

«Шпаргалки для зоологов»

Шпаргалка №1

**Обработка и анализ  
количественных данных в  
зоологии**

Игорь Игоревич Дзевеверин

# Статистика

- Задача статистики: систематизация, обработка и использование количественных данных для научных и практических выводов
- Массовость данных. Поэтому от статистических методов требуются:
  - Информационная полнота описания
  - Компактность (незначительное количество элементов описания)
- Гипотетическая природа наших знаний. Они получены индуктивным путем; существует только определенная вероятность того, что они соответствуют действительности. Поэтому необходима количественная оценка вероятности: проверка соответствия принятых гипотез эмпирическим данным.

# Модели и гипотезы

|               | Принимаем $H_0$ | Принимаем $H_1$ |
|---------------|-----------------|-----------------|
| $H_0$ истинна | +               | $\alpha$        |
| $H_1$ истинна | $\beta$         | +               |

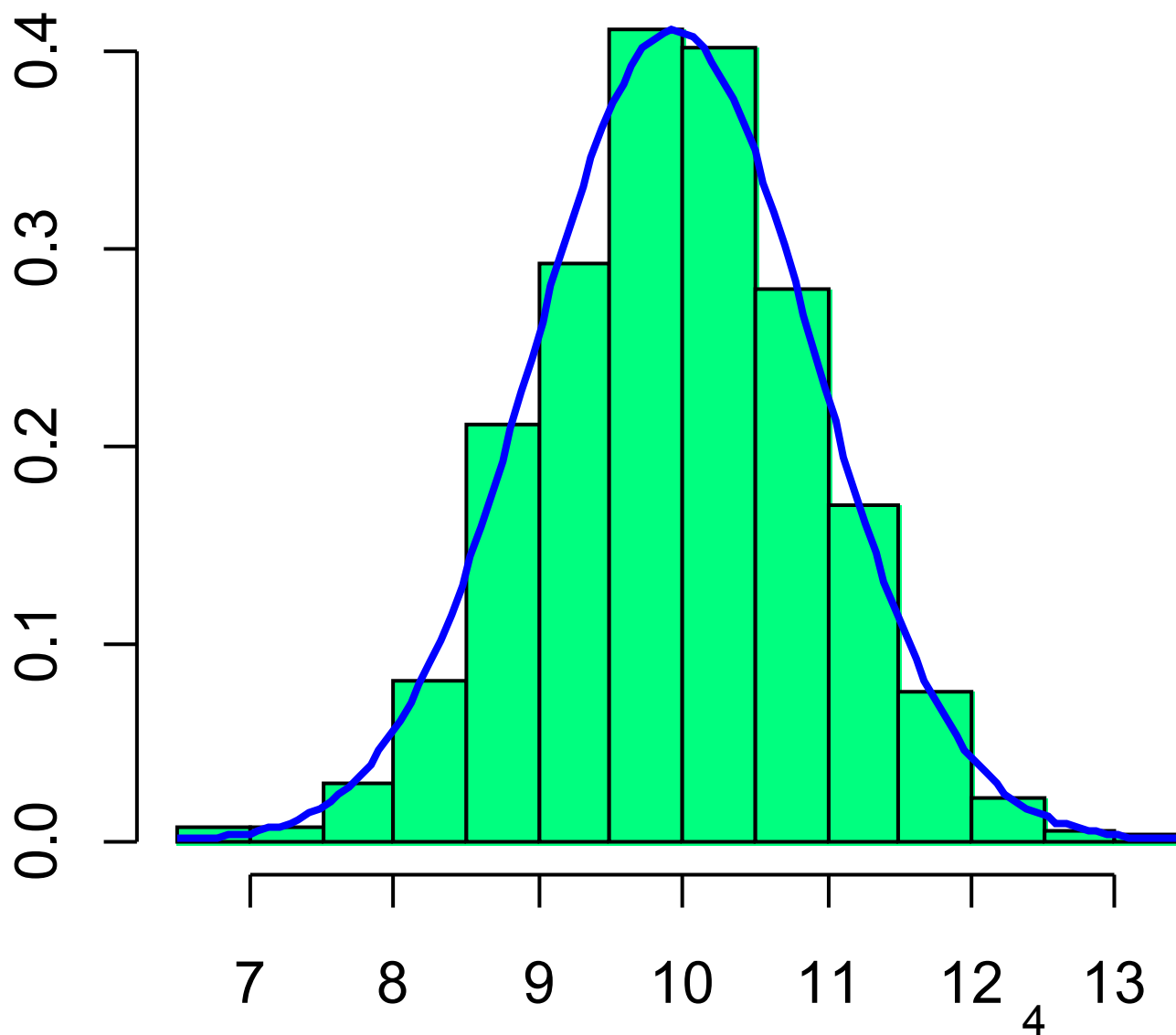
$H_0$  – нулевая гипотеза;  $H_1$  – альтернативная гипотеза

$p = \alpha$  – вероятность ошибки 1-го рода = уровень значимости

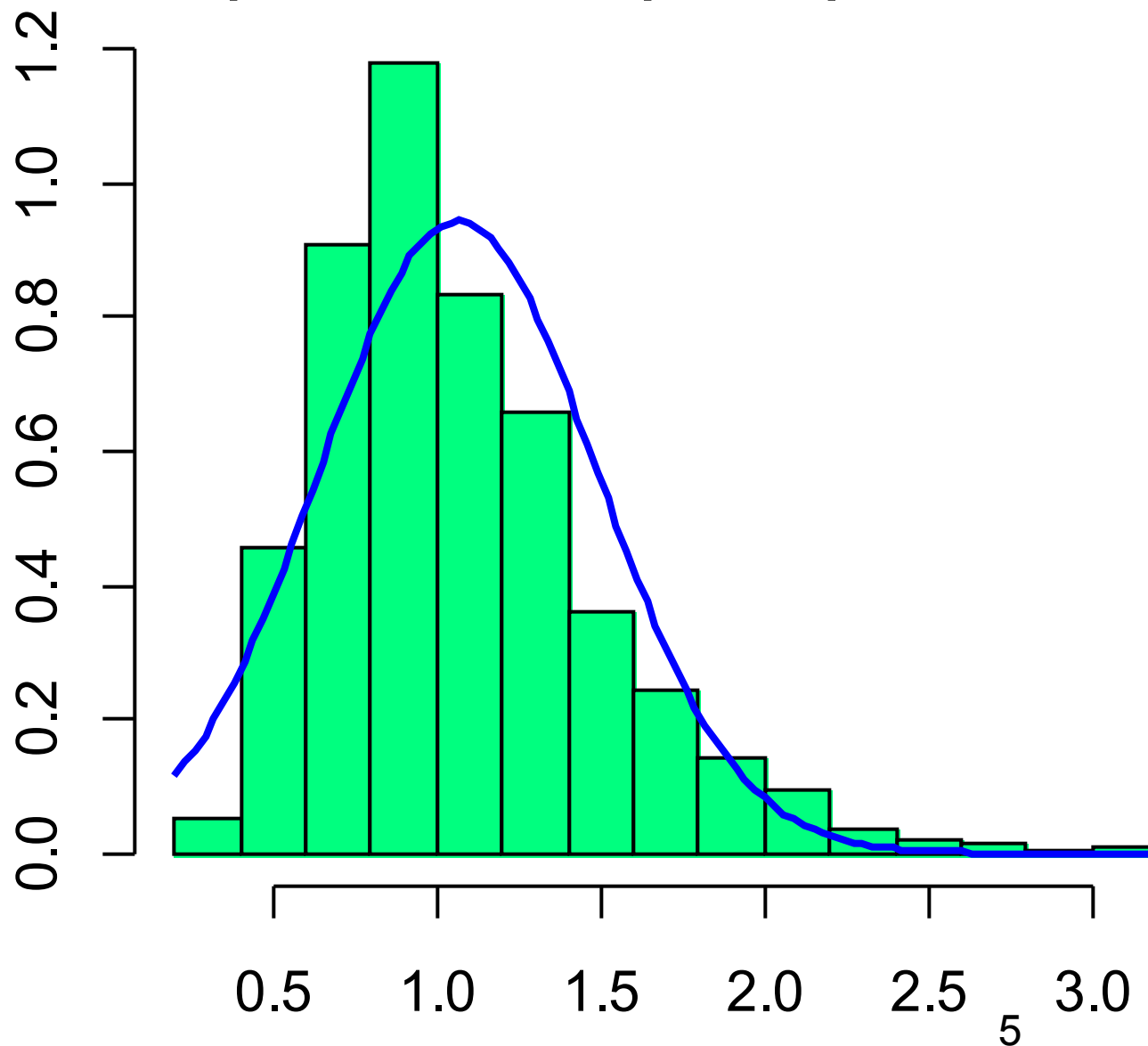
$P = 1 - \alpha$  – доверительная вероятность

Выбор между несколькими гипотезами:  
критерий Акаике и ряд других методов

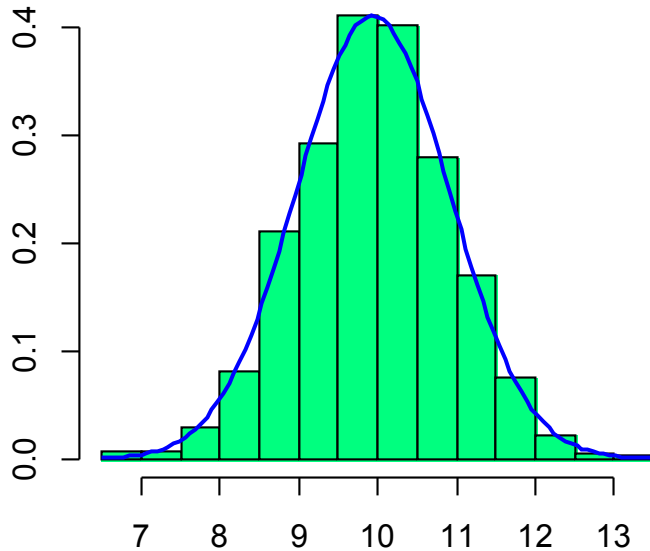
# Нормальное распределение



# Лог-нормальное распределение

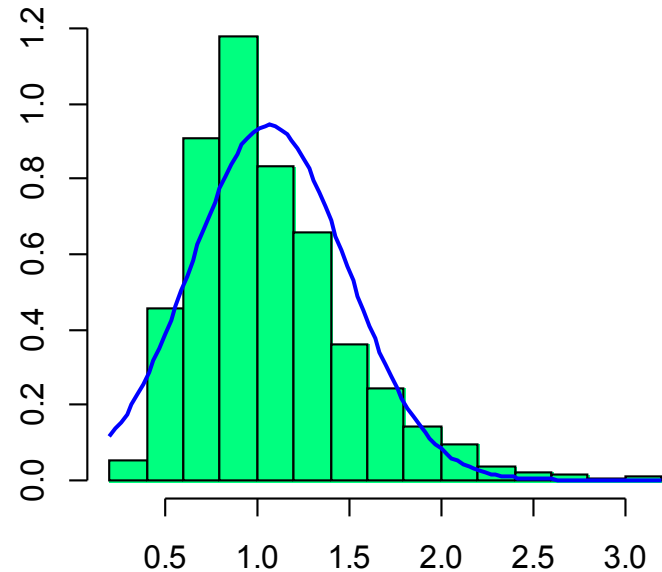


# Проверка нормальности распределения



Критерий Шапиро - Уилка

$W = 0.999$ ,  $p = 0.80$



$W = 0.94$ ,  $p < 2.2e-16$

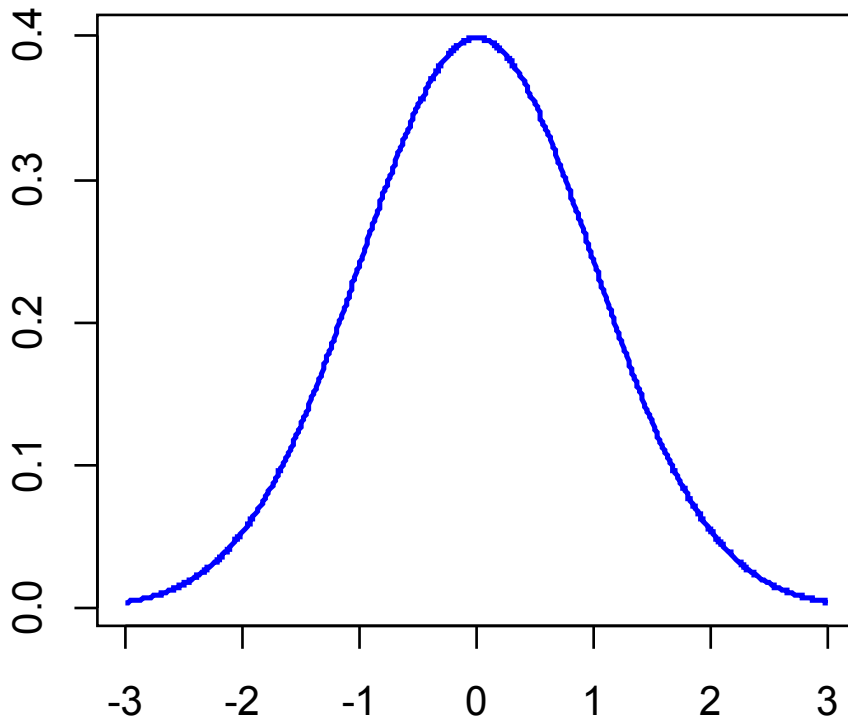
После лог-трансформации:

$W = 0.999$ ,  $p = 0.77$

# Параметрические и непараметрические методы

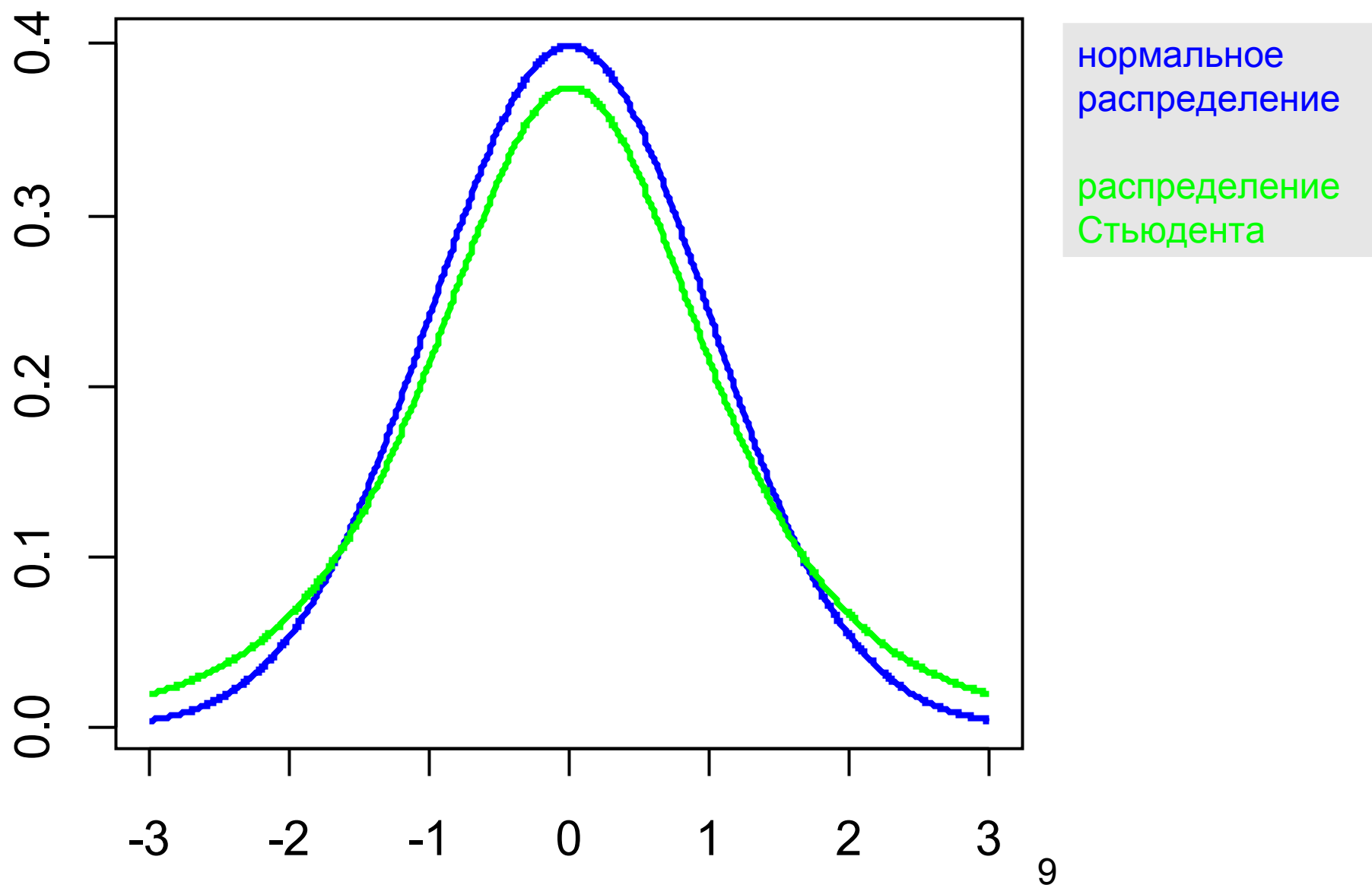
- Если распределение соответствует нормальному, используем параметрические методы
  - критерии Стьюдента и Фишера, дисперсионный анализ, корреляция Пирсона, линейный регрессионный анализ и ряд других методов
- Если распределение не соответствует нормальному, то мы можем применить
  - нормализующие трансформации (например, логарифмическую)
  - непараметрические методы (критерии Манна – Уитни, Уилкоксона, Краскелла – Уоллиса, корреляции Спирмана и Кендалла и др.)
  - процедуры ресемплинга (например, бутстрап или перестановочный тест).

# Нормальное распределение: оценки достоверности



- С вероятностью  $P=0,95$  случайная величина будет отстоять от среднего значения не более, чем на 1,96 стандартных отклонений.
- С вероятностью 0,99 - не более, чем на 2,58 стандартных отклонений.
- С вероятностью 0,999 – не более, чем на 3,29 стандартных отклонений.

# Распределение Стьюдента



# Распределение Стьюдента: малочисленные выборки

| df       | $p < 0,05$ | $p < 0,01$ |
|----------|------------|------------|
| 4        | 2,78       | 4,60       |
| 20       | 2,09       | 2,85       |
| 120      | 1,98       | 2,62       |
| $\infty$ | 1,96       | 2,58       |

На практике  $df \approx 50$  - очень малое отличие

# Оценка необходимого объема выборки

Стандартная ошибка –  $s_{\bar{X}} = \frac{s}{\sqrt{n}}$

Необходимая точность –  $\Delta = t \cdot s_{\bar{X}} = t \cdot \frac{s}{\sqrt{n}}$

Объем выборки –  $n = \frac{t^2 \cdot s^2}{\Delta^2}$

Пример:

Измерение погрешности измерения штангенциркулем.

$S = 0,108$ .  $t = 1,96$ .  $\Delta = 0,1$ .

Получим  $n = \left( \frac{1,96 \cdot 0,108}{0,1} \right)^2 \approx 4,48$

Нужно измерять 5 раз и брать среднее, чтобы получить точность 0,1 мм на 5%-ном уровне значимости.

# Шкалы измерения

- *Номинальная шкала:* объекты отличаются только именами, порядок не имеет значения.
- *Порядковая шкала:* значения переменной имеют четко определенный порядок.
- *Интервальная шкала:* можно определить, насколько одно значение больше, чем другое.
- *Шкала отношений:* есть фиксированное начало координат; можно определить, во сколько раз одно значение больше, чем другое.

# Некоторые рекомендации по составлению таблиц первичных данных

- Данные представлять в удобной для машинной обработки форме.
- Количественные данные одной размерности приводить в одной шкале.
- Там, где возможно, объединять данные из разных столбцов в одном столбце.
- Следить, чтобы таблица имела как можно меньше пропущенных значений.
- Не путать нуль и пропуск.
- С осторожностью подходить к удалению выбросов.
- Округлять данные только по завершении расчетов.
- Изучать методы векторного и матричного анализа – они необходимы для обработки табличных данных.

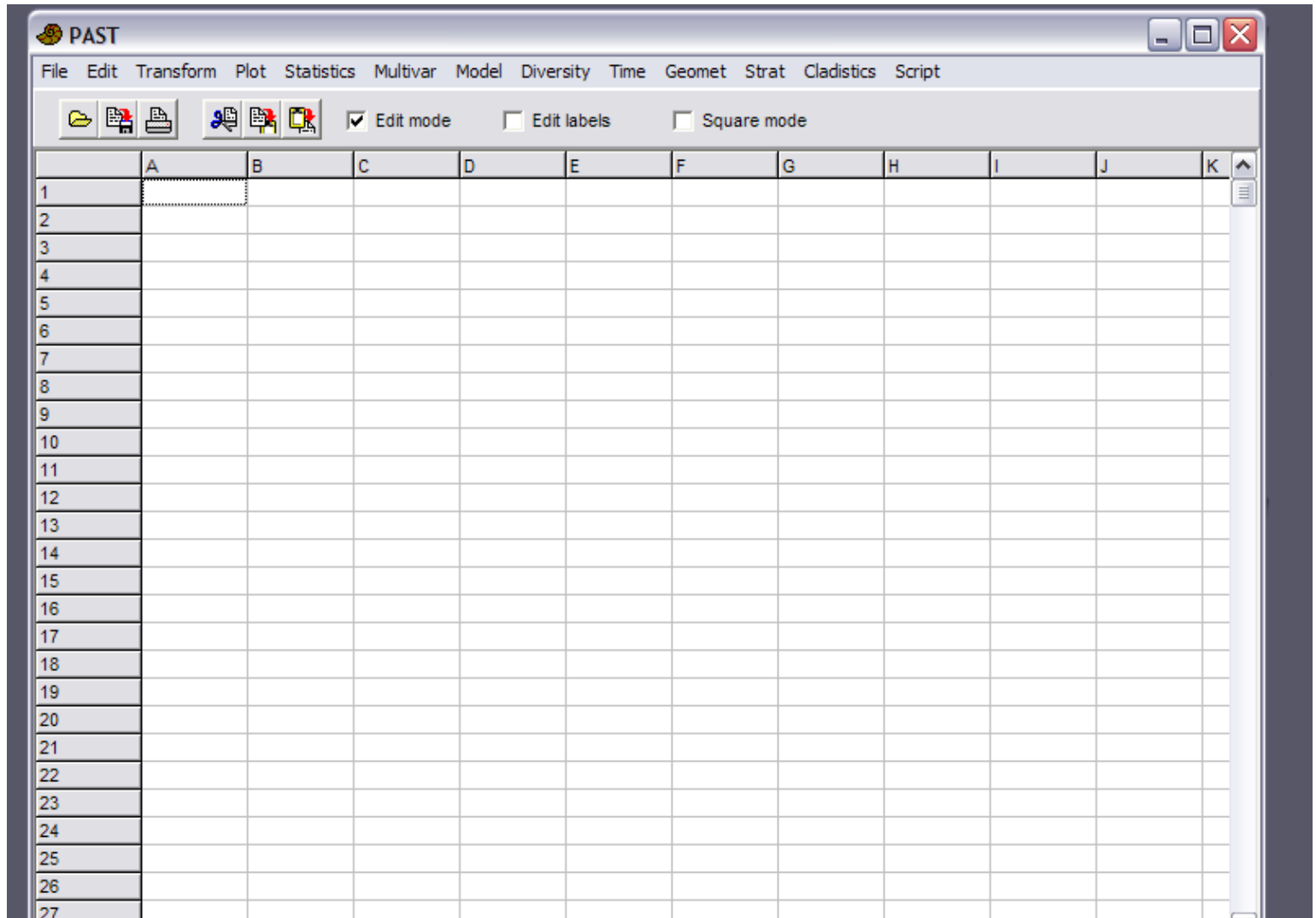
# Некоторые программы статистической обработки данных

|                        | Проприетарные         | Свободные |
|------------------------|-----------------------|-----------|
| Для математиков        | MATHEMATICA,<br>MAPLE | MAXIMA    |
| Для инженеров          | MATLAB                | Scilab    |
| Для статистиков<br>(1) | SAS, S-PLUS           | R         |
| Для статистиков<br>(2) | SPSS,<br>STATISTICA   | PAST      |

В таблицу не включены неспециализированные программы работы с электронными таблицами (Excel, Calc, Gnumeric и др.) и, наоборот, узкопрофильные программы (геоинформационные системы, геометрическая морфометрия, филогенетический анализ и др.).



# PAST – Paleontological Statistics





# R: A language and environment for statistical computing

```
RGui (32-bit)
File History Resize Windows

R Console

R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> set.seed(405)
> y <- rlnorm(1200,0,0.4)
> hist(y,prob=TRUE,col='springgreen',main="",xlab="",ylab="")
> curve(dnorm(x,mean=mean(y),sd=sd(y)),add=TRUE,col="blue",lwd=2.5)
> shapiro.test(y)

      Shapiro-Wilk normality test

data:  y
W = 0.9374, p-value < 2.2e-16

> shapiro.test(log(y))

      Shapiro-Wilk normality test

data:  log(y)
W = 0.999, p-value = 0.7655
```

