

Лабораторна работа 5. Однофакторная линейная регрессия

Постановка задачи

В самых разных областях знания возникает *задача определения зависимости между случайными величинами*, являющимися признаками одних и тех же объектов.

Например, это может быть зависимость

- между ростом и весом человека;
- между силой сигнала на входе и выходе технического устройства;
- между затратами компании на рекламу и доходом от продаж;
- между уровнем инфляции и безработицей;
- между содержанием радиоактивного вещества в растениях-медоносах и в мёде, полученном от этих растений.

На практике на значение исследуемой величины влияет множество *факторов*, но для простоты мы будем считать, что основное влияние оказывает один из них, потому и анализ будем называть *однофакторным*.

Будем считать, что оба признака, зависимость между которыми мы стараемся выявить, представимы как значения вещественных переменных. Предположим, что нам известны результаты n измерений. Каждое измерение $i (i=1, \dots, n)$ даёт пару чисел (x_i, y_i) – значения двух признаков измеряемого объекта, (например, рост – вес, затраты на рекламу – доход и т.д.), т.е. сырые данные представимы как таблица:

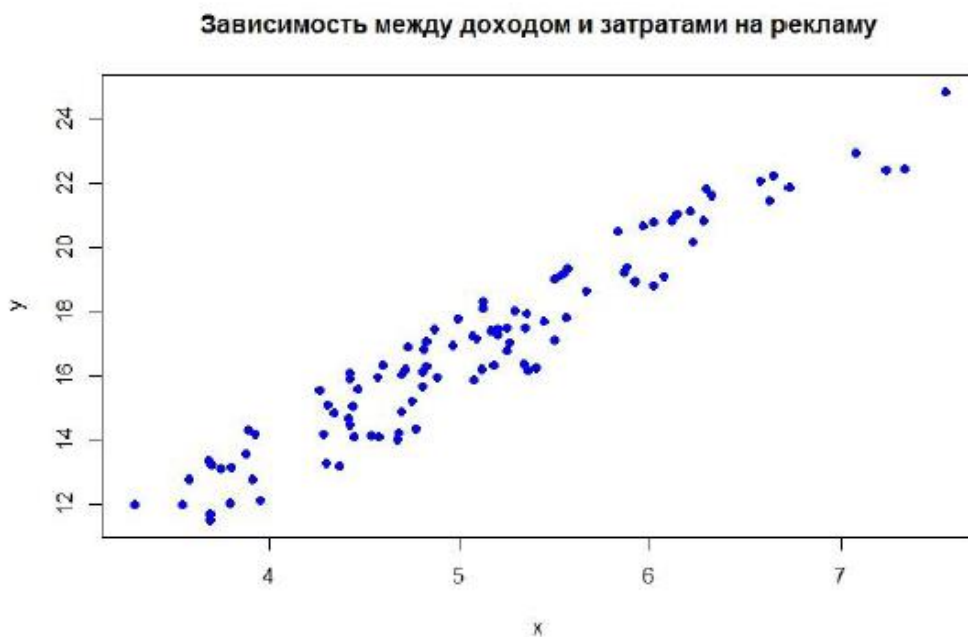
№ наблюдения, i	Значения фактора, x_i	Значения переменной отклика, y_i
1	x_1	y_1
...		
i	x_i	y_i
...		
n	x_n	y_n

Здесь каждая строка соответствует одному *объекту (наблюдению)*. Признак, который может быть непосредственно измерен (x), является *фактором (предиктором)*, прогнозируемая переменная (y) – *переменная отклика*.

Цель исследования – построить (линейную) функцию (регрессионную модель), которая позволит прогнозировать значение переменной отклика (y) по известному значению фактора (x)

Визуализация сырых данных

Построим систему координат, где по оси абсцисс будем откладывать значения фактора (x), по оси ординат – значения переменной отклика (y). Таким образом, каждому наблюдению (т.е. каждой паре) (x_i, y_i) ($i = \overline{1, n}$) соответствует точка на координатной плоскости. Если зависимость между изучаемыми признаками была бы *линейной* и отсутствовала бы случайная компонента, то все эти точки лежали бы на одной прямой. Однако из-за наличия случайного «шума» точки оказываются разбросанными по координатной плоскости в виде так называемого «облака». Пример такого облака показан на приведённом ниже рисунке.



Здесь ось абсцисс соответствует затратам на рекламу, ось ординат – объёму продаж (зафиксированному через заданное время после проведения рекламной кампании).

Поставим задачу: *найти такую линейную функцию, которая наилучшим образом отражает зависимость переменной отклика y (объёма продаж) от фактора x (затрат на рекламу).*

Эта задача называется задачей **однофакторной линейной регрессии**.

Приведём математическую формулировку задачи.

Математическая постановка задачи нахождения уравнения регрессии

Пусть Y зависит от одной переменной x . При этом предполагается, что Y принимает заданные (фиксированные) значения, а зависимая переменная x имеет случайный разброс из-за ошибок измерения, влияния неучтенных факторов и других причин. Каждому значению x соответствует некоторое вероятностное распределение случайной величины (СВ) Y . Предположим, что СВ Y «в среднем» линейно зависит от значений переменной x . Это означает, что условное математическое ожидание

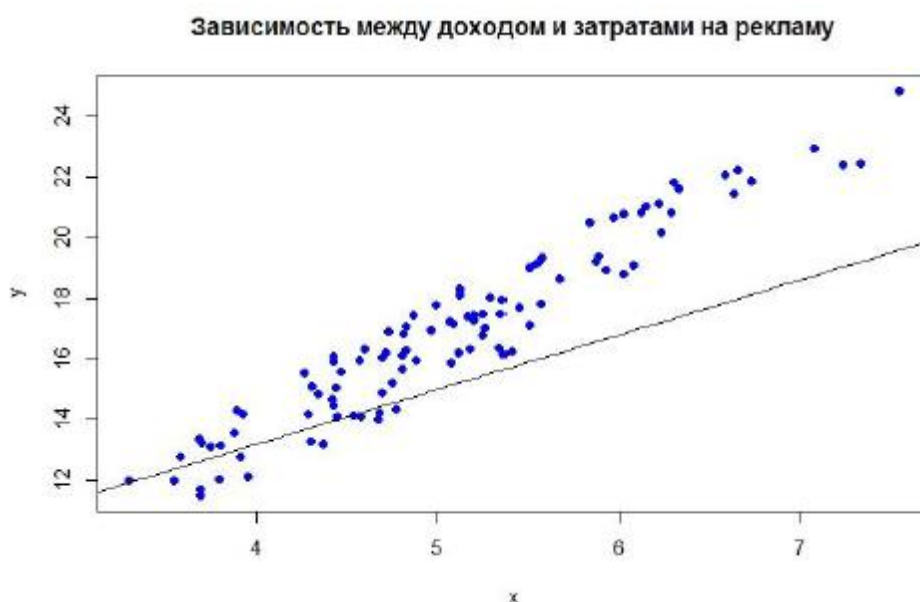
случайной величины Y при заданном значении переменной x имеет вид (y – среднее значение всех значений y для данного x)

$$M(Y|x) = \beta_0 + \beta_1 x$$

Функция переменной x , определяемая правой частью этой формулы, называется линейной регрессией Y на x , а параметры β_0, β_1 – параметрами линейной регрессии.

β_1 – тангенс угла наклона графика этой функции к оси Ox (английский термин: «*slope*»), β_0 – ордината точки пересечения этой прямой с осью Oy (англ.: «*intercept*»). Задача состоит в том, чтобы *найти такие значения переменных β_0, β_1 , при которых прямая (1) наилучшим образом проходит через облако точек $(x_i, y_i), i = 1, \dots, n$.*

Поясним смысл задачи геометрически. Зафиксируем произвольные значения β_0 и β_1 и построим соответствующую прямую:



Очевидно, построенная «наугад» прямая является не самой лучшей для данного облака точек. Формализуем понятие «качества» модели. При фиксированных β_0 и β_1 «ожидаемое» (согласно (1)) значение y при $x = x_i$ составляет $\beta_0 + \beta_1 \cdot x_i$, $i = 1, \dots, n$ (т.е. точка $(x_i, \beta_0 + \beta_1 \cdot x_i)$ лежит на построенной прямой). Но фактическое значение переменной y при $x = x_i$ составляет y_i , т.е. «ошибка» составляет $((\beta_0 + \beta_1 \cdot x_i) - y_i)$.

На практике параметры линейной регрессии неизвестны и их оценки определяются по результатам наблюдений переменных Y и x . Пусть проведено n независимых наблюдений случайной величины Y при значениях переменной X : x_1, x_2, \dots, x_n . При этом измерения величины Y дали следующие результаты: y_1, y_2, \dots, y_n . Так как эти значения имеют разброс относительно линейной регрессии, то связь между переменными Y и x можно записать в виде линейной (по параметрам β_0, β_1) регрессионной модели:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

где ε - случайная ошибка наблюдений, причем $M(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$. Значение дисперсии ошибок σ^2 неизвестно, и оценка ее определяется по результатам наблюдений.

Задача линейного регрессионного анализа состоит в том, чтобы по результатам наблюдений (x_i, y_i) , $i = \overline{1, n}$:

- 1) Получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2 линейной регрессионной модели.
- 2) Проверить статистические гипотезы о параметрах модели.
- 3) Проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

В соответствии с моделью результаты наблюдений зависимой переменной Y : y_1, y_2, \dots, y_n являются реализациями случайных величин $\beta_0 + \beta_1 x_i + \varepsilon_i$, обозначаемых y_i , $i = \overline{1, n}$.

Задача линейного регрессионного анализа решается в предположении, что случайные ошибки наблюдений ε_i и ε_j не коррелированы, имеют математические ожидания

равные нулю, и одну и ту же дисперсию, равную σ^2 , т.е. $M(\varepsilon_i) = 0$, $K_{\varepsilon_i \varepsilon_j} = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$,

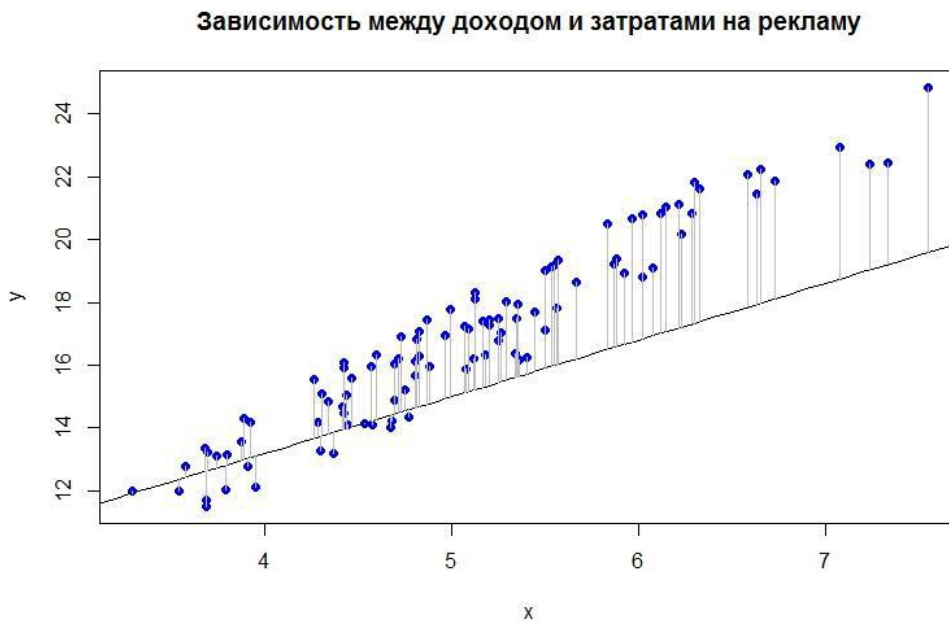
$i = \overline{1, n}$. При статистическом анализе регрессионной модели предполагается также, что случайные ошибки наблюдений $\varepsilon_i, i = \overline{1, n}$ имеют нормальное распределение, т.е. $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1, n}$. В этом случае ошибки наблюдений также являются независимыми случайными величинами. Можно определить величину ошибки для всех отмеченных точек. Линейная модель, которая наилучшим образом аппроксимирует данные – одна из тех, для которых общая ошибка выборки имеет наименьшее значение. Чтобы рассчитать ее нужно избежать позитивных и негативных значений. Это можно сделать, возведя все ошибки в квадрат и делая их положительными величинами. Линия наилучшего подбора – та, которая минимизирует квадраты разниц между рассматриваемыми значениями y и соответствующими значениями x , рассчитанными с помощью линии наилучшего подбора. Эта линия называется линией регрессии, полученной методом наименьших квадратов. Для нахождения оценок параметров модели по результатам наблюдений используется метод наименьших квадратов. По этому методу выбирают такие оценки β_0, β_1 , которые минимизируют сумму квадратов отклонений наблюдаемых значений случайных величин Y_i от их математических ожиданий, т.е.

$$\sum_{i=1}^n ((\beta_0 + \beta_1 \cdot x_i) - y_i)^2 \rightarrow \min \quad (2)$$

Метод наименьших квадратов

Принцип поиска коэффициентов регрессии путём минимизации суммы квадратов отклонений между реальными значениями признака и прогнозируемыми согласно предполагаемой форме зависимости (в нашем случае – линейной) называется *методом наименьших квадратов* (англ.: Least Square Method, LSM).

Проиллюстрируем целевую функцию задачи (2) на следующем рисунке.



Значение целевой функции задачи (2) при фиксированных значениях β_0 и β_1 равно сумме квадратов длин построенных отрезков. Из рисунка видно, что построенная прямая – не лучшая, так как можно провести прямую, обеспечивающую меньшее значение целевой функции задачи (2).

Найдём решение задачи (2). Целевую функцию задачи (2) обозначим через $\varphi(\beta_0, \beta_1)$. Очевидно, $\varphi(\beta_0, \beta_1)$ – дифференцируемая функция двух переменных. Найдём её частные производные:

$$\frac{\partial \varphi}{\partial \beta_0} = 2 \sum_{i=1}^n ((\beta_0 + \beta_1 \cdot x_i) - y_i) \quad (3)$$

$$\frac{\partial \varphi}{\partial \beta_1} = 2 \sum_{i=1}^n ((\beta_0 + \beta_1 \cdot x_i) - y_i) x_i$$

и запишем систему для поиска стационарной точки:

$$\begin{cases} \sum_{i=1}^n ((\beta_0 + \beta_1 \cdot x_i) - y_i) = 0 \\ \sum_{i=1}^n ((\beta_0 + \beta_1 \cdot x_i) - y_i) x_i = 0 \end{cases} \quad (4)$$

После несложных преобразований системы (4) получим

$$\begin{cases} n\beta_0 + \beta_1 \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0 \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \cdot \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0 \end{cases} \quad (5)$$

Введём обозначения: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $\overline{x \cdot y} = \frac{1}{n} \sum_{i=1}^n x_i y_i$ и $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$.

Поделив оба уравнения системы (5) на n , получим

$$\begin{cases} \beta_0 + \beta_1 \cdot \bar{x} - \bar{y} = 0 \\ \beta_0 \bar{x} + \beta_1 \cdot \overline{x^2} - \overline{x \cdot y} = 0 \end{cases} \quad (6)$$

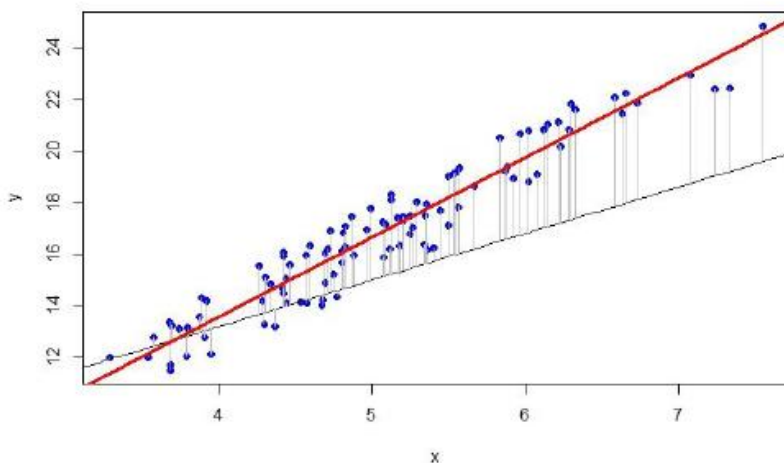
Нетрудно показать, что система (5) имеет единственное решение (β_0^*, β_1^*) , где

$$\beta_1^* = \frac{\bar{x} \cdot \bar{y} - \overline{x \cdot y}}{\overline{x^2} - \bar{x}^2} \quad \beta_0^* = \bar{y} - \beta_1^* \cdot \bar{x} \quad (7)$$

Учитывая свойства функции $\varphi(\beta_0, \beta_1)$, нетрудно показать также, что это решение (т.е., стационарная точка функции $\varphi(\beta_0, \beta_1)$) является *точкой минимума* функции $\varphi(\beta_0, \beta_1)$.

Иными словами, определяемые формулами (7) значения β_0^* и β_1^* обеспечивают получение наилучшей (в смысле задачи (2)) линейной функции, отражающей зависимость переменной отклика y от фактора x . График этой линейной зависимости называется *прямой регрессии* (y на x). На приведённом ниже рисунке эта прямая имеет красный цвет.

Зависимость между доходом и затратами на рекламу



Найдена лінійна функція дозволяє *прогнозувати* значення залежного признака (y) по заданим значенням незалежного фактора (x).

Завдання

Після кількох років роботи менеджер фірми прийняв рішення щодо збільшення грошових витрат на проведення рекламної діяльності. Статистичні дані розмірів додаткових щомісячних витрат $\{x_i\}$ фірми на рекламу та отриманих від цього прибутків $\{y_i\}$ містить таблиця (таблиця (n - номер студента за списком групи):

i	1	2	3	4	5
x_i	$5n$	$7n$	$10n$	$n+5$	$n+3$
y_i	n	$2n$	$20n$	$n+30$	$n+25$

Вважаючи, що розмір додаткового прибутку y лінійно залежить від додаткових витрат x на рекламу ($y_{\text{теоретичне}} = \beta_1^* x + \beta_0^*$) обчислити за методом найменших квадратів значення параметрів β_0^* та β_1^* за формулами (7). Також обчислити:

- 1) коефіцієнт детермінації R^2 :
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_{i \text{ теоретичне}})^2}{\sum_{i=1}^n (y_i - y_{i \text{ середє}})^2};$$
- 2) емпіричну дисперсію отриманого прибутку σ_y^2 :
$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \left(\frac{1}{n} \sum_{i=1}^n y_i \right)^2;$$
- 3) загальне квадратичне відхилення f_* щомісячних очікуваних прибутків від фактичних

$$f_* = \sum_{i=1}^n (y_i - y_{i \text{ теор}})^2;$$

- 4) побудувати графік залежності $y_{\text{теоретичне}} = \beta_1^* x + \beta_0^*$;
- 5) перевірити адекватність одержаної моделі за допомогою критерію Фішера.

$$\text{Розрахункове значення критерію Фішера } F_{\text{Фішера, розрах}} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

де R^2 – коефіцієнт детермінації, n – число даних спостережень, m – число незалежних змінних в рівнянні регресії.

Табличне значення: $F_{\text{Фішера, табл}}(0,05; k_1 = m; k_2 = n - m - 1)$

Якщо $F_{\text{Фішера, табл}} > F_{\text{Фішера, розрах}}$ модель не адекватна, у противному випадку модель є адекватною.

Таблиця F -розподілу Фішера (k_1, k_2 – ступені вільності, $\alpha = 0,05$)

$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,1	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,931
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,711
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,99	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,64	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00