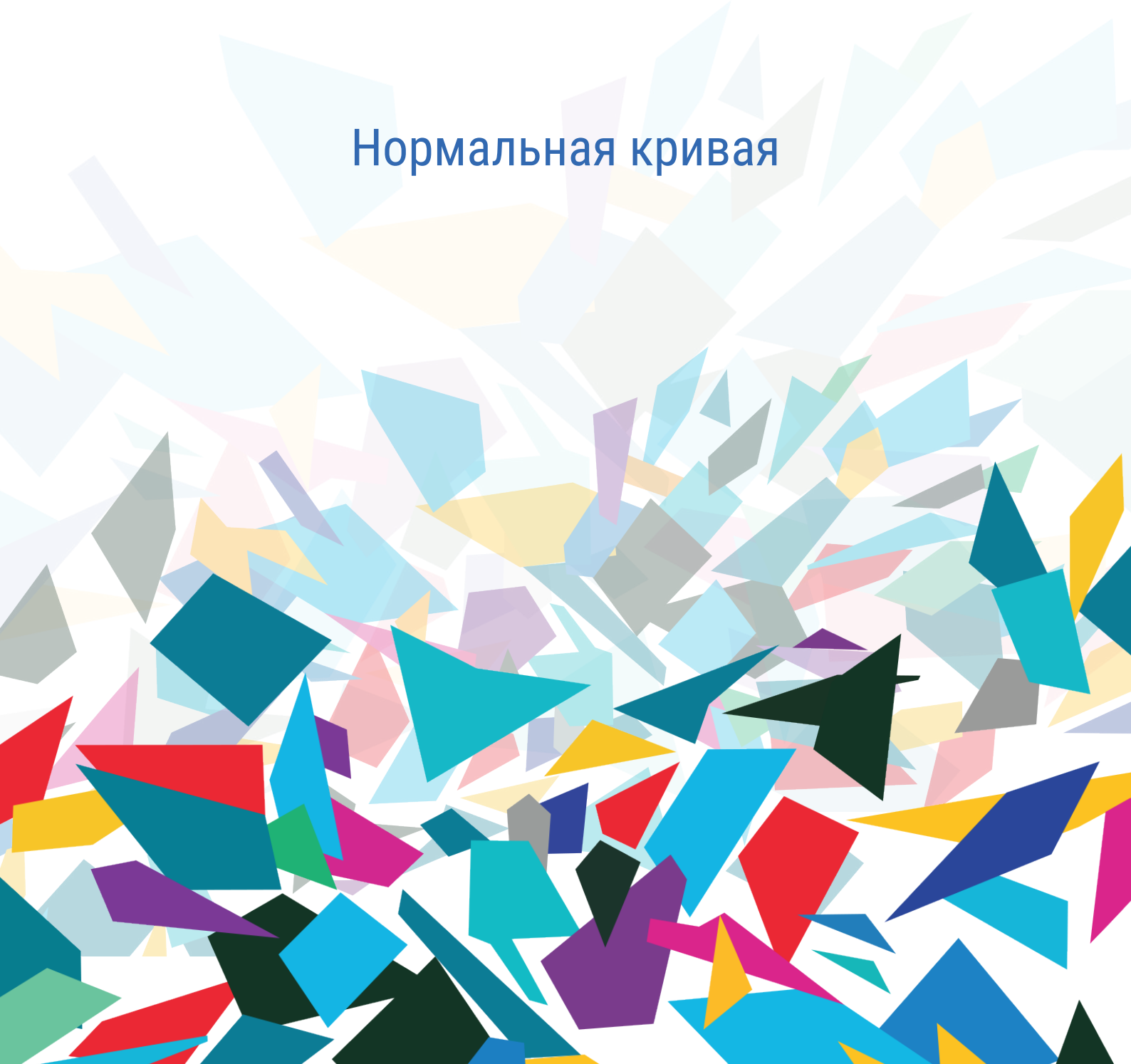
Лекция 6

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Нормальная кривая



Здравствуйте!

В результате изучения этой лекции вы будете знать и уметь:
1. Определять и объяснять концепцию нормальной кривой.
2. Преобразовать эмпирические оценки в Z-оценки и использовать Z-оценки и таблицу нормальных кривых для получения необходимой информации.
3. Выражать область кривой в терминах вероятностей.

Нормальная кривая – понятие, имеющее большое значение в статистике. В сочетании со средним и стандартным отклонением он используется для точных описательных утверждений об эмпирических распределениях. Кроме того, нормальная кривая является центральной в теории, которая лежит в основе статистики. Нормальная кривая иногда используется для оценки оценок на тестах, и вы возможно знакомы с этим приложением. В этом виде нормальную кривую часто называют «кривой колокола».

Свойства нормальной кривой

Кривая представляет собой теоретическую модель, линейный график, который является унимодальной (то есть имеет один режим или пик), идеально гладкий и симметричный (не скошенный), поэтому ее среднее значение, медиана и мода все точно такое же значение. Это колоколообразный, и его хвосты простираются бесконечно в обоих направлениях. Конечно, ни одно эмпирическое распределение не соответствует этой идеальной модели идеально, но некоторые переменные (например, результаты тестов для больших классов, стандартизированные результаты тестов) достаточно близки, чтобы допустить предположение о нормальности. В свою очередь, это предположение делает возможным одно из наиболее важных применений нормальной кривой – описание эмпирических распределений, основанное на наших знаниях теоретической нормальной кривой. Критическим моментом в отношении нормальной кривой является то, что расстояния вдоль горизонтальной оси, измеряемые в стандартных отклонениях от среднего, всегда охватывают одинаковую долю общей площади под кривой. Другими словами, расстояние от любой точки до среднего – при измерении в стандартных отклонениях – будет отрезать точно такую же пропорциональную часть площади под кривой.
Чтобы проиллюстрировать это, на рисунках 6.1 и 6.2 представлены два гипотетических распределения баллов IQ, оба нормально распределенных, для вымышленных групп мужчин и женщин:

Мужчины	Женщины
$\overline{X} = 100$ S=20 N=1000	$\overline{X} = 100$ S=10 N=1000

так что рисунки 6.1 и 6.2 изображены с двумя шкалами на горизонтальной оси графика. Верхняя шкала указывается в «единицах IQ», а нижняя - в стандартных отклонениях от среднего. Эти шкалы являются взаимозаменяемыми, и мы можем легко переходить от одного к другому. Например, для мужчин показатель IQ равен 120 это одно стандартное отклонение (помните, что для группы спаривания s – 20) выше среднего значения, а IQ 140 - на два стандартных отклонения выше среднего значения (справа). Оценки IQ слева от среднего значения отмечены

как отрицательные значения на шкале стандартных отклонений, потому что они меньше среднего. IQ 80 на одно стандартное отклонение ниже среднего, IQ 60 – на два стандартных отклонения меньше среднего и т. д.

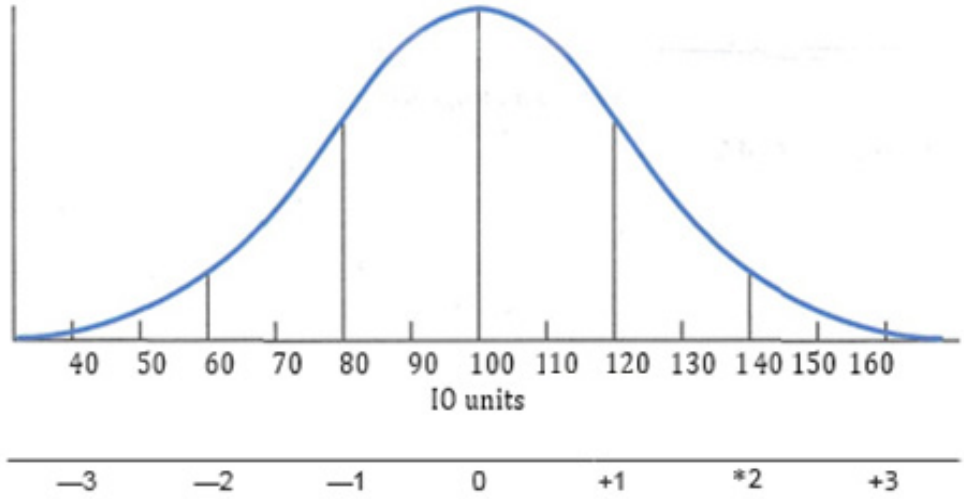


Рисунок 6.1. IQ результаты для группы мужчин.

Рисунок 6.2 обозначен аналогичным образом, за исключением того, что его стандартное- это другое значение s – = 10), маркировка происходит в разных точках. Для выборки женского пола одно стандартное отклонение выше среднего равно IQ 110, одно стандартное отклонение ниже среднего равно lQ 90 и т. д. Напомним, что на любой нормальной кривой расстояния вдоль горизонтальной оси при измерении в стандартных отклонениях всегда охватывают точно одинаковую долю общей площади под кривой.

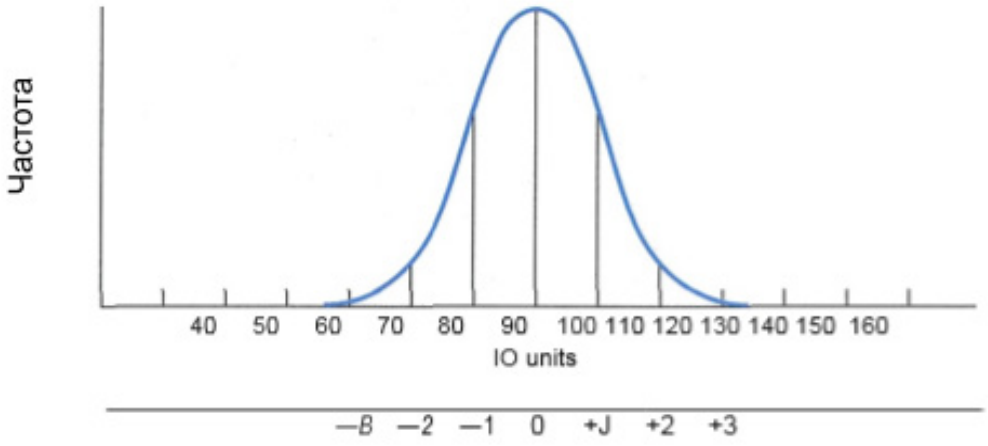


Рисунок 6.2 IQ результаты для группы женщин.

Рисунок 6.2 обозначен аналогичным образом, за исключением того, что его стандартное- это другое значение s – = 10), маркировка происходит в разных точках. Для выборки женского пола одно стандартное отклонение выше среднего равно IQ 110, одно стандартное отклонение ниже среднего равно lQ 90 и т. д. Напомним, что на любой нормальной кривой расстояния вдоль горизонтальной оси при измерении в стандартных отклонениях всегда охватывают точно одинаковую долю общей площади под кривой.

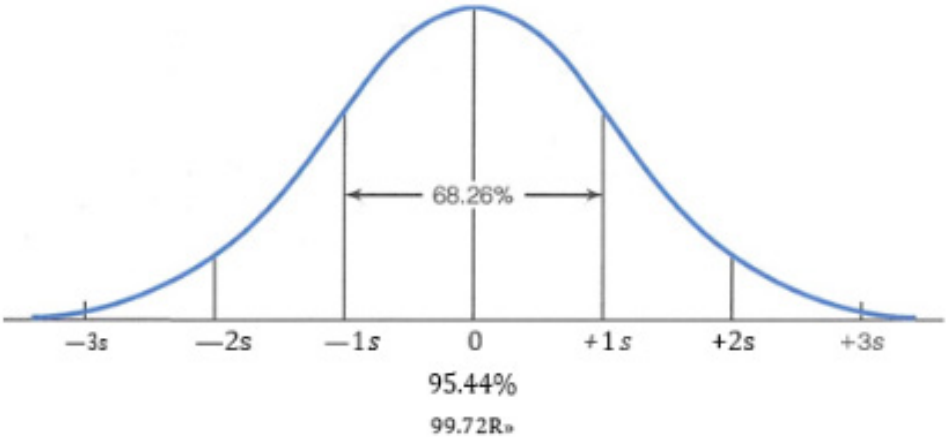


Рисунок 6.3 Области под нормальной кривой.

Ознакомимся со следующими отношениями между расстояниями от среднего значения и областями под нормальной кривой:

Между	Находится
±1 среднеквадратичное отклонение	68.26% области
± 2 среднеквадратичных отклонения	95.44% области
± 3 среднеквадратичных отклонения	99.72% области

Эти отношения отображаются графически на рисунке 6.3.

Рассмотрим эмпирические распределения, которые по крайней мере приблизительно нормальны, используя эти отношения между расстоянием от среднего значения и площадью. Положение отдельных баллов можно описать относительно среднего значения, распределения в целом или любого другого балла в распределении.

Области между баллами также могут быть выражены в количестве случаев, а не в процентах от общей площади. Например, нормальное распределение в 1000 случаев будет содержать около 683 случаев (68.267%с из 1000 случаев) между ±1 стандартным отклонением от среднего, около 954 между ±2 стандартными отклонениями и около 997 между стандартными отклонениями

±3. Таким образом, для любого нормального распределения только несколько случаев будут отличаться от среднего значенияна ±3.

Использование нормальной кривой

Мы видели, что мы можем найти области под нормальной кривой для баллов, которые в точности равны 1, 2 или 3 стандартным отклонениям выше или ниже среднего. Чтобы работать со значениями, которые не являются точными кратными стандартного отклонения, мы должны выразить исходные оценки в единицах стандартного отклонения или преобразовать их в Z оценки. Исходные оценки могут быть в любой единице измерения (футы, IQ, доллары), но Z-оценки всегда имеют одинаковые значения для их среднего значения (0) и стандартного отклонения (1).

Вычисление Z оценок

Чтобы преобразовать исходные оценки в Z-оценки как процесс изменения шкал значений, котороый строиться аналогично изменению от метров к ярдам, километров к милям или галлонов к литрам. Эти единицы измерения различны, но одинаково действительны для выражения расстояния, длины или объема. Например, миля равна 1,61 километра, поэтому два города, которые находятся на расстоянии 10 миль друг от друга, также находятся на расстоянии 16,1 километра, а гонка «5 тысяч километров» покрывает около 3,10 миль. Таким же образом, исходные (или «сырые») оценки и Z-оценки являются двумя одинаково действительными, но разными способами измерения расстояний под нормальной кривой. На рисунке 6.1, например, мы могли бы описать конкретную оценку в единицах IQ («оценка Армана была 120») или стандартных отклонений («Арман набрал одно стандартное отклонение выше среднего»).

Когда мы вычисляем Z-оценки, мы конвертируем исходные единицы измерения (IQ-показатели, дюймы, доллары и т. Д.) В Z-оценки и, таким образом, «стандартизируем» нормальную кривую для распределения со средним значением 0 и стандартным отклонением. из 1.

Среднее эмпирическое нормальное распределение будет преобразовано в 0, его стандартное отклонение в 1, и все значения будут выражены через Z-оценки:

$$Z = \frac{X_i - \overline{X}}{S}$$

Эта формула преобразует любую оценку (Xi) из эмпирического нормального распределения в эквивалентную оценку Z. Рассмотрим на примере изображенного на рисунке 6.1, где

$$Z = \frac{120-100}{20} = \frac{20}{20} = \pm 1.00$$

Z оценка положительная 1,00 указывает на то, что исходный балл находится на единицу стандартного отклонения выше (справа от) среднего. Отрицательный результат упал бы ниже (слева) от среднего значения.
Теоретическая нормальная кривая была очень подробно описана статистиками. Области, связанные с любой Z-оценкой, были точно определены и организованы в виде таблицы. Эта

таблица нормальных кривых или таблица Z-показателей представлена как Приложение А.

Таблица нормальных кривых состоит из трех столбцов, с оценками Z в левом столбце (столбец a), областями между показателем z и средним значением в середине (столбец b) и областями за пределами показателя Z в правой части столбец (столбец c). Чтобы найти область между любой оценкой Z и средним значением, спускаемся вниз по столбцу Z-оценки, пока не найдем оценку. Например, переходите вниз к столбцу a либо в Приложении А, либо в Таблице 6.1, пока не найдем Z балл + 1,00. Запись в столбце b показывает, что «Площадь между средним и z» равна 0,3413.

В таблице представлены области в виде пропорций, но мы можем легко перевести их в проценты, умножив их на 100. Мы могли бы сказать, что «доля в 0,3413 от общей площади под кривой лежит между Z-показателем 1,00 и средним» или «34,137c от общей площади находится между Z-показателем 1,00 и средним».

Таблица 6.1 Определение области нормальной кривой (Приложение А).

(a) Показатель Z	(b) Областями между показателем Z и средним значением	(c) Областями за пределами показателя Z
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
.	.	.
.	.	.
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
.	.	.
.	.	.
.	.	.
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
.	.	.
.	.	.

Нахождение общей площади выше и ниже оценки

До сих пор мы использовали таблицу нормальных кривых, чтобы найти области между оценкой Z и средним значением. Приложение А также можно использовать для поиска других типов областей в эмпирических распределениях, которые по крайней мере приблизительно имеют нормальную форму. Например, предположим, что нам нужно определить общую площадь ниже баллов двух субъектов мужского пола в распределении, описанном на рисунке 6.1.

Рассматриваемый субъект имеет оценку 117 (X1 = 117),
что эквивалентно оценке Z равной + 0,85

$$Z = \frac{X_i - \overline{X}}{s} = \frac{117-100}{20} = \frac{17}{20} = +0.85$$

Знак «плюс» указывает на то, что оценка должна быть расположена выше (справа от) среднего значения. Чтобы найти область ниже положительной оценки Z, область между оценкой и средним значением (см. Столбец b) должна быть добавлена к области ниже среднего. Как мы уже отмечали ранее, нормальная кривая симметрична (не скошена), и ее среднее значение будет равно ее медиане. Следовательно, площадь ниже среднего (как и в среднем) будет 50%. На рисунок 6.4. нас интересует заштрихованная область. Изучив таблицу нормальных кривых, мы находим, что область между оценкой и средним значением (см. Столбец b) составляет 30,23% с общей площади. Следовательно, область ниже Z-балла +0,85 составляет 80,23% с (50,00%+ 30,23%). Этот значит, что данный балл набрали более 80,23% опрошенных.

Второй субъект имеет оценку IQ 73 (X2 = 73),
что эквивалентно оценке Z - 1,35

$$Z = \frac{X_i - \overline{X}}{s} = \frac{73-100}{20} = -\frac{27}{20} = -1,35$$

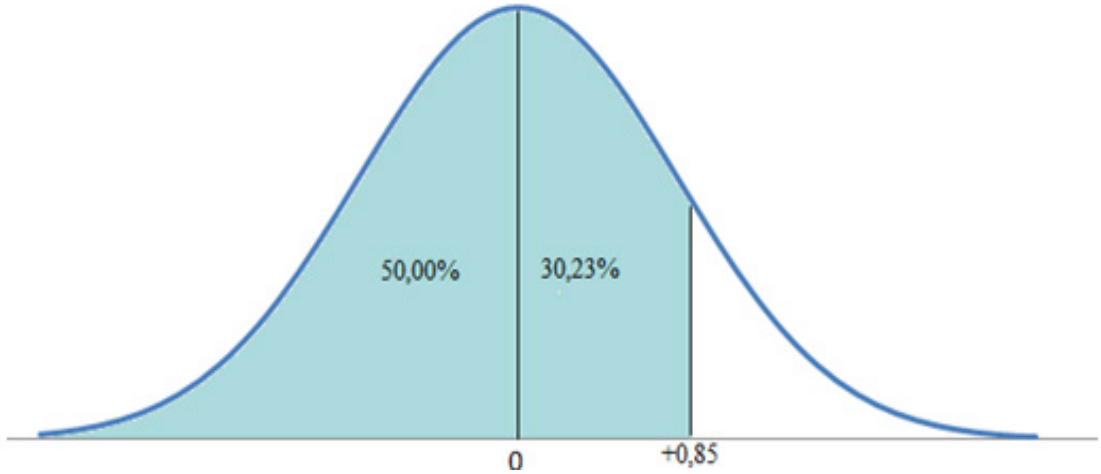


Рисунок 6.4 Нахождение области для положительной оценки Z.

Чтобы найти область ниже отрицательного значения, мы используем правый столбец или «Область за пределами Z». Область, которая нас интересует, изображена на рисунке 5.5, и мы должны определить размер заштрихованной области. «Площадь за пределами» (см. Столбец c), балл -1,35 составляет 0,0885, что можно выразить как 8,85%. Второй субъект (X2 = 73), т.е. субъект набрал более 8,85% баллов в тестиру емой группе.

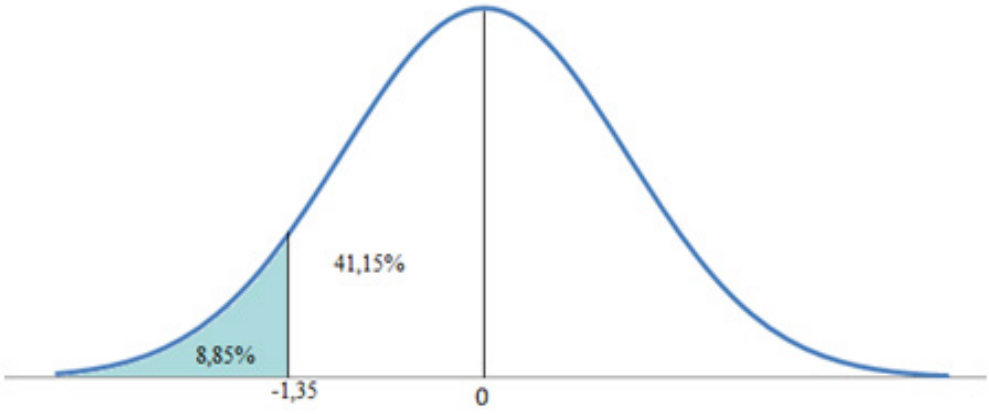


Рисунок 6.5 Нахождение области отрицательной оценки Z.

Нахождение областей между двумя результатами

В некоторых случаях нам нужно будет определить область между двумя баллами. Когда оценки находятся на противоположных сторонах от среднего значения, область между ними можно найти, сложив области между каждым результатом и средним значением. Используя в качестве примера распределение IQ по мужчинам, если мы хотим знать область между оценками IQ от 93 до 112, мы конвертируем обе оценки в оценки Z, находим область между каждой оценкой и средним значением в Приложении A, и сложите эти две области вместе.

Первое значение IQ, равное 93, преобразуется в значение Z, равное -0,35:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{93-100}{20} = -\frac{7}{20} = -0,35$$

Второй показатель IQ (112) достигает +0,60:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{112-100}{20} = \frac{12}{20} = +0,60$$

Обе оценки приведены на рисунке 6.6. Нас интересует общая заштрихованная площадь. Общая площадь между этими двумя показателями составляет 13,68% + 22,57% или 36,25%. Таким образом, 36,25% от общей площади (или около 363 из 1000 случаев) находится между показателями IQ 93 и 112.

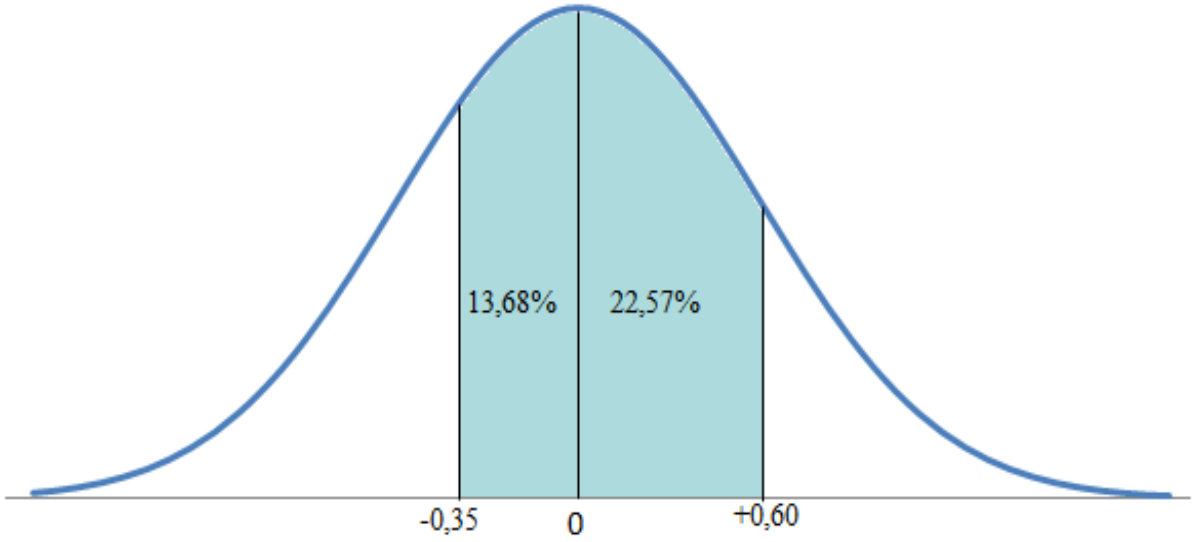


Рисунок 6.6 Нахождение области между двумя значениями.

Когда интересующие баллы находятся на одной и той же стороне от среднего значения, необходимо соблюдать другую процедуру, чтобы определить область между ними. Например, если бы нас интересовала область между баллами 113 и 121, мы бы начали с преобразования этих баллов в Z оценки:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{113-100}{20} = \frac{13}{20} = +0,65$$

$$Z = \frac{X_i - \overline{X}}{s} = \frac{121-100}{20} = \frac{21}{20} = +1,05$$

Баллы отмечены на рисунке 6.7. Нас интересует заштрихованная область, чтобы найти область между двумя показателями с одной и той же стороны от среднего значения. Найдем область между каждой оценкой и средним значением, а затем вычтем меньшую область из большей. Используя столбец b Приложения А, мы видим, что 24,22% с от общей площади лежит между z-оценкой +0,65 и средним значением, а 35,31% этой области находится между Z-баллом + 1,05 и средним. Следовательно, площадь между этими двумя показателями составляет 35,31% - 24,22% или 11,09% от общей площади. Если бы обе оценки были ниже среднего, то расчиты согласно той же методике.

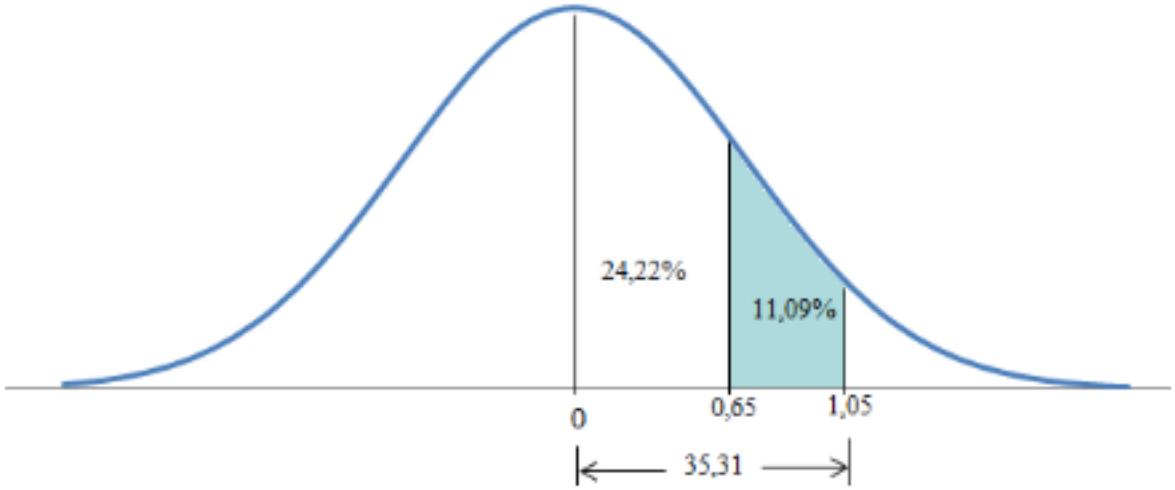


Рисунок 6.7 Нахождение области между двумя значениями

Использование нормальной кривой для оценки вероятностей

К этому моменту мы рассматривали теоретическую нормальную кривую как способ описания областей выше, ниже и между значениями. Мы также видели, что эти области могут быть преобразованы в число случаев выше, ниже и между значениями.

Сейчас мы рассмотрим использование нормальной кривой для оценки вероятностей или шансов того, что определенные события произойдут. Мы найдем эти вероятности, используя те же методы, которые использовались для поиска областей. Единственная новая идея состоит в том, что области под нормальной кривой (или в Приложении А) можно рассматривать как вероятности. Прежде чем мы рассмотрим эти механизмы, давайте рассмотрим, что подразумевается под «вероятностью».

Существуют такие вопросы, как: Какова вероятность дождя? Сдам ли я тест, если я не учусь? Какая комбинация сейчас будет в покере? и т.д.

Чтобы оценить вероятность события, мы должны сначала определить, что будет означать «успех». Предыдущие примеры содержат несколько разных определений успеха (дождь, вытаскивание определенной карты и проходные баллы). Для определения вероятности должна быть установлено отношение, в которой числитель равен числу событий, которые могут составить успех, а знаменатель равен общему числу возможных событий, при которых успех теоретически может возникнуть:

$$\text{Вероятность} = \frac{\text{успешные события}}{\text{общее число возможных событий}}$$

Для примера предположим, что мы хотим знать вероятность выбора конкретной карты - скажем, короля червей - в одном тираже из хорошо перетасованной колоды карт. Наше определение успеха является конкретным (рисунок короля сердец) и с предоставленной информацией мы можем составить отношение. Только одна карта удовлетворяет нашему определению успеха, поэтому число событий, которые составляют успех, равно 1; это значение будет числителем дроби. Есть 52 возможных события (то есть 52 карты в колоде), поэтому знаменатель будет 52. Таким образом, доля равна 1/52, что представляет вероятность выбора короля черв на одном

розыгрыше из колодца. перемешанная колода карт. Наша вероятность успеха 1/ 52.

Мы можем оставить эту дробь такой, какая она есть, или выразить ее несколькими другими способами. Например, мы можем выразить это как отношение шансов, инвертировав долю, показывая, что шансы выбора короля червей на одном розыгрыше равны 52: 1 (или вероятность пятьдесят два к одному). Мы можем выразить дробь как пропорцию, разделив числитель на знаменатель. Для нашего примера соответствующая пропорция равна 0,0192, что является пропорцией всех возможных событий, которые удовлетворяли бы нашему определению успеха. В общественных науках вероятности обычно выражаются в виде пропорций, и мы будем следовать этому правилу. Используя P для обозначения «вероятности», вероятность получения короля червей (или любой конкретной карты) можно выразить как:

$$P(\text{король червь}) = \frac{\text{успех } 1}{\text{событий } 52} = \frac{1}{52} = 0,0192$$

Интерпретируем результат: в долгосрочной перспективе события, которые мы определяем как успехи, будут иметь определенную пропорциональную связь с общим числом событий. Вероятность 0,0192 для выбора короля червей в одном розыгрыше действительно означает, что при тысячах вариантов выбора по одной карте за раз из колоды из 52 карт с хорошими перетасовками доля успешных розыгрышей будет равна 0,0192. Или, на каждые 10 000 розыгрышей, 192 будет королем червей, а остальные 9808 выборов будут другими картами.

Таким образом, когда мы говорим, что вероятность розыгрыша короля червей в одном розыгрыше равна 0,0192, мы, по сути, применяем к одному розыгрышу наши знания о том, что произойдет за тысячи розыгрышей.
Как и в пропорции, вероятности колеблются от 0,00 (что означает, что событие не имеет абсолютно никаких шансов на возникновение) до 1,00 (определенность). Когда значение вероятности увеличивается, вероятность того, что определенное событие произойдет, также возрастает. Вероятность 0,0192 близка к 0, и это означает, что событие выпадение карты короля червь вряд ли или маловероятно.

Эти методы могут быть использованы для установления простых вероятностей в любой ситуации, в которой мы можем указать количество успехов и общее количество событий. Например, один кубик имеет шесть граней или граней, каждая из которых имеет различное значение в диапазоне от 1 до 6. Поэтому вероятность получения любого конкретного числа (скажем, 4) за один бросок кубика равна
p (число 4) =1/6 = 0.1667.

Вероятность и нормальная кривая

Сочетание этого подхода к вероятности с нашими знаниями о теоретической кривой нормалей позволяет нам оценить вероятность выбора случая, который имеет оценку в определенном диапазоне. Например, предположим, что мы хотели оценить вероятность того, что случайно выбранный субъект из распределения показателей IQ мужчин будет иметь IQ балл от 95 до среднего балла 100. Наше определение успеха здесь будет отбор любого предмета с баллом в указанном спектре. Обычно мы затем устанавливаем дробь с числителем, равным количеству предметов с оценками в определенном диапазоне, и знаменателем, равным общему количеству

предметов. Однако, если эмпирическое распределение является нормальным по форме, мы можем пропустить этот шаг, потому что вероятности в пропорциональной форме уже указаны в Приложении А. То есть области в Приложении А можно интерпретировать как вероятности.

Чтобы определить вероятность того, что случайно выбранный случай будет иметь оценку от 95 до среднего значения, мы конвертируем исходную оценку в оценку Z:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{95-100}{20} = -\frac{5}{20} = -0,25$$

Используя приложение А, мы видим, что область между этой оценкой и средним значением составляет 0,0987. Это вероятность, которую мы ищем. Вероятность того, что случайно выбранный случай получит оценку от 95 до 100, равна 0,0987 (или округляется до 0,1 или 1 из 10). Таким же образом можно оценить вероятность выбора субъекта из любого диапазона баллов. Обратите внимание, что методы оценки вероятностей точно такие же, как и при поиске областей.

Чтобы рассмотреть дополнительный пример, какова вероятность того, что случайно выбранный мужчина будет иметь IQ менее 123? Мы найдем вероятности точно так же, как мы нашли области. Оценка (Xi,) выше среднего, мы найдем вероятность, добавив площадь в столбце b к 0.5000. Сначала мы находим Z-оценку:

$$Z = \frac{X_i - \bar{X}}{s} = \frac{123-100}{20} = \frac{23}{20} = +1,15$$

Затем используя значения в столбеце b Приложения А, найдем область между этой оценкой и средним значением. Затем добавем площадь (0,3749) к 0,5000. Вероятность выбора мужчины с IQ менее 123 составляет 0,3749 + 0,5000 или 0,8749. Округляя это значение до 0,88, мы можем сказать, что шансы 0,88 (очень высокие), что мы выберем мужчину с показателем IQ в этом диапазоне. Технически, помните, что эта вероятность выражает то, что произойдет в долгосрочной перспективе: на каждые 100 мужчин, выбранных из этой группы в течение бесконечного числа испытаний, 58 будут иметь IQ баллов менее 123, а 12 - нет. Подчеркнем, очень важный момент о вероятностях и нормальной кривой. Очень высока вероятность того, что любой случай, случайно выбранный из нормального распределения, будет иметь значение, близкое по значению к среднему. Форма нормальной кривой такова, что большинство случаев сгруппированы вокруг среднего значения и уменьшаются по частоте, когда мы смещаемся дальше вправо или влево - от среднего значения. Фактически, учитывая то, что мы знаем о нормальной кривой, вероятность того, что случайно выбранный случай будет иметь оценку в пределах ±1 стандартного отклонения от среднего, составляет 0,6826. Подводя итог, можно сказать, что 68 из 100 случаев - или немногим более двух третей всех случаев - отобранных в долгосрочной перспективе, будут иметь оценку между ±1 стандартным отклонением или Z оценками среднего значения. Вероятности высокая, что любой случайно выбранный случай будет иметь значение, близкое по значению к среднему.

Выводы

1.Нормальная кривая в сочетании со средним и стандартным отклонением может быть использована для построения точных описательных утверждений об эмпирических распределениях, которые обычно искажены.

2. Чтобы работать с теоретической нормальной кривой, мы должны преобразовать необработанные оценки в их эквивалентные оценки Z. Z-оценки позволяют нам находить области под теоретической нормальной кривой (Приложение A).

3. Мы рассмотрели три варианта использования теоретической нормальной кривой: нахождение суммарных оценок выше и ниже балла, выявление областей между двумя баллами и выражение этих областей в качестве вероятностей. Последнее использование нормальной кривой особенно уместно, потому что логическая статистика имеет центральное значение для оценки вероятностей способом, очень похожим на процесс, описанный нами.