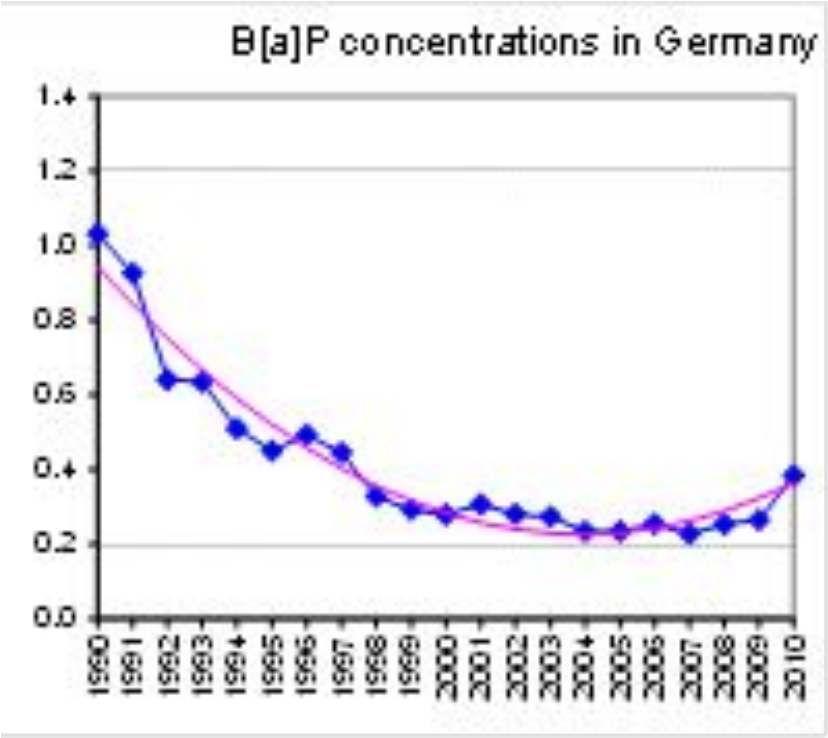
Статистика

Тренды

- Тренд (от англ. trend — *тенденция*) — это долговременная тенденция изменения исследуемого временного ряда.
- Тренды могут быть описаны различными уравнениями — линейными, логарифмическими, степенными и так далее.
- Методы оценки
- Параметрические — рассматривают временной ряд как гладкую функцию При этом сначала выявляют один либо несколько допустимых типов функций, затем различными методами оценивают параметры этих функций, после чего на основе проверки критериев адекватности выбирают окончательную модель тренда.
- Непараметрические — это разные методы сглаживания исходного временного ряда —скользящие средние (простая, взвешенная), экспоненциальное сглаживание. Они полезны в случае, когда для оценки тренда не удается подобрать подходящую функцию.

Анализ ряда данных

- Для анализа тренда необходимо разложить временные ряды на сумму регулярной составляющей (тренда) и остатка (шума).
- $y_t = T_t + \omega_t$, $t = 1, \ldots, N$,



Анализ ряда данных (продолжение)

- Для анализа тренда временных рядов необходимо выполнить следующие шаги:
- Шаг 1. Обнаружение тенденции и ее характер. На этом этапе нужно убедиться, что тренд существует и определяет характер тренда (увеличение, уменьшение или смешение).
- Шаг 2. Идентификация типа тренда. На этом этапе следует выбрать тип тренда, подходящий для описания общих тенденций рассматриваемых временных рядов (например, линейного тренда, экспоненциального тренда и т. д.). Ниже приводятся возможные типы тенденций.
- Шаг 3. Количественная оценка тренда. На этом этапе выполняется выбор основных параметров, описывающих тренд выбранного типа.
- Шаг 4. Расчеты и интерпретация полученных результатов.

Тест Манна-Кендалла

- Непараметрический тест для определения наличия монотонной, статистически значимой тенденции.
- Для многолетних рядов данных без явно выраженных сезонных колебаний.
- Для временных рядов с менее чем 10 значений используется S – статистика (Gilbert (1987)), для временных рядов от 10значений используется нормальное приближение (normal approximation) или Z статистика
- Основан на статистике S или Z, рассчитанной как разность между возрастающими и уменьшающимися парами значений в исследуемом временном ряду

S-статистика Теста Манна-Кендалла

- Если временной ряд состоит из 9 или менее значений то результаты расчета по формулам сравниваются непосредственно с теоретическим распределением, полученный Манном и Кендалом (Gilbert, 1987).
- Полученные значения сравниваются с определенными табличными значениями и в результате подтверждается или опровергается нулевая гипотеза (гипотеза, что тренда нет).

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}(x_j - x_k),$$

$$\operatorname{sgn}(x_j - x_k) = \begin{cases} 1 & if\ x_j - x_k > 0 \\ 0 & if\ x_j - x_k = 0 \\ -1 & if\ x_j - x_k < 0 \end{cases}.$$

Significance level α	required n
0.1	≥ 4
0.05	≥ 5
0.01	≥ 6
0.001	≥ 7

Аппроксимационный тест (Z-статистика) Теста Манна-Кендалла

- Для временного ряда из 10 и более значений.
- Для проведения данной проверки рассчитывается S и ее дисперсия (с учетом возможности наличия «связанных» или «равных» значений).
- Если вычисленное значение статистики Z превышает соответствующий порог по абсолютной величине, то предполагается, что серия имеет тенденцию на соответствующем уровне достоверности.

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{VAR(S)}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \dfrac{S+1}{\sqrt{VAR(S)}} & if\ S < 0 \end{cases}$$

Метод Сенса

- Использует линейную модель для оцени наклона тренда (т.е. в случаях, если предполагается что тренд линейный).
- Распределение «остатков» предполагается постоянной во времени.
- Не чувствителен к ошибочным значениям и «выбросам».
- Для каждой пары рядом стоящих чисел рассчитывается угол наклона Qi.
- Если в временном ряду есть n значений xj, мы получаем столько же, сколько N = n (n-1) наклона Qi.
- Оценкой склона Сена является медиана значений Qi. Значения N Qi оцениваются наименьшего до самого большого

$$Q_i = \frac{x_j - x_k}{j - k}, \qquad j > k.$$

$$Q = Q_{[(N+1)/2]}, \text{ if } N \text{ is odd}$$

$$Q = \frac{1}{2}\left(Q_{[N/2]} + Q_{[(N+2)/2]}\right), \text{ if } N \text{ is even.}$$

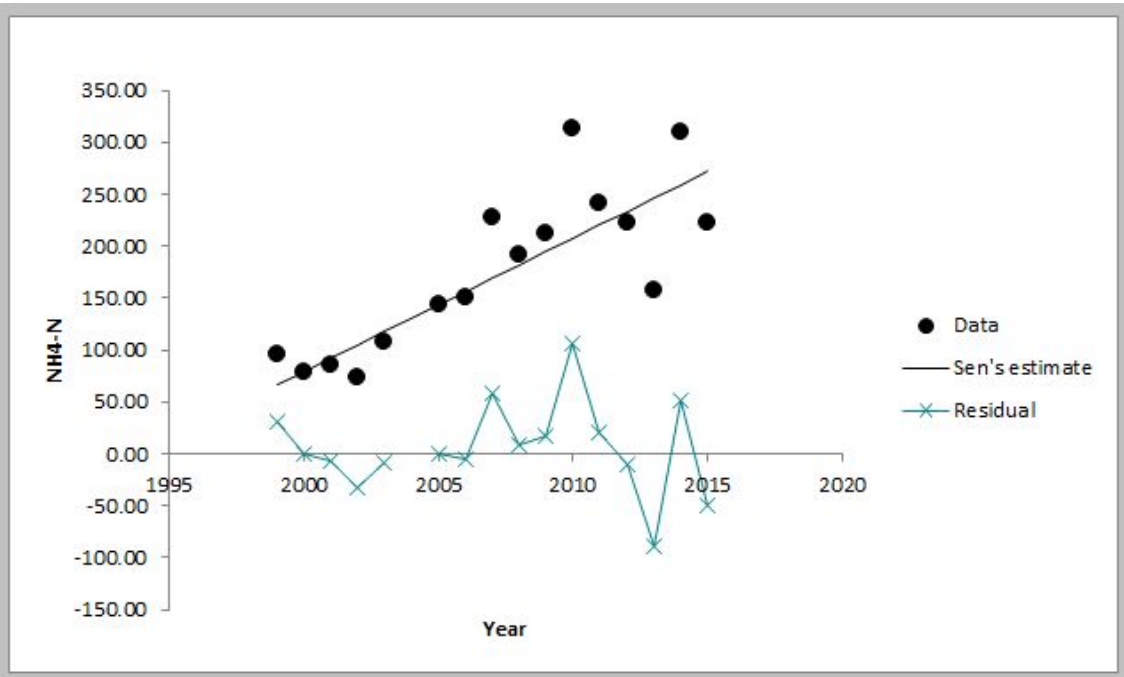
TREND STATISTICS

Pinega 1999-2013

				Mann-Kendall trend			Sen's slope estimate									
Time series	First year	Last Year	n	Test S	Test Z	Signific.	Q	Qmin99	Qmax99	Qmin95	Qmax95	B	Bmin99	Bmax99	Bmin95	Bmax95
SO4-S	1999	2015	16		1.22		4.522	-10.159	13.915	-3.719	11.158	202.82	329.10	123.42	261.80	146.8
NO3-N	1999	2015	16		2.93	**	4.047	1.241	6.661	1.787	6.371	46.90	72.04	22.23	69.31	25.3
NH4-N	1999	2015	16		3.38	***	12.891	4.731	19.455	7.765	17.403	66.05	132.61	18.30	98.14	38.6
Na	1999	2015	16		0.95		8.907	-21.747	53.369	-13.355	23.590	216.11	580.85	58.88	458.15	104.7
Mg	1999	2015	16		1.13		2.151	-4.662	6.049	-1.484	5.223	56.20	129.05	20.77	89.53	24.4
Ca	1999	2015	16		1.40		12.444	-16.829	29.222	-3.508	22.663	178.45	475.43	82.88	358.86	97.1
Cl	1999	2015	16		0.95		6.947	-20.661	30.970	-8.284	22.981	377.23	661.01	199.33	536.66	237.0
H	1999	2015	16		-3.29	**	-193.858	-316.066	-53.456	-276.215	-103.182	3710.91	5075.29	2245.13	4676.78	2792.1
K	1999	2015	16		0.14		1.145	-13.390	12.466	-10.446	8.890	229.11	399.11	131.87	370.65	150.7

Pinega 1999-2013

TsNumber	3
Name	NH4-N
Years	1999 - 2015
n	16
Test S	
Test Z	3.38
Signific.	***
Q	1.29E+01
Qmin99	4.73E+00
Qmax99	1.95E+01
Qmin95	7.76E+00
Qmax95	1.74E+01
B	6.60E+01
Bmin99	1.33E+02
Bmax99	1.83E+01
Bmin95	9.81E+01
Bmax95	3.87E+01

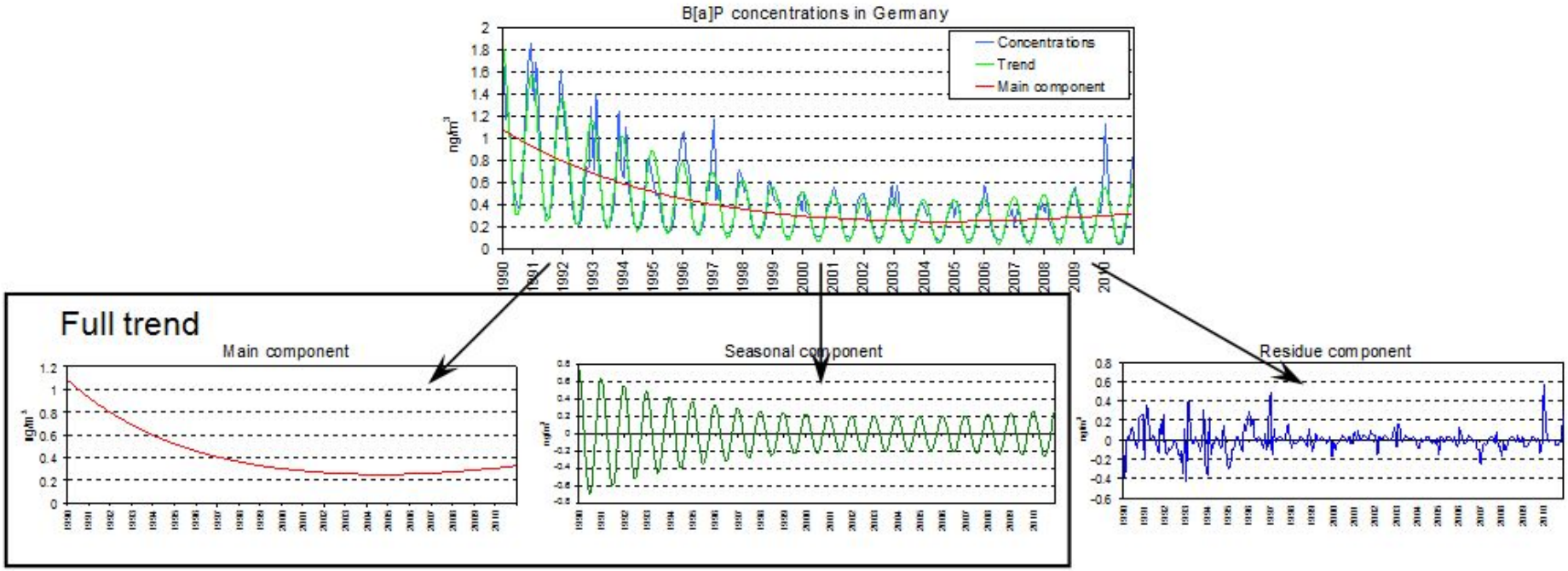


Анализ ряда данных с явно выраженной сезонной составляющей (один из методов)

- Тренд (Tt) можно разложить на несколько компонентов тренда, описывающих разные типы поведения исследуемых величин(y_t) во времени, например «основной» тренд и сезонную составляющую.
- $C = C_{main} + C_{seas} + \omega,$
- $C_{main,t} = a_1 \cdot \exp(- t / \tau_1) + a_2 \cdot \exp(- t / \tau_2),$ (15)
- $C_{seas,t} = a_1 \cdot \exp(- t / \tau_1) \cdot (b_{11} \cdot \cos(2\pi \cdot t - \varphi_{11}) + b_{12} \cdot \cos(4\pi \cdot t - \varphi_{12}) + ...) + a_2 \cdot \exp(- t / \tau_2) \cdot (b_{21} \cdot \cos(2\pi \cdot t - \varphi_{21}) + b_{22} \cdot \cos(4\pi \cdot t - \varphi_{22}) + ...).$

- $C_t = a_1 \cdot \exp(- t / \tau_1) \cdot (1 + b_{11} \cdot \cos(2\pi \cdot t - \varphi_{11}) + b_{12} \cdot \cos(4\pi \cdot t - \varphi_{12}) + ...)$

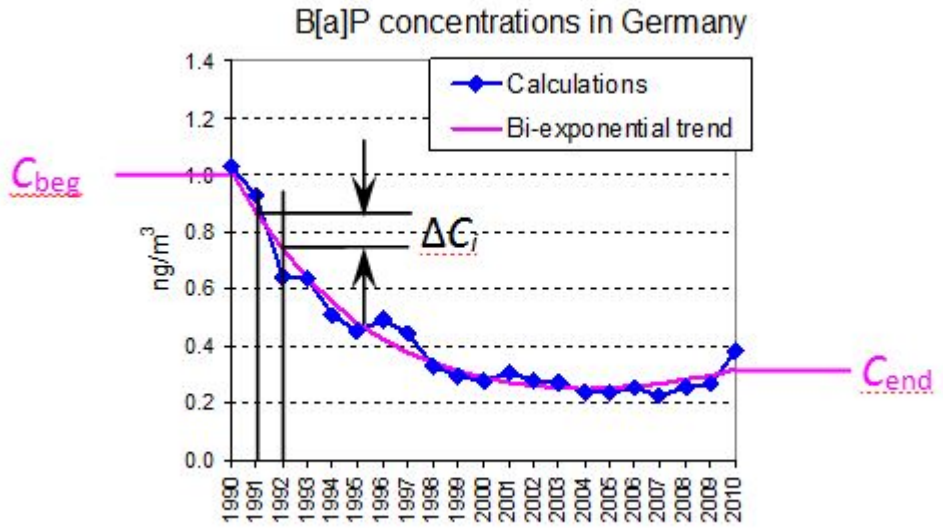
  $+ a_2 \cdot \exp(- t / \tau_2) \cdot (1 + b_{21} \cdot \cos(2\pi \cdot t - \varphi_{21}) + b_{22} \cdot \cos(4\pi \cdot t - \varphi_{22}) + ...) + \omega_t,$



- Количество гармоник можно рассчитать с использованием F-статистики. Для анализа содержания ТМ и СОЗ оптимальным считается 2 гармоники. Использование двух гармоник позволяет избежать в некоторых случаях таких артефактов как отрицательных значений расчетного тренда для концентрации воздуха.

Пример Количественной оценки тренда

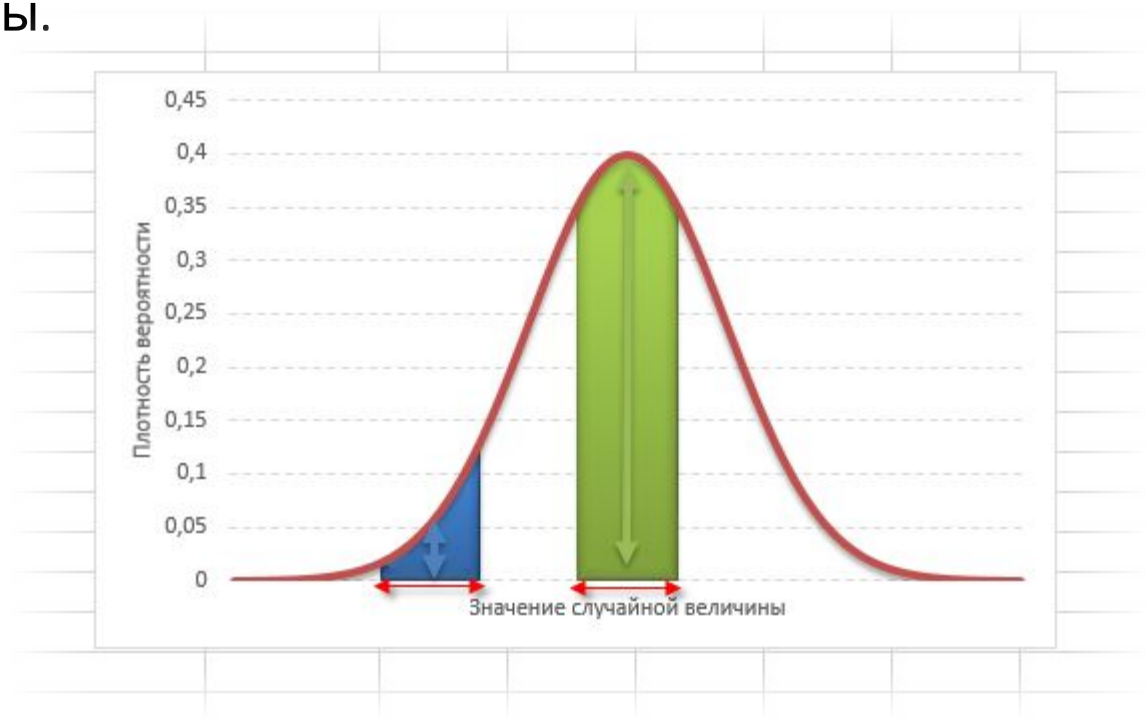
- total reduction: $R_{tot} = (C_{beg} - C_{end}) / C_{beg} = 1 - C_{end} / C_{beg}$,
- annual reduction for year i: $R_i = \Delta C_i / C_i = 1 - C_{i+1} / C_i$,



- Значения остаточной составляющей ω следуют величине основного компонента, ymain, соответственного остаточную компоненту можно нормализовать по основной компоненте
- В качестве характеристики остаточной составляющей можно использовать следующую величину по сравнению с основной:

Нормальное (Гаусса) распределение

- это функция, которая описывает тенденцию высокой концентрации значений около центра
- Кривая Гаусса по форме несколько напоминает колокол, поэтому график нормального закона часто еще называют **колоколообразной кривой**.
- Вероятность того, что случайная величина окажется около центра гораздо выше, чем то, что она сильно отклонится от середины.

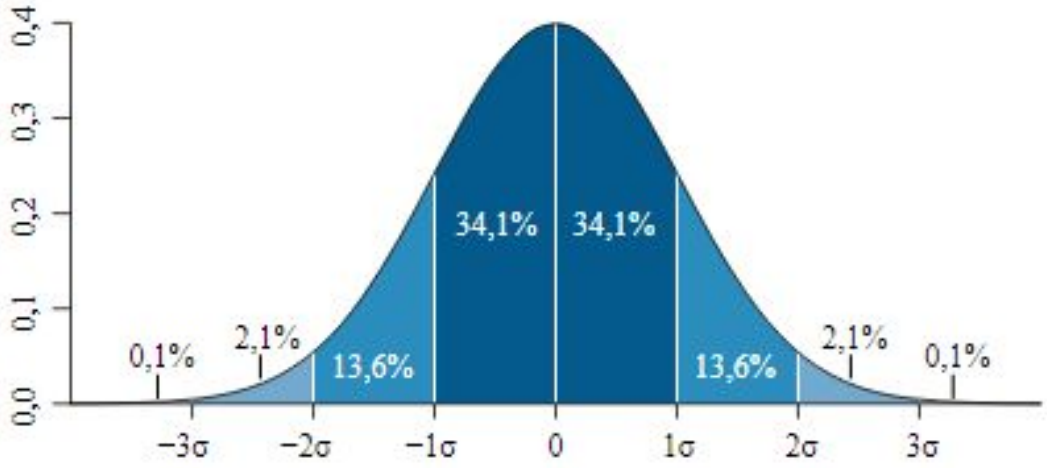


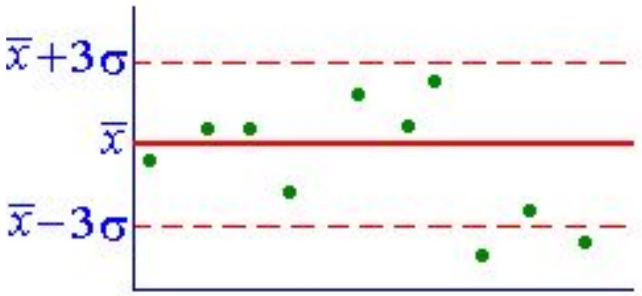
$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

- Параметр *m* (матожидание) определяет центр распределения, которому соответствует максимальная высота графика. Дисперсия σ^2 характеризует размах вариации, то есть «размазанность» данных.

Правило трёх сигм

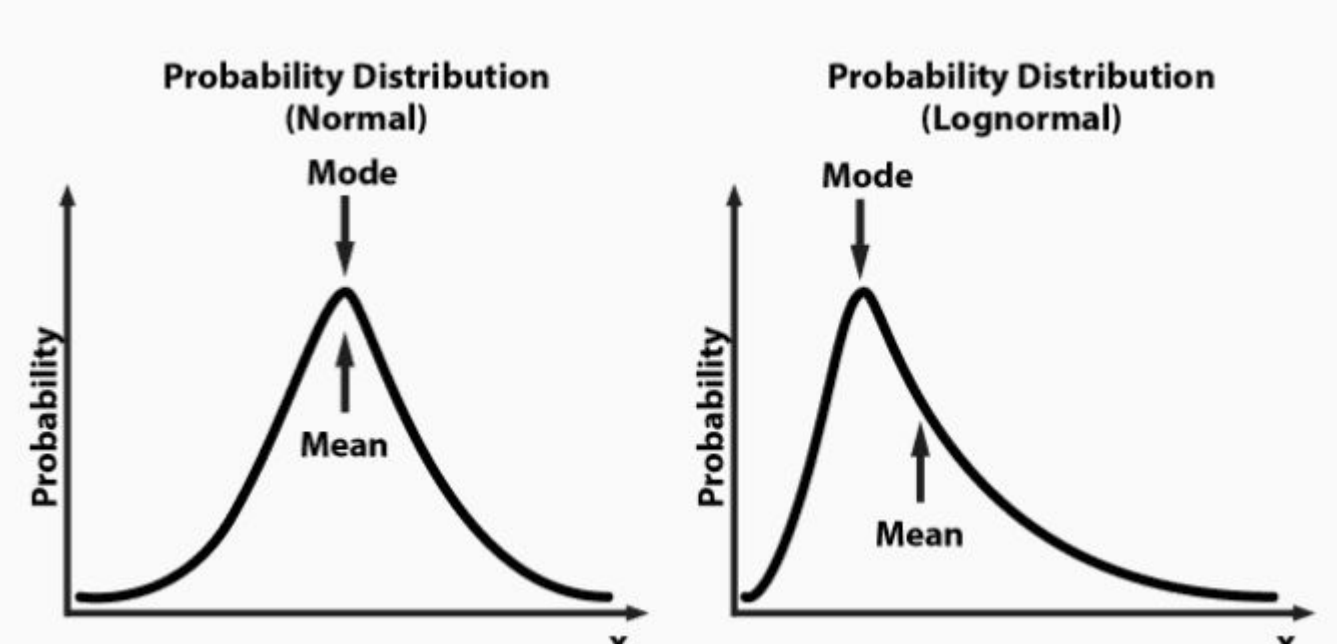
- Вероятность того, что случайная величина отклонится от своего математического ожидания на большую величину, чем утроенное среднее квадратичное отклонение, практически равна нулю. Правило справедливо только для случайных величин, распределенных по
- нормальному закону.





Логнормальное распределение

- случайная величина X имеет логнормальное распределение с параметрами μ, σ, если X = exp(Y), где Y имеет нормальное распределение с параметрами μ, σ. Случайная величина с логнормальным распределением является непрерывной, и принимает только положительные значения. Графики плотности (привязан к левой вертикальной оси ординат).



Оценка показателя повторяемости методики анализа

- Рассчитывают среднее арифметическое и выборочную дисперсию результатов единичного анализа содержания компонента, полученных в условиях повторяемости (параллельных определений).

$$S^2_{m,l} = \frac{\sum_{i=1}^{N} (X_{m,l,i} - X_{m,l})^2}{N-1}$$

Критерий Кохрена

- Рассчитывается для выборки и сравнивается с табличными значениями. Если рассчитанного значение выше табличного, то соответствующая дисперсия исключается из дальнейшего расчета.
- Не исключенные из расчетов дисперсии считают однородными и по ним оценивают средние квадратические отклонения, характеризующие повторяемость результатов единичного анализа (параллельных определений).

$$G_{m(\max)} = \frac{(S^2_{m,l})_{\max}}{\sum_{i=1}^{L} S^2_{m,l}}$$

Оценка показателя правильности методики анализа

- Рассчитывают значение смещения - как разность между средним значением результатов анализа , и аттестованным значением.
- Далее проверяют значимость вычисленных значений по критерию Стьюдента.

$$t_m = \frac{|\Theta_m|}{\sqrt{\frac{S_m^2}{L} + \frac{\Delta_{o,m}^2}{3}}}$$

In the game of darts or archery…

Accuracy = distance from the center of the target
Precision = size of the arrow cluster