**Департамент политологии, 2017-18 уч. год**
**Математика и статистика, часть 1.**
**Лекция 7. Элементы теории вероятностей, часть 3 (04.12.17)**
*И. А. Хованская, Р. Я. Будылин, И. В. Щуров, Д. А. Филимонов, К. И. Сонин (РЭШ)*

# 1 Формула полной вероятности

Рассмотрим следующую ситуацию. Допустим, в некотором вузе есть три факультета: математический, физический и химический. Факультеты отличаются по числу студентов — например, на математическом факультете учится 1/10 студентов всего вуза, на физическом — 3/10, а на химическом — оставшиеся 6/10 (то есть 3/5) всех студентов вуза (см. табл. 1).[1] Мы также знаем, какова доля отличников на каждом факультете. Например, на математическом факультете каждый четвертый является отличником, на физическом — каждый второй студент отличник, а на химическом 3/4 студенов являются отличниками. Спрашивается, какова доля отличников в целом по вузу?

| факультет | матфак | физфак | химфак |
|---|---|---|---|
| какая часть студентов учится на факультете | 1/10 | 3/10 | 6/10 |
| доля отличников на факультете | 1/4 | 1/2 | 3/4 |

Таблица 1: Распределение студентов по факультетам

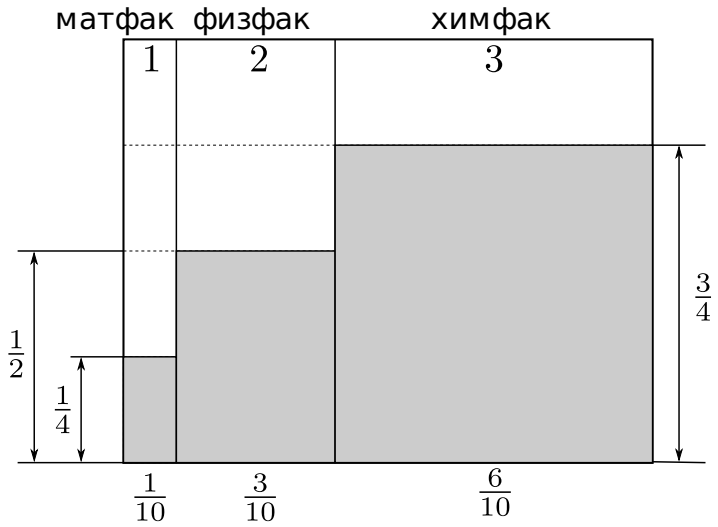


Рис. 1: Распределение студентов по факультетам. Закрашенные прямоугольники соответствуют отличникам. Их высоты — условным вероятностям.

Допустим для простоты, что всего в вузе учится 1000 студентов. Тогда на математическом факультете будет $1000 \times \frac{1}{10} = 100$ студентов. Среди них четвертая часть (то

[1] Конечно, мы считаем, что каждый студент учится только на одном факультете, и каждый учится на каком-нибудь факультете из этих трёх.

есть $1000 \times \frac{1}{10} \times \frac{1}{4} = 100 \times \frac{1}{4} = 25$ человек) являются отличниками, и это составляет $25/1000 = 0.025$ от студентов всего вуза. Нетрудно видеть, что это число не зависит от того, сколько студентов всего учится в вузе — если их $N$ человек, то среди них $N \times \frac{1}{10}$ учится на математическом факультете, отличников с математического факультета $N \times \frac{1}{10} \times \frac{1}{4}$ человек, и их доля среди всех $N$ студентов вуза равна $N \times \frac{1}{10} \times \frac{1}{4} \times \frac{1}{N} = \frac{1}{10} \times \frac{1}{4} = 0.025$. (Вспомните задачи про фонды и компании, которые владели какими-то частями друг друга...)

Здесь также можно привести простую геометрическую интерпретацию (см. рис. 1): на рисунке, отличникам с математического факультета соответствует крайний левый закрашенный прямоугольник, ширина которого равна 1/10, а высота — 1/4; общая доля отличников с математического факультета таким образом равна площади этого прямоугольника, то есть 0.025.

Иными словами, можно записать:

$$\begin{aligned}\text{Доля отличников с мат. факультета среди студентов всего вуза} = \\ = \text{доля студентов этого факультета среди студентов всего вуза} \times \\ \times \text{ доля отличников на математическом факультете}\end{aligned} \quad (1)$$

На физическом факультете учится $1000 \times \frac{3}{10} = 300$ студентов. Отличников среди них половина, то есть $300 \times \frac{1}{2} = 150$ человек. Это составляет $150/1000 = 0.15$ от всех студентов вуза (средний закрашенный прямоугольник на рисунке). Наконец, на третьем — химическом — факультете учится $1000 \times \frac{6}{10} = 600$ студентов. Из них три четверти отличники, то есть $600 \times \frac{3}{4} = 450$ студентов, что составляет $450/1000 = 0.45$ от всех студентов вуза (крайний правый закрашенный прямоугольник).

Поскольку каждый отличник учится ровно на одном факультете, чтобы найти число всех отличников, достаточно сложить полученные числа: всего отличников $25 + 150 + 450 = 625$ человек. Доля отличников по всему вузу составляет $625/1000 = 0.625$. Заметим, что к тому же результату мы бы пришли, если бы складывали *доли отличников от всего вуза* по каждому факультету: $0.025 + 0.15 + 0.45 = 0.625$. Такую долю составляет закрашенная часть большого прямоугольника на рисунке 1 от всего большого прямоугольника. Ответ не изменился бы, если бы в вузе училось не 1000 студентов, а любое другое число (проверьте!).

Переведем теперь рассмотренную задачу на язык теории вероятностей. Допустим, мы выбираем случайного студента вуза. Пространство элементарных исходов здесь совпадает со множеством всех студентов (по-прежнему будем считать, что их 1000 человек). Можно рассмотреть различные события — например, событие $H_1$ — «студент учится на математическом факультете», событие $H_2$ — «студент учится на физическом факультете», событие $H_3$ — «студент учится на химическом факультете», а также событие $A$ — «студент является отличником». Каковы вероятности таких событий?

Событию $H_1$ («студент учится на математическом факультете») благоприятствует столько исходов, сколько студентов учится на матфаке. Вероятность события $H_1$ в таком случае равна отношению числа студентов математического факультета к числу студентов всего вуза, то есть будет в точности совпадать с долей студентов мат. факультета во всём вузе — по условию, она равна 1/10. Аналогично обстоит ситуация с событиями $H_2$ и $H_3$.

Событию $A$ («студент оказался отличником») благоприятствует столько исходов, сколько отличников учится в вузе (в примере — 625 человек). Таким образом, вероятность случайно выбрать отличника из всего вуза — это в точности доля студентов-отличников во всем вузе (то есть 0.625).

Допустим теперь, что мы выбрали случайного студента среди студентов математического факультета. Какова в таком случае вероятность, что он окажется отличником? С точки зрения теории вероятностей, это будет *условная вероятность события $A$ при условии $H_1$* (то есть $p(A|H_1)$). Нетрудно видеть, что эта вероятность в точности равна доле отличников среди студентов математического факультета (то есть $1/4$).

Предположим, что мы знаем только те данные, которые приведены в условии задачи. Иными словами, нам даны вероятности событий $H_1$, $H_2$, $H_3$, а также условные вероятности события $A$ при условиях $H_1$, $H_2$, $H_3$ (то есть $p(A|H_1)$, $p(A|H_2)$, $p(A|H_3)$). Мы также знаем, что события $H_1$, $H_2$, $H_3$ попарно несовместны (то есть один студент может учиться только на одном факультете) и вместе образуют всё пространство элементарных исходов (любой студент учится хотя бы на одном факультете). Иными словами, эти события разбивают всё пространство элементарных исходов на несколько непересекающихся частей — в этом случае такие события часто называют *гипотезами*, и говорят, что они образуют «полную систему событий». Тот факт, что гипотезы попарно несовместны, и что в объединении они дают всё пространство элементарных событий очень важен — он означает, что всегда реализуется в точности одна гипотеза.

Нетрудно видеть, что по указанным данным можно найти вероятность события $A$ (то есть долю отличников среди студентов всего вуза) — точно таким образом, как мы это сделали выше. Запишем приведенное решение в терминах теории вероятностей:

Напомним, что $p(A|H_1)$ — это доля отличников на математическом факультете, $p(H_1)$ — доля студентов математического факультета на всем вузе, и значит $p(A|H_1) \times p(H_1)$ — доля отличников с математического факультета среди всех студентов вуза (см. формулу (1)). Иными словами, это то же самое, что $p(AH_1)$ — вероятность одновременного выполнения $A$ и $H_1$ (в примере — вероятность выбрать одного из 25 отличников с математического факультета — она составляет 0.025). Аналогично $p(A|H_2) \times p(H_2)$ — доля отличников со физического факультета среди всех студентов вуза, $p(A|H_3) \times p(H_3)$ — доля отличников с химического факультета среди всех студентов вуза. Складывая эти доли, получаем долю всех отличников вуза, то есть $p(A)$.

$$p(A) = p(A|H_1) \times p(H_1) + p(A|H_2) \times p(H_2) + p(A|H_3) \times p(H_3), \tag{2}$$

Говоря человеческим языком, это означает, что доля всех отличников в вузе это сумма долей отличников каждого факультета от всех студентов вуза.

Это рассуждение можно записать более формально. Действительно, напомним, что по определению условной вероятности,

$$p(A|H_1) = \frac{p(AH_1)}{p(H_1)}, \tag{3}$$

где $p(AH_1)$ — вероятность одновременного выполнения $A$ и $H_1$. Умножая обе части этого равенства на $p(H_1)$, получим:

$$p(AH_1) = p(A|H_1) \times p(H_1). \tag{4}$$

С другой стороны, поскольку события $H_1$, $H_2$, $H_3$ не пересекаются и вместе образуют всё пространство элементарных исходов, событие $A$ можно представить как объединение кусочков, лежащих в $H_1$, $H_2$, $H_3$, то есть $A = AH_1 + AH_2 + AH_3$. События $AH_1$, $AH_2$, $AH_3$ тоже не пересекаются, и значит их вероятности можно складывать: $p(A) = p(AH_1 + AH_2 + AH_3) = p(AH_1) + p(AH_2) + p(AH_3)$. Учитывая (4), имеем: $p(A) = p(A|H_1) \times p(H_1) + p(A|H_2) \times p(H_2) + p(A|H_3) \times p(H_3)$.

Доказанная нами формула (2) называется *формулой полной вероятности*. В общем случае гипотез может быть не три, а больше, и формула выглядит так.

Пусть имеется полная система событий $H_1, H_2, ..., H_n$: все эти события попарно несовместны и их сумма составляет всё пространство элементарных исходов $\Omega$. Пусть нам известны вероятности $p(H_1),\ p(H_2), \ldots,\ p(H_n)$ событий (гипотез) $H_1,\ H_2, \ldots,\ H_n$. Разумеется, $p(H_1) + p(H_2) + \cdots + p(H_n) = 1$. Пусть мы знаем условные вероятности выполнения события $A$ при условии выполнения каждого из событий $H_1,\ H_2, \ldots,\ H_n$, т. е. $p(A|H_1),\ p(A|H_2), \ldots,\ p(A|H_n)$. Тогда вероятность события $A$ (полная вероятность) составляет

$$p(A) = p(H_1)p(A|H_1) + p(H_2)p(A|H_2) + \cdots + p(H_n)p(A|H_n)$$

**Пример:**

1. На столе лежат три монетки: две обыкновенные и одна с орлами с двух сторон. Наудачу выбирается одна монетка и подбрасывается. С какой вероятностью выпадет орел?

Мы могли выбрать либо обычную монетку, либо монетку с двумя орлами. В каждом из этих случаев не сложно найти вероятность выпадения орла, поэтому обозначим за $H_1$ событие (гипотезу) — мы взяли обычную монетку, за $H_2$ событие (гипотезу) — мы взяли монетку с двумя орлами. Интересующее нас событие $A$ — выпадение орла на выбранной монетке. Итак, имеем:

$$p(H_1) = 2/3, \quad p(A|H_1) = 1/2$$
$$p(H_2) = 1/3, \quad p(A|H_2) = 1$$

Значит,

$$p(A) = p(H_1)p(A|H_1) + p(H_2)p(A|H_2) = \frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 = \frac{2}{3}.$$

## 2 Формула Байеса

Вернемся к задаче об отличниках. Пусть мы выбрали случайным образом студента, и он оказался отличником. Теперь нас интересует вероятность того, что он учится, скажем, на математическом факультете. Итак, задача в отыскании условной вероятности того, что студент учится на первом факультете, при условии того, что он отличник. Что мы знаем? Вероятности $p(H_1)$, $p(H_2)$, $p(H_3)$ того, что студент учится на каждом из трех факультетов, вероятности (условные!) $p(A|H_1),\ p(A|H_2),\ p(A|H_3)$ того, что студент, учась на каждом из этих факультетов, будет отличником. Нас же интересует условная вероятность $p(H_1|A)$, что студент, *оказавшийся* отличником, учится на математическом факультете.

Иными словами, нам нужно вычислить, какую долю составляет площадь левого закрашенного прямоугольника на рис. 2 (выделен более темным цветом), соответствующего отличникам с математического факультета (то есть событию $AH_1$), от площади всей закрашенной области (соответствующей всем отличникам, т.е. событию $A$).

Напомним, что по формуле полной вероятности, $p(A)$ представляется в виде суммы трех слагаемых, каждое из которых выражает долю отличников с соответствующего факультета от числа всех студентов вуза (на рисунке — площади соответствующих закрашенных прямоугольников):

$$p(A) = p(A|H_1) \times p(H_1) + p(A|H_2) \times p(H_2) + p(A|H_3) \times p(H_3)$$

В этом месте есть кажущееся несоответствие условию задачи. Ведь мы *знаем*, что студент оказался отличником, т. е. событие $A$ произошло, зачем же мы ищем вероятность этого

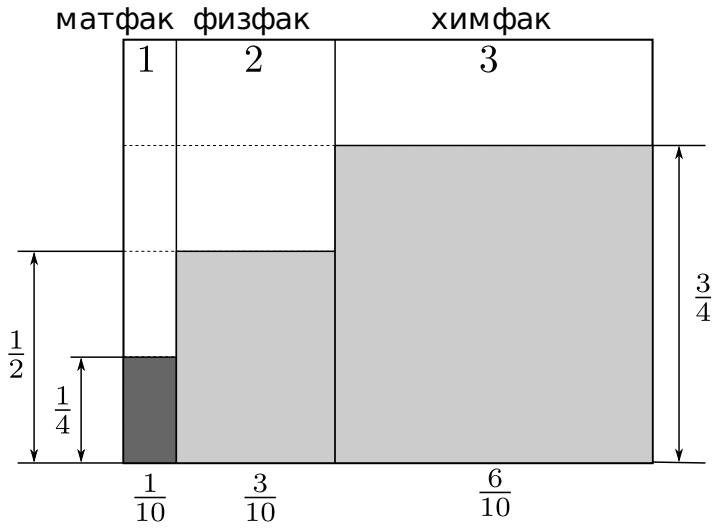


Рис. 2: Формула Байеса: вычисление условной вероятности $H_1$ при условии $A$ соответствует вычислению доли площади левого тёмного прямоугольника во всей закрашенной области

события? На самом деле, здесь $p(A)$ *априорная* вероятность события $A$, просто вероятность выбрать отличника, выбирая случайного студента.

Нас же сейчас интересует, какую долю составляет первое слагаемое $(p(A|H_1) \times p(H_1))$ в этой сумме — это и есть условная вероятность выбрать отличника с математического факультета, выбирая из всех отличников. Иными словами, нас интересует отношение

$$p(H_1|A) = \frac{p(A|H_1) \times p(H_1)}{p(A)} = \frac{p(A|H_1) \times p(H_1)}{p(A|H_1) \times p(H_1) + p(A|H_2) \times p(H_2) + p(A|H_3) \times p(H_3)}$$

Вот мы и получили *формулу Байеса* (в одной из своих записей). В рассмотренном примере интересующая нас вероятность есть отношение 25 отличников с математического факультета, ко всем 625 отличникам, то есть $25/625 = 0.04$ (или, переходя от студентов к долям от всего вуза, $0.025/0.625 = 0.04$).

Эти рассуждения также можно записать более формально. По определению условной вероятности

$$p(H_1|A) = \frac{p(H_1 A)}{p(A)}$$

Величину $p(A)$ мы нашли при помощи формулы полной вероятности $p(A) = p(H_1)p(A|H_1)+p(H_2)p(A|H_2)+p(H_3)p(A|H_3)$.

Осталось найти $p(H_1 A)$. Это вероятность того, что случайно выбранный студент окажется отличником с математического факультета. Согласно (4), $p(H_1 A) = p(A H_1) = p(A|H_1) \times p(H_1)$.

Итак, мы получили формулу Байеса:

$$p(H_1|A) = \frac{p(A|H_1) \times p(H_1)}{p(A)} = \frac{p(A|H_1)}{p(A|H_1) \times p(H_1) + p(A|H_2) \times p(H_2) + p(A|H_3) \times p(H_3)}$$

Она, разумеется, верна в случае любого количества гипотез:

$$p(H_i|A) = \frac{p(H_i) \times p(A|H_i)}{p(H_1)p(A|H_1) + p(H_2)p(A|H_2) + ... + p(H_n)p(A|H_n)} \tag{5}$$

**Примеры:**

1. На столе лежат три монетки: две обыкновенные и одна с орлами с двух сторон. Наудачу выбирается одна монетка и подбрасывается. На подброшенной монетке выпал орел. С какой вероятностью мы кидали обычную монетку?

Пусть

$H_1$ событие (гипотеза) — мы взяли обычную монетку,

$H_2$ событие (гипотеза) — мы взяли монетку с двумя орлами,

Гипотезы $H_1$ и $H_2$ образуют полную систему событий.

Событие $A$ — на монетке выпал орел.

Тогда

$p(H_1) = 2/3, \qquad p(A|H_1) = 1/2$

$p(H_2) = 1/3, \qquad p(A|H_2) = 1$

Значит,

$$p(A) = \frac{p(H_1)p(A|H_1)}{p(H_1)p(A|H_1) + p(H_2)p(A|H_2)} = \frac{\frac{2}{3} \times \frac{1}{2}}{\frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times 1} = \frac{1/3}{2/3} = \frac{1}{2}$$

Это и неудивительно: хотя на странной монетке орлы встречаются в два раза чаще, чем на обычной, зато на нашем столе обычные монетки встречаются в два раза чаще, чем странные.

2. Тест на ВИЧ выдает верный результат в 95% случаев. Известно, что ВИЧ заражена 1/1000 часть всего населения. Человек получил положительный результат теста на ВИЧ (то есть по результатам теста он заражен). С какой вероятностью он действительно заражен?

Пусть

$H_1$ событие (гипотеза) — человек заражен ВИЧ,

$H_2$ событие (гипотеза) — человек не заражен ВИЧ.

Гипотезы $H_1$ и $H_2$ образуют полную систему событий.

Событие $A$ — тест дал положительный результат.

Тогда

$p(H_1) = 0,001, \qquad p(A|H_1) = 0,95$

$p(H_2) = 0,999, \qquad p(A|H_2) = 0,05$

Значит,

$$p(H_1|A) = \frac{p(H_1)p(A|H_1)}{p(H_1)p(A|H_1) + p(H_2)p(A|H_2)} = \frac{0,001 \times 0,95}{0,001 \times 0,95 + 0,999 \times 0,05} \approx 0,0189$$

Результат кажется удивительным: ведь ошибка теста всего 5%, как же могло получиться, что вероятность заболевания при положительном результате теста так мала? Дело в том, что среди людей, получивших положительный результат теста, гораздо больше получивших ложный положительный результат (5% от 99,9% населения), чем людей, действительно зараженных ВИЧ (0,1% населения), а тем более, действительно больных, получивших положительный результат теста — см. рисунок 3.

Рис. 3: Применение формулы Байеса в случае теста на ВИЧ. Тонкая полоска слева — люди, зараженные ВИЧ. Темный закрашенный прямоугольник — те из них, кто получил положительный результат теста. Длинный закрашенный прямоугольник справа — здоровые люди, получившие положительный результат теста. Доля больных людей мала по сравнению со всеми людьми, получившими положительный результат.