $I(n)$, нужно пользоваться более эффективным критерием согласия, а если его нет – совершенствовать технологию производства и выборочного контроля (уменьшать C_1, C_2 и z).

Обсудим теперь проблему выбора уровня значимости. Его определение, приведенное выше, не совсем удачно, ибо правильной по воле случая может оказаться любая из альтернативных гипотез. Чтобы избежать двусмысленности, уровнем значимости данной гипотезы назовем условную вероятность ее отвергнуть при условии, что она (данная гипотеза) верна. В рамках такого определения уровни значимости гипотез вычисляются по формулам (5), (6), в которых фигурирует m^*, определяемое с помощью (10) для n, найденного из условия (19). Впрочем, сделать это можно лишь из чистого любопытства, так как для принятия (отклонения) проверяемой гипотезы эти вычисления не нужны. В рамках подхода, основанного на минимизации $I(n)$, истинными независимыми переменными, задающими объем выборки и критическое значение критерия согласия, являются η, p_1, p_2 и z.

Укажем в этой связи, что учебные задачи, в которых в той или иной форме предлагается с помощью готовых таблиц проверить, укладываются ли данные n, α и некоторое значение критерия согласия в рамки определенной гипотезы [2], являются профанацией, поскольку нет соображений, позволяющих выбрать уровень значимости, если для альтернативной гипотезы не конкретизирован закон распределения критерия. В частности, в рассмотренном примере отсутствие информации о p_2 не позволяет даже убедиться в информативности выборки (см. условия (12), (13)). Без чего статистическая проверка гипотезы не может считаться корректной.

ЛИТЕРАТУРА

1. Гнеденко, Б.В. Курс теории вероятностей / Б.В. Гнеденко. – М.: Наука, 1969. – 400 с.
2. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике / В.Е. Гмурман. – М.: Высш. шк., 1975. – 334 с.

Поступила 26.09.2011

SUBSTANTIATION OF THE SAMPLE SIZE AND THE CRITICAL VALUE OF THE CRITERION OF THE CONSENT OF THE WHEN TESTING STATISTICAL HYPOTHESES

S. EKHILEVSKIY, O. GOLUBEVA, N. GURIEVA

In work it is shown that the true independent variable defining volume of sample, a kind of criterion of the consent and its critical value, is comprehensible risk. Also there are no the reasons, allowing to minimize risk if for an alternative hypothesis the law of distribution of criterion isn't concretized. At such approach the significance value and capacity of criterion arise at a stage of intermediate calculations as a by-product and as independent concepts aren't required for the decision of a problem on check of statistical hypotheses.