**Методические указания по лабораторной работе. М.В.Морозов (docentmorozov.ru)**

**Образцов И.М., группа МГП-МСМХСIII**
**Горный университет**
**Санкт-Петербург**
**Кафедра МКП**

Дисциплина: **Математические методы моделирования в геологии**

**Лабораторная работа 1.** Статистическое распределение. Оценка параметров распределения.
**Лабораторная работа 2**. Проверка гипотезы о нормальном распределении случайной переменной.
**Лабораторная работа 3**. Знакомство с программой STATISTICA.

**Задание.** Проверить гипотезу о соответствии выборочного распределения нормальной модели и рассчитать числовые характеристики случайной переменной.
1. Определить статистические характеристики распределения **металла** (M, г/т) по результатам литохимического опробования почв (горизонт B) Забайкальского края, доказательно описать характер распределения, степень его соответствия нормальному.
2. Изучить распределение логарифмов содержаний теми же способами.
3. Определить наличие аномальных значений и повторить статистику после их удаления для выбранной модели распределения (нормальная или логнормальная).

| **Общий объем выборки** | 1803 | | |
|---|---|---|---|
| **Статистические параметры** | **По содержаниям M, г/т** | **По логарифмам содержаний** | |
| | | **Логарифмы LgM** | **Антилогарифмы $10^{LgM}$** |
| Мода | 0.0005 | -3.301 | 0.0005 |
| 1. Наша цель – определить истинную моду распределения, но программа в качестве моды предлагает лишь значение с максимальной частотой (frequency), в данном случае – 126 значений. Т.е. для программы «мода» - это самый высокий столбец гистограммы. При ограниченной выборке и неидеальном распределении самое частое значение не обязательно соответствует истинной моде, поэтому здесь значение «моды» проще игнорировать. Распределение реальных частот удобнее наблюдать на гистограмме. В настоящей работе «мода» используется для учебных целей: мы сравниваем ее с медианой и средним значением, чтобы оценить полезность «моды».<br>2. Обратите внимание: поскольку логарифм – монотонная функция, то мода по содержаниям всегда равна антилогарифму моды по логарифмам.<br>3. В нашем примере самое частое значение – это минимальное значение. Оно соответствует нижнему пределу обнаружения (НПО) при химическом анализе и поэтому не имеет физического смысла, т.к. всем значениям ниже НПО принудительно присваивается одно и то же значение. По сути, этот столбец гистограммы – «сборный» из всех столбцов, которые располагаются ниже НПО и тем самым являются недоступными для определения. См. гистограммы (ниже). | | | |
| Среднее арифметическое | **0.064** | -1.562 | 0.027 |
| Доверительный интервал среднего, 95% | 0.059 – 0.069 | -1.593 – -1.532 | – |
| Стандартная ошибка среднего | 0.003 | 0.016 | – |

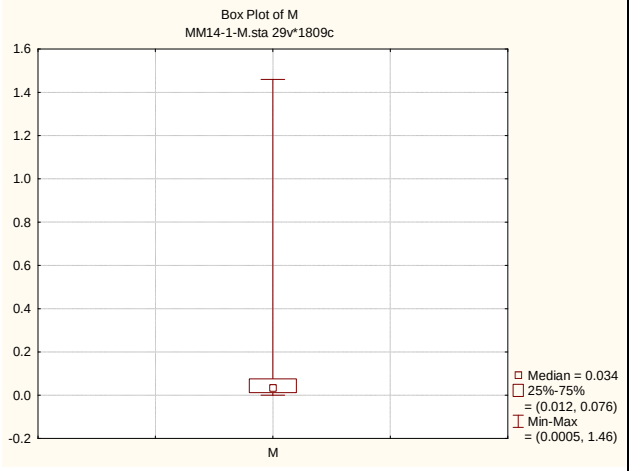
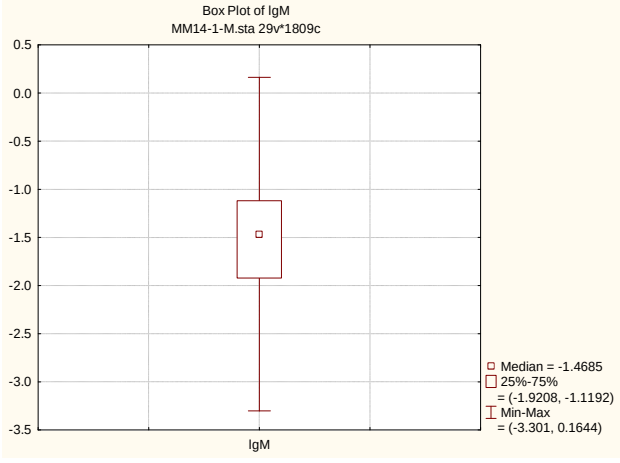
| | | | |
|---|---|---|---|
| Если распределение нормально – то физический смысл имеет левое значение (0.064 г/т), если логнормально – то правое (0.027 г/т). | | | |
| Медиана | **0.034** | -1.469 | 0.034 |
| 1. Обратите внимание: поскольку логарифм – монотонная функция, то медиана по содержаниям всегда равна антилогарифму медианы по логарифмам.<br>2. Медиана почти в два раза **отличается** от среднего арифметического по содержаниям, но значения по логарифмам близки. Следовательно, распределение ближе к логнормальному, чем к нормальному. | | | |
| Среднее геометрическое | 0.027 | – | – |
| 1. Среднее геометрическое точно соответствует антилогарифму среднего арифметического по логарифмам, как и должно быть.<br>2. Среднее геометрическое близко к значению медианы. Следовательно, распределение стремится к логнормальному. | | | |
| Как мы видим, изучив медиану и средние, значение «моды» не несет для нас никакой полезной информации. | | | |
| Дисперсия | 0.011 | 0.440 | – |
| Стандартное отклонение | 0.107 | 0.663 | 4.602 |
| Минимальное значение | 0.001 | -3.301 | – |
| Нижний (25%) квартиль | 0.012 | -1.921 | – |
| Верхний (75%) квартиль | 0.076 | -1.119 | – |
| Максимальное значение | 1.460 | 0.164 | – |
| Межквартильное расстояние | 0.064 | 0.802 | – |
| Робастная оценка стандартного отклонения | 0.047 | 0.594 | 3.928 |
| 1. Квартили распределены неравномерно: расстояние от верхнего квартиля до максимального значения (область верхних 25%) превышает межквартильное расстояние (50% выборки), что соответствует непропорционально вытянутому к верхним значениям «усу» ящика с усами. Иными словами, распределение сильно асимметричное, с правой асимметрией (как и у логнормального).<br>2. Стандартное отклонение сильно завышено относительно его робастной оценки для содержаний. Следовательно, распределение резко не нормальное. В то же время обе оценки близки для логарифмов значений, следовательно, распределение ближе к логнормальному, но все же логарифмирование не нормализует его, т.к. разница остается существенной (0.66 - 0.59, т.е. >10%). Можно предположить наличие аномальных значений, не позволяющих считать распределение нормальным даже для логарифмов. Это также подтверждает большая величина стандартного отклонения для логарифмов (~0.6 – полпорядка) и, соответственно, его антилогарифма (~4.6), т.е. дисперсия значений аномально велика. | | | |
| Асимметрия | 6.075 | -0.921 | |
| Стандартная ошибка асимметрии | 0.058 | 0.058 | |
| Эксцесс | 55.31 | 1.14 | |
| Стандартная ошибка эксцесса | 0.115 | 0.115 | |

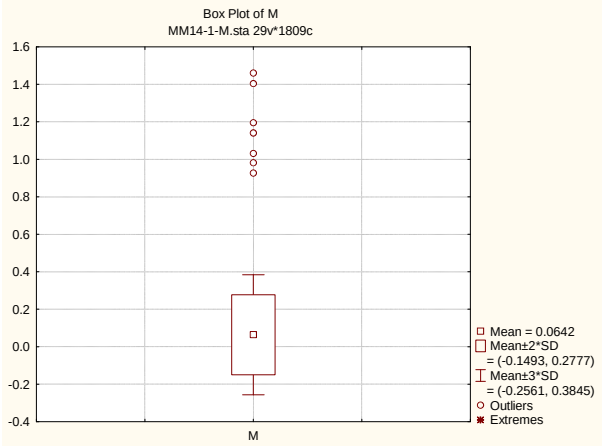
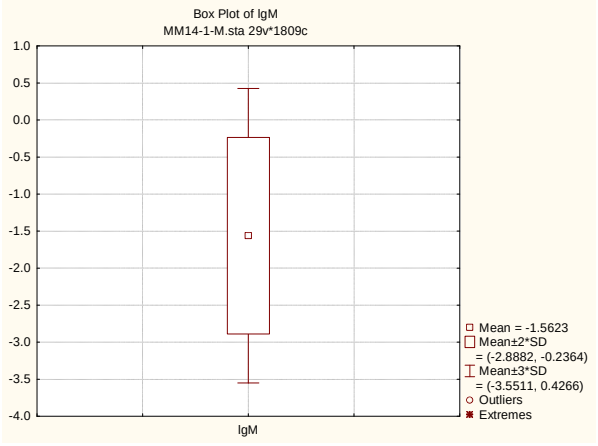
| | | | |
|---|---|---|---|
| Коэффициент вариации, % | 166.3 | – | – |
| **Оценка характера распределения** | | | |
| Значения медианы, моды, средних значений (для содержаний и логарифмов), позволяют сделать вывод о близости распределения к логнормальному. | | | |
| Сравнение величины стандартного отклонения и его робастной оценки позволяет сделать вывод, что распределение близко к логнормальному и, кроме того, содержит аномальные значения. | | | |
| Коэффициент вариации превышает 100%, что исключает возможность нормального распределения и говорит о большом разбросе значений. Причиной может быть логнормальный характер распределения и (или) наличие аномальных мод. | | | |
| Коэффициент асимметрии ($K_A$) | 105.3 >> 3<br>Сокр. формула.<br>105.4 >> 3<br>Точная формула. | 16.0 > 3<br>По сокращенной и точной формулам. | |
| Коэффициент эксцесса ($K_E$) | 479.4 >> 3<br>Сокр. формула.<br>480.1 >> 3<br>Точная формула. | 9.9 > 3<br>По сокращенной и точной формулам. | |
| 1. Асимметрия распределения положительная, как и логнормального распределения. Логарифмирование сильно сокращает асимметрию, но не приводит ее в 95% доверительный интервал нормального распределения (<3). Вывод: распределение близко к логнормальному, но содержит аномальные значения.<br>2. Эксцесс не соответствует нормальному распределению, но логарифмирование сильно уменьшает его, хотя и не приводит в 95% доверительный интервал нормального распределения (<3). Вывод: распределение близко к логнормальному. | | | |
| Критерий Колмогорова-Смирнова | D=0.27538, p<0.01<br>С доверительной вероятностью 99% (>95%) распределение не является нормальным. | D=0.07634, p<0.01<br>С доверительной вероятностью 99% (>95%) распределение не является нормальным. | |
| Критерий Лиллиефорса | p<0.01<br>С доверительной вероятностью 99% (>95%) распределение не является нормальным. | p<0.01<br>С доверительной вероятностью 99% (>95%) распределение не является нормальным. | |
| Критерий Шапиро-Уилка | W=0.50379, p=0.0000<br>Распределение не является нормальным. | W=0.92647, p=0.0000<br>Распределение не является нормальным. | |
| Критерий хи-квадрат | 1885.79725, 3 степени свободы (корректированы), p = 0.00000<br>Распределение не является нормальным. | 582.58971, 5 степеней свободы (корректированы), p = 0.00000<br>Распределение не является нормальным. | |

| **Наличие аномальных значений** | | | |
|---|---|---|---|
| Минимально-нормальное значение (Mean-3SD) | -0.256 | -1.735 | |
| Максимально-нормальное значение (Mean+3SD) | 0.384 | -1.389 | |
| Количество аномальных точек отрицательных аномалий | 0 | 0 | |
| Количество аномальных точек положительных аномалий | 34 | 0 | |
| В случае принятия логнормальной модели с доверительной вероятностью 99% можно считать, что аномальные точки отсутствуют. Однако данная грубая оценка не учитывает реального распределения частот, которое можно оценить только с помощью гистограммы или графика на вероятностной бумаге, поэтому решение об исключении аномальных точек будет принято только после анализа соответствующих графиков. | | | |

**Ящики с усами (квартильный Min-Max, параметрический 3-сигма)**

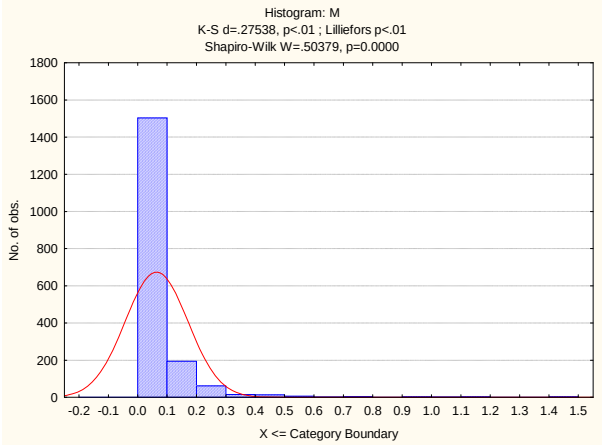
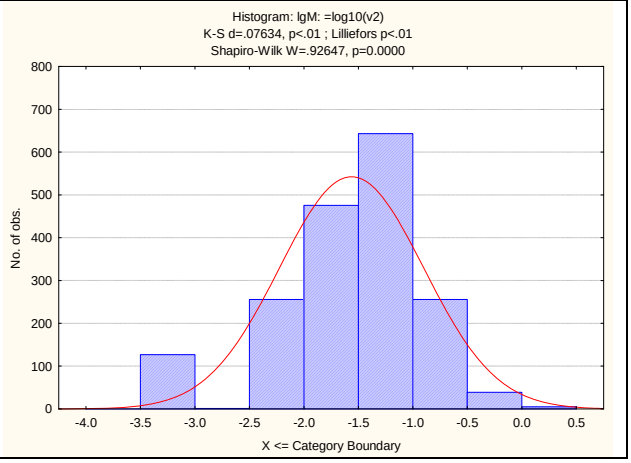
Графики представляют наглядное отображение параметров, уже обсужденных выше.

| | |
|---|---|
|  |  |
| Робастный ящик с усами (медиана, квартили, минимум и максимум) демонстрирует резкую правую асимметрию, характерную для логнормального распределения. | После логарифмирования ящик с усами становится симметричен, что указывает на высокую вероятность логнормального распределения исходных значений. |

| | |
|---|---|
|  |  |
| Параметрический ящик с усами (среднее, 2 и 3 сигма) демонстрирует наличие аномальных точек, что может быть вызвано логнормальным распределением и (или) наличием аномальных мод в области высоких содержаний. | После логарифмирования аномальные значения не наблюдаются, что с высокой вероятностью свидетельствует о логнормальности исходного распределения и об отсутствии существенных аномальных мод. |

**Гистограмма**

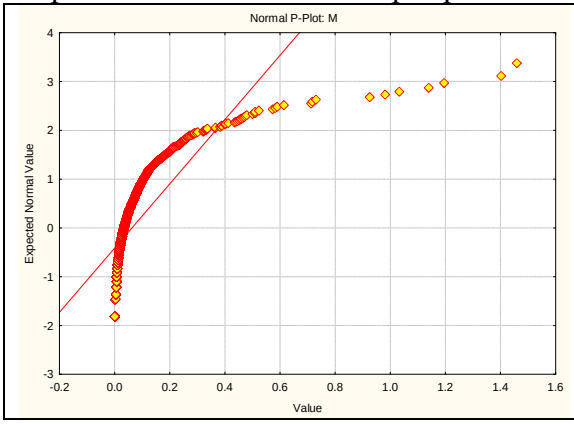
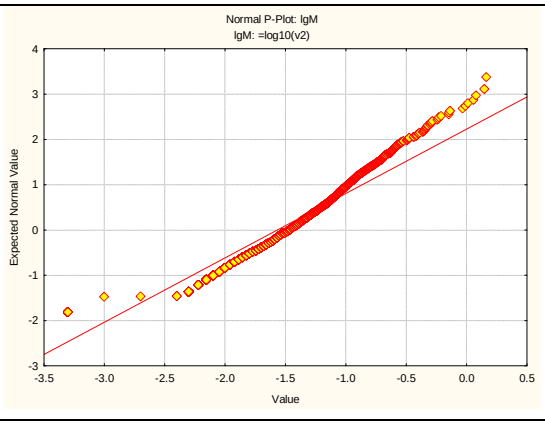
Графики представляют собой наглядное отображение параметров, уже обсужденных выше, а также дают информацию о реальном распределении частот, асимметрии, наличии дополнительных мод и т.п.

| | |
|---|---|
|  |  |

Логарифмирование приближает гистограмму к нормальному распределению.

**График на вероятностной бумаге**

График позволяет принять решение о соответствии основной массы наблюдений нормальной модели и выборе границ отсечения аномальных значений.

| | |
|---|---|
|  |  |

В исходных значениях наблюдается положительная (т.е. правая) асимметрия и сильное отклонение распределения от нормальной модели. После логарифмирования распределение приближается к нормальному, но сохраняется отклонение в области малых (<-2.5) и высоких (>0.2) значений, которые можно принять в качестве аномальных.

**Итоговые выводы.**
Доказательные выводы на основе изучения графиков. Выбор модели: нормальное или логнормальное.
**В результате сопоставления всех характеристик распределения сделан вывод о соответствии основного распределения логнормальному закону. Выявлены аномальные значения в области низких и (предположительно) высоких содержаний.**

**Удаление аномальных значений.**
Выбор минимально-нормального и максимально-нормального значений содержаний и их логарифмов (могут быть установлены визуально по графику на вероятностной бумаге).
**Границами нормального распределения являются логарифмы значений от -0.5 до 0.2, что соответствует содержаниям (антилогарифмы) 0.004 и 0.199 г/т.** Замечание: верхняя граница установлена только для примера, т.к. в данной выборке отсутствуют значения выше 0.199 г/т.

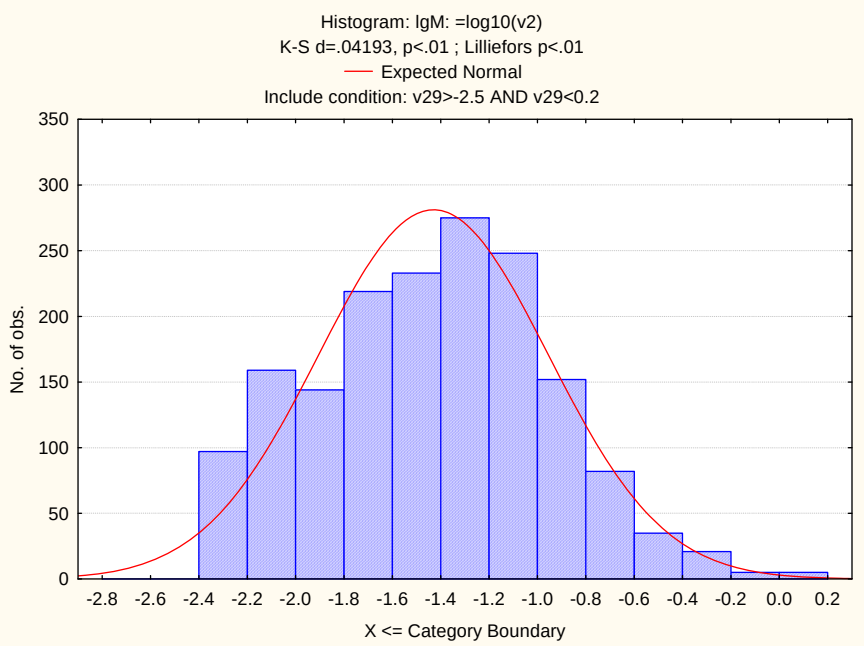
**Параметры очищенного распределения.**
Расчетные параметры, критерии нормальности и графики для очищенной выборки (с исключенными аномальными значениями) в той модели, которая была выбрана (нормальное – по содержаниям, логнормальное – по логарифмам содержаний).
Для логнормального распределения дать антилогарифмы необходимых параметров, а также рассчитать коэффициент вариации по содержаниям, но с удаленными аномальными значениями.

Очищенная выборка содержит 1675 значений (92.9% исходной выборки). 128 значений признаны аномальными (все в области низких значений). Параметры очищенной выборки:
Среднее арифметическое: -1.430 (антилогарифм 0.037 г/т).
95% доверительный интервал среднего: от -1.453 до -1.407 (0.035-0.039 г/т).
Стандартная ошибка: 0.012.
Медиана: -1.420 (антилогарифм 0.038 г/т).
Дисперсия: 0.226.
Стандартное отклонение: 0.475 (антилогарифм 2.99).
Минимальным значением является граница аномальных значений: -0.5 (0.004 г/т).
Максимальное значение соответствует исходной выборке: 0.164 (1.460 г/т).
Нижний квартиль: -1.796 (антилогарифм 0.016 г/т).
Верхний квартиль: -1.092 (антилогарифм 0.081 г/т).
Межквартильное расстояние: 0.704.
Робастная оценка стандартного отклонения: 0.522 (антилогарифм 3.33).
Стандартное отклонение, как меньшая величина, предпочтительно для оценки разброса.
Коэффициент вариации (по исходным значениям содержаний): 86.6% (не соответствует нормальной модели «< 33.3 %»).

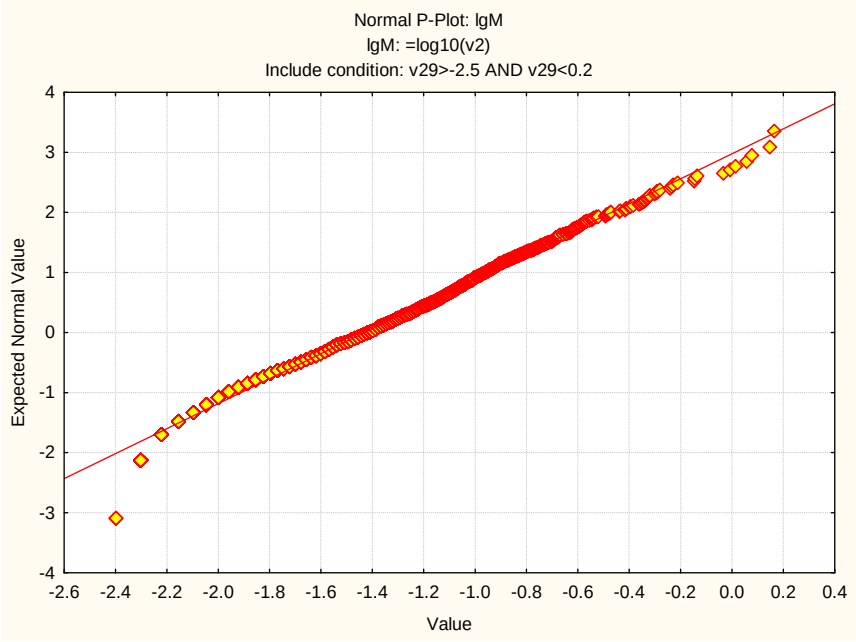
| | |
|---|---|
| Асимметрия | -0.164 |
| Стандартная ошибка асимметрии | 0.060 |
| Коэффициент асимметрии ($K_A$) | 2.74 < 3 |
| Эксцесс | -0.365 |
| Стандартная ошибка эксцесса | 0.120 |
| Коэффициент эксцесса ($K_E$) | 3.05 ≈ 3 |

Коэффициент асимметрии не превышает 3, коэффициент эксцесса незначительно превышает 3, следовательно, распределение логарифмов практически соответствует нормальной модели. Аналогичное заключение позволяют сделать гистограмма и график на вероятностной бумаге. Однако критерии согласия не соответствуют нормальной модели для доверительной вероятности 95%.

Критерий хи-квадрат: 250.85672 (5 степеней свободы после корректировки), p = 0.00000.




Наиболее существенное отличие распределения логарифмов от нормальной модели – наличие слабой дополнительной моды при -2.1 (бимодальное распределение).




**Вывод:** очищенная выборка близка к логнормальному распределению за исключением области низких значений, где достигнут нижний предел обнаружения (большая часть точек < НПО удалена). Логнормальная модель распределения содержаний может быть принята с достаточной надежностью.

Последние исправления внесены 24.02.2014 23:58:16. М.В.Морозов.