ПРАКТИКА ДЛЯ ПРАКТИКОВ

О статистической обработке результатов педагогических измерений

М.Е. Бершадский

В данной статье я хотел бы обратить внимание читателей на необходимость статистической обработки результатов педагогических измерений, которые выполняются при мониторинге учебного процесса при оценке учебных достижений учащихся, а также сдвигов в их интеллектуальном и физическом развитии, при оценке эффективности педагогических инноваций.

К сожалению, работники образования привыкли сравнивать результаты диагностики, опираясь только на сравнение числа «неудов», «троек», «четвёрок» и «пятёрок», полученных школьниками, не догадываясь, что для оценки результатов массовых измерений, выполненных на основе нечётких критериев отнесения испытуемых к различным классам, необходимо применять статистические методы сравнения.

На результаты измерения тех или иных свойств учащихся и их учебных достижений влияет множество разнообразных факторов: время тестирования и его длительность; личность учителя; освещённость и обстановка кабинета; погодные условия; степень утомления учеников и их физическое состояние; валидность теста; значимость его результатов для будущего учеников и многое другое. Оценив работы учащихся количественным баллом, а затем сравнивая достижения учащихся с их предыдущими показателями или с результатами других классов, учитель часто забывает, что он пользовался весьма размытыми критериями выставления оценок и часто долго размышлял, какой оценки достоин ученик, — «три с плюсом» или «четыре с минусом». На самом деле, как и при любом другом измерении, результаты педагогических измерений всегда выполняются с погрешностями, поэтому при сопоставлении их данных нужно устанавливать степень достоверности сходства или различия показателей, описывающих те или иные характеристики учащихся. Эта задача и решается с помощью методов математической статистики.

Предположим, что читатель решил применить одну из приглянувшихся ему образовательных технологий и хочет убедиться в том, что она действительно позволяет достичь декларируемых целей. Для этого необходимо сравнить результаты учащихся до и после применения технологии. Поскольку прирост результатов может быть вызван и другими причинами, то для корректного обоснования выводов о результативности технологии необходимо выбрать ещё и контрольную группу учащихся (группа, в которой применяется технология, называется экспериментальной).

На первом этапе проводится начальная диагностика в обеих группах для измерения значения того признака, который должен измениться под воздействием технологии. Если технология предназначена для изменения уровня признака (уровень развития интеллекта, активность, измеряемая количеством ответов, количество решённых задач, рост и вес учащихся и т. д.), то нужно сравнить его значения в обеих группах с помощью соответствующего статистического критерия, каждый из которых имеет свою область применения и ограничения, например, с помощью Q-критерия Розенбаума, описание которого приведено в этой статье.

Дальнейшие действия зависят от результата сравнения. Если группы статистически не отличаются по уровню признака, то в экспериментальном классе можно начинать применять образовательную технологию, одновременно придерживаясь традиционной системы в контрольном классе и обеспечивая одинаковые условия обучения. Через определённый интервал времени, зависящий от предполагаемой длительности достижения целей, проводится повторное измерение значений признака в обеих группах и сравнение его уровней с помощью того же статистического критерия. Если в экспериментальном классе он окажется выше, чем в контрольном, то тогда можно утверждать, что его прирост объясняется именно применённой технологией обучения.

Впрочем, существование прироста нужно специально доказывать. Очень часто на практике педагоги оценивают его «на глазок», прямо сравнивая значения измеряемого признака. Если более высоких значений стало больше, то делается вывод, что уровень признака возрос. На самом деле задача сравнения уровня признака в одном и том же классе в разное время тоже должна решаться с применением статистических методов. Это можно сделать, пользуясь, например, Т-критерием Вилкоксона.

Таким образом, обоснованные заключения о динамике результатов можно делать только с помощью статистических методов. Однако в педагогической среде бытует мнение об их чрезмерной сложности. Попробую развенчать это представление, подробно описав процедуру применения уже упоминавшегося выше Q-критерия Розенбаума.

Этот критерий применяется для сравнения данных педагогических измерений в двух выборках (группах уча-

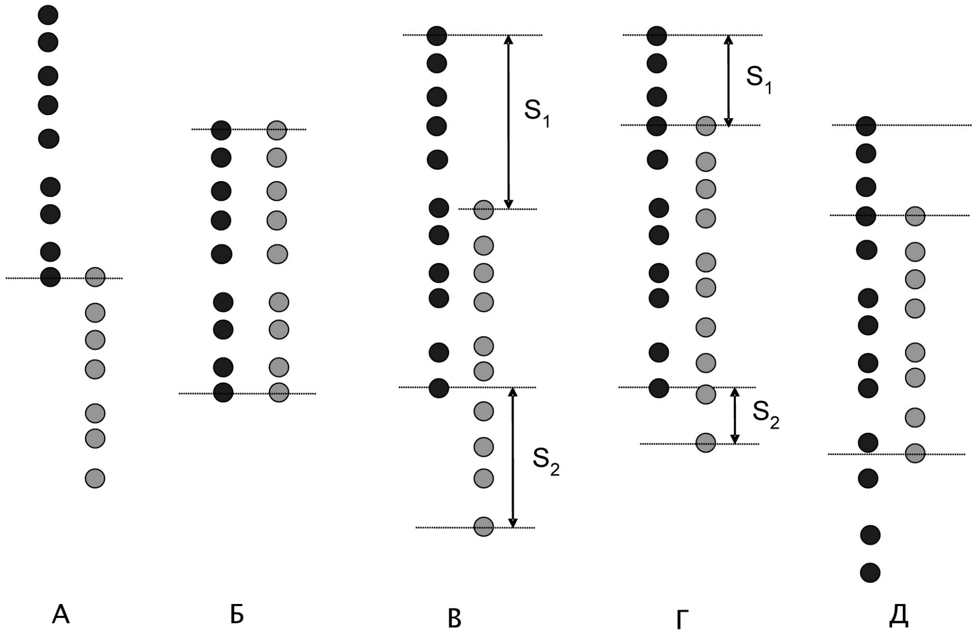

Рис. 1. Распределение значений признака

щихся), в результате которых можно упорядочить учащихся по величине количественно измеряемого признака. Например, можно упорядочить учащихся по величине коэффициента интеллекта (КИ), времени бега на какую-либо дистанцию, ростовым и весовым показателям и т. д. Этот метод — один из наиболее простых и наглядных способов выявления достоверных различий между уровнями измеряемого признака в двух группах учащихся.

Поясним суть Q-критерия Розенбаума с помощью графического представления. На рисунке «Распределение значений признака» изображены возможные соотношения между диапазонами значений признаков в двух группах. Каждый из диапазонов представлен вертикальным рядом точек. Верхняя точка соответствует максимальному значению признака в группе, нижняя — минимальному. Так как диапазон значения признака в двух группах может не совпадать, то отрезки могут быть сдвинуты по высоте. Если они не перекрывают друг друга (положение А на рисунке), т. е. минимальное значение признака в первой группе больше максимального значения во второй, то различия в уровне признака достоверны и Q-критерий Розенбаума применять не нужно. Если отрезки совпадают (положение Б на рисунке), то уровни признака совпадают, поэтому Q-критерием Розенбаума пользоваться не имеет смысла. Необходимость в его применении возникает тогда, когда соотношение между диапазонами признака принимает вид, соответствующий положению В.

Через S_1 обозначим разность между числом учащихся, характеризующихся наибольшим значениями признака в первой группе, и числом учащихся, имеющих наибольшее значение признака, во второй группе. Через S_2 — разность между числом учащихся первой группы, имеющих наименьшее значение признака в этой группе, и числом учеников во второй группе, характеризующихся минимальным значением признака в своей группе. Например, в одном классе самый высокий мальчик имеет рост 172 см, а во втором классе трое подростков имеют рост выше 172 см. Тогда $S_1 = 3$. Пусть самый маленький рост ребёнка в первом классе равен 164 см, а во втором двое мальчиков имеют рост, превышающий эти данные. Тогда $S_2 = 2$.

Если область перекрывания отрезков уменьшить (положение Г на рисунке), т. е. сблизить показания учащихся в двух группах, то значения S_1 и S_2 уменьшатся, поэтому уменьшится и их сумма S. Таким образом, величина S зависит от степени расхождения значений признака в сравниваемых группах. Чем ближе результаты, тем меньше S. На вычислении S и сравнении его с некоторым критическим значением, вычисленным заранее с помощью методов статистики (для применения метода в практической деятельности разбираться в способах вычисления критических значений различных критериев не обязательно) и основано применение Q-критерия Розенбаума. Вычисленные критические значения критерия заносятся в специальные таблицы (см. таблицы 1 и 2).

Таблица 1

Критические значения Q-критерия Розенбаума для уровня значимости 0,05

n	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
$\rho = 0,05$																
11	6															
12	6	6														
13	6	6	6													
14	7	7	6	6												
15	7	7	6	6	6											
16	8	7	7	7	6	6										
17	7	7	7	7	7	7	7									
18	7	7	7	7	7	7	7	7								
19	7	7	7	7	7	7	7	7	7							
20	7	7	7	7	7	7	7	7	7	7						
21	8	7	7	7	7	7	7	7	7	7	7					
22	8	7	7	7	7	7	7	7	7	7	7	7				
23	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	8	8	8	8	8	8	8	8	8	8	7	7	7	7		
25	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7	
26	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7	7

Их величины зависят от числа учащихся в группах (n_1 и n_2) и от уровня значимости *p*, который характеризует вероятность ошибочного вывода на основе сравниваемых данных. В педагогике традиционно выделяют два уровня значимости — $p = 0,05$ и $p = 0,01$ (впрочем, более жёсткие требования на уровне значимости 0,01 встречаются редко; их обычно применяют в психологии). Если уровень значимости $p \leq 0,05$, то вероятность ошибочного вывода из результатов сравнения не превышает 5% (надёжность вывода составляет не менее 95%). При уровне значимости $p \leq 0,01$ вероятность ошибки не превышает 1%, или надёжность вывода больше 99%.

При выборе Q-критерия Розенбаума для сравнения данных нужно быть осторожным (это справедливо и по отношению ко всем статистическим критериям), так как существуют несколько ограничений на его применение.

1. Необходимо, чтобы данный показатель мог принимать широкий спектр значений, иначе сравнение окажется бессмысленным. Например, нельзя пользоваться данным критерием для сравнения числа оценок по обычной пятибалльной шкале. Чем больше разброс численных значений, тем более достоверными оказываются выводы о совпадении или различии значений признака в сравниваемых группах.

2. Q-критерием Розенбаума можно пользоваться, если число испытуемых в каждой группе не меньше одиннадцати. Кроме того, объёмы

Таблица 2

Критические значения Q-критерия Розенбаума для уровня значимости 0,01

n	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
ρ = 0,01																
11	9															
12	9	9														
13	9	9	9													
14	9	9	9	9												
15	9	9	9	9	9											
16	9	9	9	9	9	9										
17	10	9	9	9	9	9	9									
18	10	10	9	9	9	9	9	9								
19	10	10	10	9	9	9	9	9	9							
20	10	10	10	10	9	9	9	9	9	9						
21	11	10	10	10	9	9	9	9	9	9	9					
22	11	11	10	10	10	9	9	9	9	9	9	9				
23	11	11	10	10	10	10	9	9	9	9	9	9	9			
24	12	11	11	10	10	10	10	9	9	9	9	9	9	9		
25	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9	
26	12	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9

выборок (количество учащихся в группах) не должны сильно отличаться друг от друга. Можно воспользоваться тремя правилами, которые позволяют определить возможность применения Q-критерия Розенбаума[1]:

· если в каждой из выборок менее 50 учащихся, то разность между числом учащихся в группах не должна быть меньше десяти;

· если в каждой из выборок более 51 учащегося, но меньше 100, то разность между числом учащихся в группах не должна быть больше 20;

· если в каждой из выборок более 100 учащихся, то одна из выборок может быть больше другой не более чем в полтора-два раза.

3. Диапазон значений, которые принимает признак в одной группе, не должен перекрывать всего множества значений, принимаемых им во второй группе.

Воспользуемся ещё раз графическим представлением критерия и изобразим диапазоны значений признака в двух группах, соответствующие описанной ситуации (см. положение Д на рисунке). Видно, что наибольшее значение признака в первой группе, *больше* максимального значения во второй. Одновременно с этим минимальное значение в первой группе *меньше* наименьшего значения во второй. При таком соотношении диапазонов значений признака в двух группах Q-критерием Розенбаума пользоваться нельзя.

Продолжим описание процедуры применения Q-критерия Розенбаума. Целью применения этого критерия является сравнения уровня величины измеряемого признака в двух группах, Эти величины могут либо совпадать, либо отличаться. Поэтому мы можем выдвинуть только два предположения, называемых статистическими гипотезами, которые соответствуют этим двум возможностям. Статистическая гипотеза об отсутствии различий в уровне признака называется нулевой, так как при равенстве значений их разность равна нулю. Обозначается нулевая гипотеза символом H_0. Гипотеза о значимости различий между уровнем признака называется альтернативной. Она обозначается H_1. Общепринятая формулировка гипотез для Q-критерия Розенбаума выглядит следующим образом.

H_0: Уровень признака в выборке 1 не превышает уровня признака в выборке 2.

H_1: Уровень признака в выборке 1 превышает уровень признака в выборке 2.

Вместо слов «уровень признака» в гипотезах указывается названия конкретного признака (уровень интеллекта, скорость реакции, количество ответов и т. д.).

Для проверки гипотез необходимо сопоставить экспериментальные данные и определить S_1 и S_2. Зная эти числа, следует вычислить эмпирическое значение Q-критерия Розенбаума: $Q_{эмп} = S_1 + S_2$. Для принятия решения полученное эмпирическое значение критерия нужно сравнить с критическими значениями критерия для уровней значимости 0,05 и 0,01, которые определяются по соответствующим таблицам 1 и 2. Для наглядности это лучше сделать в графичес-

[1] *Гублер Е.В.* Вычислительные методы анализа и распознавания патологических последствий. Л.: Медицина, 1978. С. 75.

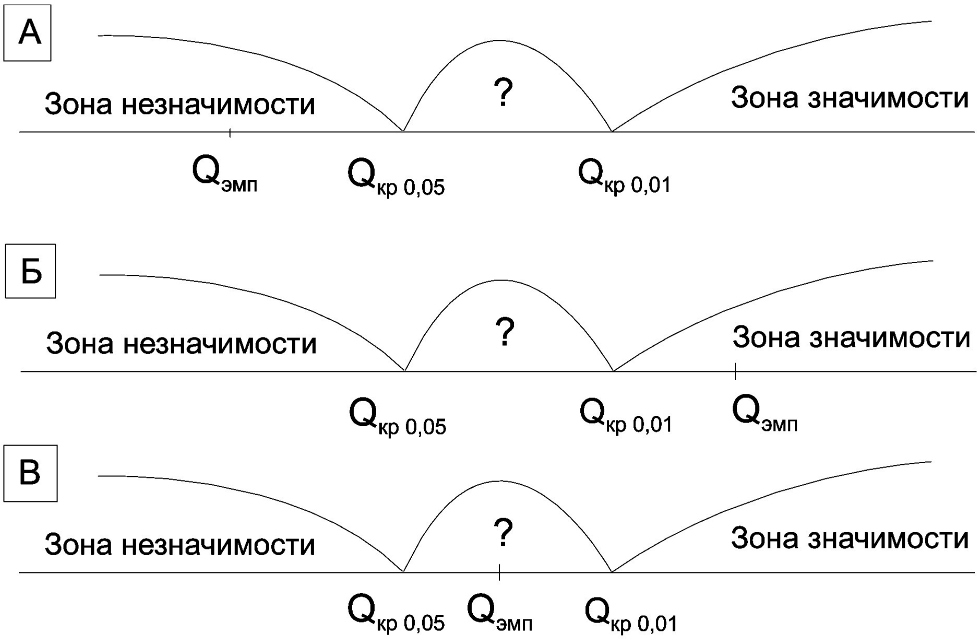

Рис. 2. Графическая интерпретация принятия решения

кой форме (см. рисунок «Графическая интерпретация принятия решения»).

На числовой оси, которую принято называть осью значимости, откладываются два критических значения критерия $Q_{кр0,05}$ и $Q_{кр0,01}$, и эмпирическое значение критерия Qэмп. Точки, соответствующие критическим значениям, делят числовую ось на три интервала. Слева от значения $Q_{кр0,05}$ располагается зона незначимости. Если эмпирическое значение критерия Qэмп попадает в эту область, то различия в уровне признака между сравниваемыми группами признаются несущественными, т. е. принимается нулевая гипотеза H_0 (пример А на рисунке 2). Справа от значения $Q_{кр0,01}$ находится область значимости. Если эмпирическое значение критерия Qэмп попадает в неё, то различия в уровне признака между сравниваемыми группами признаются существенными и принимается альтернативная гипотеза H_1 (пример Б на рисунке 2). Строго говоря, оба этих вывода являются вероятностными. В первом случае можно утверждать, что с вероятностью более 95% учащиеся в группе 1 не превышают учащихся в группе 2 по уровню сравниваемого признака. При попадании в зону значимости можно сделать вывод, что с вероятностью более 99% учащиеся в группе 1 превышают учащихся в группе 2 по уровню сравниваемого признака.

Если эмпирическое значение критерия находится между критическими значениями ($Q_{кр0,05} < Q_{эмп} < Q_{кр0,01}$), то мы попадаем в зону неопределённости (пример В на рисунке 2). Уже можно отбросить нулевую гипотезу о равенстве уровня признаков, но принимать альтернативную гипотезу ещё рано, так как вероятность ошибочного решения достигает 5%. Впрочем, как уже было отмечено выше, в педагогических измерениях этой ошибкой пренебрегают и принимают альтернативную гипотезу и при данном соотношении между критическими и эмпирическим критериями. Учитывая, что Q-критерий Розенбаума имеет не слишком высокую мощность, я не советую читателю поступать таким образом. Более надёжно принимать решение в пользу альтернативной гипотезы только в том случае, если $Q_{эмп} > Q_{кр0,01}$.

В качестве примера рассмотрим задачу на сравнение уровня интеллектуального развития ученики двух восьмых классов. Первичные результаты измерения коэффициента интеллекта (КИ) с помощью теста Амтхауэра представлены в таблице 3.

На первый взгляд кажется, что результаты учащихся 8 «А» класса чуть выше, чем во втором классе параллели. Попробуем проверить это предположение.

Сформулируем гипотезы о возможном соотношении между уровнями КИ в первой и во второй группах. Существуют только две альтернативы. Либо уровень КИ в 8 «А» классе действительно выше, чем в 8 «Б», либо статистически они не отличаются. В данном случае под уровнем признака понимается уровень развития интеллекта учащихся. Поэтому нулевая и альтернативная гипотезы выглядят следующим образом.

H_0: Учащиеся 8 «А» класса не превосходят учащихся 8 «Б» класса по уровню интеллекта.

H_1: Учащиеся 8 «А» класса превосходят учащихся 8 «Б» класса по уровню интеллекта.

На следующем шаге для сопоставления данных нужно занести их в новую таблицу, в которой значения КИ располагаются не произвольно, как в первой таблице, а в порядке возрастания, едином для обеих сравниваемых групп. Результат этого действия изображён в таблице 4.

По таблице 4 определяем число учащихся в 8 «А» классе, имеющих КИ выше наибольшего значения в 8 «Б» классе. Обозначим это число S_1. Затем определяем по таблице количество учащихся в 8 «Б» классе, имеющих КИ ниже наименьшего значения в 8 «А» классе. Обозначим это число S_2. Для данного случая $S_1 = 4$, а $S_2 = 6$. На самом деле составлять эту таблицу не обязательно. Можно воспользоваться приёмом цветных карточек, вырезанных из бумаги. Заготовим 24 карточек, например синего цвета, для учащихся 8 «А» и напишем на них значения КИ учеников этого класса. Для учащихся 8 «Б» заготовим 23 карточки любого другого цвета, например зелёного. На каждой из них запишем КИ одного из учеников этого класса. Затем, не обращая внимания на цвет, разложим все карточки в порядке возрастания КИ. Тогда количество синих карточек, лежащих выше последней зелёной, даст нам число S_1, а число зелёных карточек, расположенных ниже последней синей — число S_2.

Вернёмся к обработке данных примера. Вычисляем эмпирическое значение Q-критерия Розенбаума по

Таблица 3

Первичные данные по измерению КИ

Ученики 8 «А»			Ученики 8 «Б»		
№	Код имени испытуемого	Коэффициент интеллекта (КИ)	№	Код имени испытуемого	Коэффициент интеллекта (КИ)
1	Алексей М.	108	1	Алла А.	81
2	Виктор С.	93	2	Елизавета Б.	83
3	Тамара К.	107	3	Светлана К.	102
4	Анна Л.	86	4	Мария Г.	106
5	Елена П.	90	5	Александр К.	109
6	Александр Т.	106	6	Алла С.	85
7	Тимофей Л.	120	7	Татьяна М.	84
8	Андрей В.	104	8	Максим С.	98
9	Людмила С.	112	9	Вера Т.	93
10	Ирина К.	87	10	Мария Н.	97
11	Кирилл В.	91	11	Андрей М.	87
12	Дмитрий Б.	126	12	Николай К.	120
13	Дарья В.	95	13	Антон С.	114
14	Ксения Р.	98	14	Анатолий М.	82
15	Иван Н.	114	15	Вадим З.	90
16	Ангелина Г.	123	16	Евгений Б.	92
17	Евгения Ф.	127	17	Никита К.	86
18	Илья В.	89	18	Артём К.	105
19	Анна Д.	97	19	Сергей П.	89
20	Андрей К.	99	20	Денис В.	101
21	Ольга Т.	128	21	Владимир Д.	94
22	Анастасия Б.	92	22	Вячеслав Д.	103
23	Дмитрий Ш.	96	23	Кирилл Н.	80
24	Алла С.	94			

формуле $Q_{эмп} = S_1 + S_2 = 4 + 6 = 10$. По таблицам 1 и 2 определяем критические значения критерия при $n_1 = 24$ (число учащихся 8 «А» класса), $n_2 = 23$ (число учащихся 8 «Б» класса), соответствующих уровням значимости 0,05 и 0,01: $Q_{кр0,05} = 7$, $Q_{кр0,01} = 9$. Изобразим соотношение между эмпирическим и критическими значениями на рисунке с помощью оси значимости (см. рисунок «Графическая иллюстрация принятия решения»).

В данном примере $Q_{эмп} > Q_{кр0,01}$, поэтому принимается альтернативная гипотеза, т. е. с вероятностью более 99% можно утверждать, что учащиеся 8 «А» класса превосходят учащихся 8 «Б» класса по уровню интеллекта.

Таблица 4

Упорядоченные данные по измерению КИ

Ученики 8 «А»			Ученики 8 «Б»		
№	**Код имени испытуемого**	**Коэффициент интеллекта (КИ)**	**№**	**Код имени испытуемого**	**Коэффициент интеллекта (КИ)**
1	Ольга Т.	128			
2	Евгения Ф.	127			
3	Дмитрий Б.	126			
4	Ангелина Г.	123			
5	Тимофей Л.	120	1	Николай К.	120
6	Иван Н.	114	2	Антон С.	114
7	Людмила С.	112			
			3	Александр К.	109
8	Алексей М.	108			
9	Тамара К.	107			
10	Александр Т.	106	4	Мария Г.	106
			5	Артём К.	105
11	Андрей В.	104			
			6	Вячеслав Д.	103
			7	Светлана К.	102
			8	Денис В.	101
12	Андрей К.	99			
13	Ксения Р.	98	9	Максим С.	98
14	Анна Д.	97	10	Мария Н.	97
15	Дмитрий Ш.	96			
16	Дарья В.	95			
17	Алла С.	94	11	Владимир Д.	94
18	Виктор С.	93	12	Вера Т.	93
19	Анастасия Б.	92	13	Евгений Б.	92
20	Дмитрий Б.	91			
21	Елена П.	90	14	Вадим З.	90
22	Илья В.	89	15	Сергей П.	89
23	Ирина К.	87	16	Андрей М.	87
24	Анна Л.	86	17	Никита К.	86
			18	Алла С.	85
			19	Татьяна М.	84
			20	Елизавета Б.	83
			21	Анатолий М.	82
			22	Алла А.	81
			23	Кирилл Н.	80

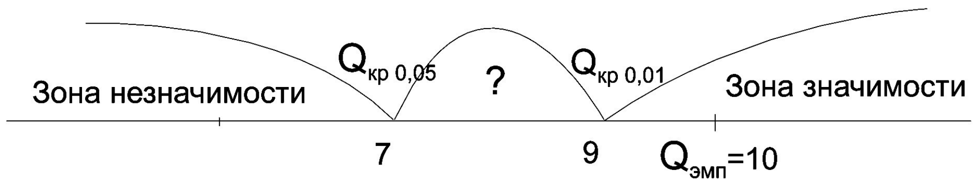

Рис. 3. Графическая иллюстрация принятия решения

Подчеркнём, что Q-критерий Розенбаума позволяет сравнить только две группы учащихся по уровню признака. Для сравнения большего числа групп необходимо воспользоваться другими статистическими критериями.

Надеюсь, что данный образец применения Q-критерия Розенбаума поможет читателю воспользоваться им в практической деятельности для оценки результативности педагогических инноваций.

Для получения дополнительной информации об использовании статистических методов рекомендую читателю познакомиться с литературой по их применению в педагогике и психологии[2].

[2] *Сидоренко Е.В.* Методы математической обработки в психологии. Спб.: ООО «Речь», 2000; Гласс Дж., Стенли Дж. Статистические методы в педагогике / Пер. с англ. М.: Прогресс, 1976.