Cộng hòa xã hội chủ nghĩa Việt Nam

Độc lập – Tự do – Hạnh phúc

**Kính gởi:**

**Đại học Quốc Gia Tp Hồ Chí Minh**

**Viện Toán học Việt Nam**

**Thuyết minh dự án**

**Dự án Phát triển nền tảng thống kê nguồn mở**

**1. Tổng quan dự án**

* *Tên dự án*: Dự án phát triển nền tảng (framework) thống kê nguồn mở
* *Tổ chức đăng ký*: Cty TNHH Lập trình Hướng Dương
  + *GPKD*: 0311286418
  + *Địa chỉ*: 42/2 Nguyễn Đình Khơi, phường 4, quận Tân Bình, Tp. Hồ Chí Minh
  + *Điện thoại*: 0975250362
* *Dự kiến kinh phí*: 5,49 tỷ
* *Dự kiến thời gian*: 55 tháng

**2. Nội dung dự án**

**2.1. Tính cấp thiết:**

Thống kê là ngành toán ứng dụng, hầu như xuất hiện trong nhiều ngành từ khoa học công nghệ đến kinh tế giáo dục, rất cần thiết trong quá trình phát triển kinh tế xã hội. Khi xử lý dữ liệu thống kê khổng lồ, sự hỗ trợ của phần mềm máy tính là một nhu cầu tất yếu. Qua khảo sát, hiện trên thế giới có rất nhiều tổ chức và cá nhân phát triển các phần mềm thống kê từ đơn giản đến phức tạp, từ tổng quát đến chuyên sâu. Các phần mềm nguồn mở đang vượt trội so với phần mềm thương mại, trước tiên do miễn phí sau đến tính mở dễ phát triển và mở rộng. Tuy nhiên có những điểm lưu ý sau:

1. Một số phần mềm không phát triển đầy đủ 5 chủ đề chính của thống kê: ước lượng, kiểm định giả thuyết, hồi quy, phân tích phương sai và quản lý chất lượng. Một số phần mềm chuyên dụng trong một lĩnh vực, không có tính tổng quát.
2. Hầu như các phần mềm đều hỗ trợ giao diện người dùng cuối, đây là ưu điểm cũng nhược điểm. Khi hỗ trợ giao diện người dùng cuối, phần mềm rất dễ sử dụng đối với người dùng nhưng rất khó cho nhà phát triển mở rộng tính năng hoặc xây dựng phần mềm mới từ phần mềm đó.
3. Phần mềm R bên cạnh hỗ trợ giao diện người dùng cuối còn cung cấp một ngôn ngữ lập trình để mở rộng tính năng. Đây là điểm xuất sắc của R nhưng có một nhược điểm các nhà triển chỉ có thể mở rộng R hay xây dựng các linh kiện chức năng chạy trên nền R, họ không thể phát triển R hoặc tích hợp chính R vào những ứng dụng chuyên biệt. Điều này có nghĩa là một ứng dụng nào đó có sử dụng R thì mất đi tính thuần nhất và rất khó bảo trì trừ những ứng dụng phân tán hay đa tạp tích hợp nhiều nhiều nền, nhiều ngôn ngữ. Hơn nữa ngôn ngữ R không đủ sức mạnh và tính năng như các ngôn ngữ lập trình chuyên dụng như Java, C#, C++.
4. Các phần mềm chưa hỗ trợ thống kê dữ liệu nhiều chiều đặc biệt trong trường hợp dữ liệu bị thiếu.

Với những điểm lưu ý trên, một phần mềm thống kê với tính năng chính vừa là một nền tảng (framework, platform) như R vừa là một thư viện nguồn mở hỗ trợ cả người dùng cuối và nhà phát triển là một nhu cầu tất yếu khi thống kê đã phát triển sâu rộng vào tin học. Đề án này như là một sự thúc đẩy cho các nhà phát triển xây dựng các phần mềm thống kê mới trên nền tảng này với những ứng dụng chuyên biệt như y tế, dịch tễ, tài chính, môi trường, sinh học…

**2.2. Mục tiêu dự án**

Dự án gồm 3 mục tiêu:

1. Phát triển một nền tảng thống kê hỗ trợ bao quát 5 lĩnh vực chính của thống kê: ước lượng, kiểm định giả thuyết, hồi quy, phân tích phương sai và quản lý chất lượng.
2. Nền tảng cũng là một thư viện mã nguồn giúp các chuyên gia tin học và thống kê phát triển một phần mềm thống kê mới. Các nhà phát triển chỉ cần thông thạo một ngôn ngữ lập trình như Java hay C#, không cần học ngôn ngữ mới.
3. Như đã đề cập, dự án là một sự thúc đẩy lĩnh vực phát triển phần mềm thống kê nói riêng và khoa học thống kê nói chung. Vì vậy, dự án sẽ đi kèm cùng với một bộ sách thống kê làm kim chỉ nam cho lộ trình phát triển. Người dùng, nhà phát triển có thể đối chiếu giữa phần mềm và bộ sách để hiểu rõ thêm về tính ứng dụng và triết lý thống kê đằng sau những con số.
4. Hỗ trợ ước lượng tham số và kiểm định giả thuyết đối với dữ liệu nhiều chiều. Dữ liệu nhiều chiều càng trở nên phổ biến đặc biệt trong ngành sinh học và tài chính. Phân tích dữ liệu nhiều chiều là một quá trình phức tạp liên quan đến đại số ma trận và giải tích nên các phần mềm thống kê ít hỗ trợ ước lượng tham số và kiểm định giả thuyết nhiều chiều. Dự án có tham vọng giải quyết những vấn đề liên quan dữ liệu nhiều chiều mà không phá vỡ tính thống nhất với thống kê dữ liệu một chiều. Do đó quá trình hoạch định kiến trúc hệ thống rất quan trọng. Hơn nữa sự thiếu hụt mẫu trong dữ liệu nhiều chiều thường xảy ra do cấu trúc dữ liệu nhiều chiều phức tạp, quá trình thu thập mẫu dễ xảy ra sai sót. Phần mềm sẽ xử lý dữ liệu bị thiếu bằng sự kết hợp giữa thống kê và máy học (machine learning), trí tuệ nhân tạo.

Với 3 mục tiêu trên, thay vì đứng trên vai người khổng lồ bằng việc sử dụng các phần mềm nguồn mở như R, chúng ta tự biến mình thành người khổng lồ để người khác đứng trên vai.

**2.3. Nội dung thực hiện và kết quả**

Dự án tập trung viết một bộ sách hai tập cùng xây dựng một nền tảng (thư viện) phần mềm thống kê. Tập I bộ sách khảo sát các chủ đề liên quan đến ước lượng tham số gồm 6 chương không kể chương đầu gồm những khái niệm cơ bản.

1. Các khái niệm cơ bản về xác suất thống kê.
2. Mẫu và thống kê mẫu.
3. Ước lượng điểm (point estimation) và các pp ước lượng điểm như: MLE, MAP, moment…
4. Ước lượng khoảng (interval estimation).
5. Ước lượng Bayesian: MAP, EM,…
6. Phương pháp Monte Carlo.
7. Ước lượng tham số với dữ liệu không đầy đủ (Estimation of incomplete data).

Tập II bộ sách khảo sát các chủ đề liên quan đến kiểm định giả thuyết gồm 8 chương cùng một chương mở rộng.

1. Kiểm định giả thuyết với trung bình mẫu (sample mean), phương sai mẫu (sample variance), tỉ lệ mẫu (sample proportion). Kiểm định giả thuyết với hai thống kê mẫu. Các chủ đề liên quan đến xác định cỡ mẫu, độ lỗi *α* và độ lỗi *β*.
2. Lý thuyết hồi quy gồm: hồi quy một biến, hồi quy nhiều biến, hồi quy logistics. Các kiểm định đối với hệ số hồi quy. Hồi quy phi tuyến là phần mở rộng.
3. Phân tích phương sai (ANOVA) với một yếu tố (factor) và nhiều yếu tố.
4. Kiểm định phi tham số.
5. Kiểm định giải thuyết bằng phương pháp Bayesian
6. Kiểm định giả thuyết với dữ liệu nhiều chiều.
7. Quản lý chất lượng (quality control).
8. Phân tích dữ liệu nhiều chiều như PCA, SVD, discriminant analysis, canonical correlation analysis. Chương 14 là chương mở rộng.

Phần mềm thống kê gồm 9 modules:

1. Mẫu và thống kê mẫu đối với dữ liệu vô hướng và đa chiều
2. Ước lượng tham số
3. Kiểm định giả thuyết có tham số và phi tham số
4. Hồi quy.
5. Phân tích phương sai.
6. Kiểm định giả thuyết đối với dữ liệu nhiều chiều.
7. Quản lý chất lượng
8. Phân tích dữ liệu nhiều chiều.
9. Giao diện người dùng và trình phân tích cú pháp dòng lệnh.

Trong module thứ 9, trình phân tích cú pháp dòng lệnh có vai trò như shell trong các hệ điều hành cho phép người dùng, nhà phát triển tương tác và lập trình qua dòng lệnh. Phần mềm cung cấp một dạng lập trình script (kịch bản). Trong bản thuyết minh này, chúng ta dùng thuật ngữ module để chỉ một thành tố (component) hay đơn vị phần mềm, nhiều module hợp thành phần mềm phức hợp. Module khi hiện thực hóa, có thể là một gói Java, một .net assembly. Module là chương tương ứng trong sách.

Thư viện đại số ma trận + giải tích + xác suất + máy học

Mẫu và thống kê mẫu

Ước lượng tham số

Kiểm định giả thuyết

Hồi quy

Phân tích phương sai

Quản lý chất lượng

Phân tích dữ liệu nhiều chiều

Kiểm định giả thuyết nhiều chiều

Giao diện người dùng và Shell

*Kiến trúc tổng quan phần mềm thống kê*

Tóm lại, kết quả dự án gồm 3 sản phẩm:

* Nền tảng thống kê hỗ trợ bao quát 5 lĩnh vực chính của thống kê: ước lượng, kiểm định giả thuyết, hồi quy, phân tích phương sai và quản lý chất lượng. Trong đó thống kê dữ liệu nhiều chiều và phân tích dữ liệu nhiều chiều là giá trị gia tăng của sản phẩm.
* Tài liệu hướng dẫn sử dụng và phát triển phần mềm
* Bộ sách gồm hai tập khảo sát ước lượng và kiểm định thống kê, trong đó những công trình và phát kiến mới được đưa vào bộ sách như là giá trị gia tăng.

**3. Phương án triển khai**

**3.1. Phương án thực hiện**

Dự án gồm nhiều module phức tạp liên kết nhau nên trong phương án triển khai, bước thiết kế rất quan trọng. Dự án gồm 37 bước được thực hiện trung bình trong 55 tháng.

|  |  |  |
| --- | --- | --- |
| **Bước** | **Nội dung** | **Thời gian** (tháng) |
| 1 | Tổng thiết kế phối hợp chương sách và module phần mềm. Kết quả là bản tổng thiết kế. | 1 |
| 2 | Phác họa các chương sách. | 1 |
| 3 | Thiết kế phần mềm bao gồm: môi trường phát triển, các module, use case, *test case*, giao diện, quản lý dự án, time line, etc. | 2 |
| 4 | Xây dựng sơ đồ lớp. | 3 |
| 5 | * Viết phần tóm tắt (abstract) 14 chương sách. * Bắt đầu viết tài liệu hướng dẫn cho module 1. | 3 |
| 6 | Cài đặt toàn bộ môi trường phát triển phần mềm cho dự án bao gồm: công cụ lập trình, trình chủ quản lý phiên bản (svn), hoạch định các gói (package), assembly. | 1 |
| 7 | * Viết chương 1: Mẫu và thống kê mẫu. * Phát triển module 1: Mẫu và thống kê mẫu đối với dữ liệu vô hướng và đa chiều. | 1 |
| 8 | * Hoàn thành và kiểm thử module 1. * Hoàn thành tài liệu hướng dẫn cho module 1, bắt đầu viết tài liệu hướng dẫn cho module 2. | 1 |
| 9 | * Viết chương 2, 3: Ước lượng điểm và ước lượng khoảng. * Phát triển module 2: Ước lượng tham số. | 2 |
| 10 | * Viết chương 4: Ước lượng Bayesian. * Nâng cấp module 2: Hỗ trợ ước lượng Bayesian. | 1 |
| 11 | * Viết chương 5: Phương pháp Monte Carlo. * Nâng cấp module 2: Cài đặt giải thuật Monte Carlo. | 2 |
| 12 | * Viết chương 6: Ước lượng tham số với dữ liệu không đầy đủ. Kết thúc tập I. * Nâng cấp module 2: Hỗ trợ ước lượng ước lượng tham số với dữ liệu không đầy đủ. | 2 |
| 13 | * Duyệt lại tập I bộ sách, chỉnh sửa và bổ sung. * Hoàn thành và kiểm thử module 2. * Hoàn thành tài liệu hướng dẫn cho module 2, bắt đầu viết tài liệu hướng dẫn cho module 3. | 1 |
| 14 | * Viết chương 7: Kiểm định giả thuyết và những chủ đề liên quan. Đây là chương cơ sở cho kiểm định giả thuyết, rất quan trọng. * Phát triển một phần module 3: Kiểm định giả thuyết có tham số với dữ liệu một chiều. | 1 |
| 15 | * Viết tiếp chương 7: Kiểm định giả thuyết và những chủ đề liên quan. * Nâng cấp và kiểm thử một phần module 3: Hỗ trợ công thức xác định cỡ mẫu, độ lỗi *α* và độ lỗi *β*. * Hoàn thành tài liệu hướng dẫn cho module 2, bắt đầu viết tài liệu hướng dẫn cho module 4. | 1 |
| 16 | * Viết chương 8: Lý thuyết hồi quy, phần một biến. * Phát triển module 4 (lưu ý module 3 chưa kết thúc): Phần hồi quy một biến và kiểm định giả thuyết các hệ số hồi quy một biến. | 1 |
| 17 | * Viết tiếp chương 8: Lý thuyết hồi quy, phần đa biến và logistic. * Nâng cấp module 4: phần hồi quy đa biến và logistic. Kiểm định giả thuyết các hệ số hồi quy đa biến biến. | 2 |
| 18 | * Chỉnh sửa và bổ sung chương 8. * Hoàn thành và kiểm thử module 4. * Hoàn thành tài liệu hướng dẫn cho module 4. Bắt đầu viết tài liệu hướng dẫn cho module 5. Lưu ý, tài liệu hướng dẫn cho module 3 chưa kết thúc. | 1 |
| 19 | * Viết chương 9: Phân tích phương sai (ANOVA), một yếu tố (factor). * Phát triển module 5: Phân tích phương sai hỗ trợ một yếu tố. | 1 |
| 20 | * Viết tiếp chương 9: Phân tích phương sai (ANOVA), nhiều yếu tố. * Nâng cấp module 5: Phân tích phương sai hỗ trợ nhiều yếu tố. | 2 |
| 21 | * Chỉnh sửa và bổ sung chương 9. * Hoàn thành và kiểm thử module 5. * Hoàn thành tài liệu hướng dẫn cho module 5. Viết tiếp tài liệu hướng dẫn cho module 3. | 1 |
| 22 | * Viết chương 10: Kiểm định giả thuyết phi tham số. * Nâng cấp module 3: Kiểm định giả thuyết phi tham số với dữ liệu một chiều. | 2 |
| 23 | * Viết chương 11: Kiểm định giả thuyết bằng phương pháp Bayesian. * Nâng cấp module 3: Kiểm định giả thuyết bằng phương pháp Bayesian. | 2 |
| 24 | * Chỉnh sửa và bổ sung chương 10 và 11. * Hoàn thành và kiểm thử module 3, phần kiểm định giả thuyết phi tham số và kiểm định giả thuyết bằng phương pháp Bayesian. * Hoàn thành tài liệu hướng dẫn cho module 3. Bắt đầu viết tài liệu hướng dẫn cho module 6. | 1 |
| 25 | * Viết chương 12: Kiểm định giả thuyết với dữ liệu nhiều chiều. * Phát triển module 6: Kiểm định giả thuyết với dữ liệu nhiều chiều. | 2 |
| 26 | * Viết tiếp chương 12: Kiểm định giả thuyết với dữ liệu nhiều chiều trong trường hợp phi tham số. * Nâng cấp module 6: Kiểm định giả thuyết với dữ liệu nhiều chiều trong trường hợp phi tham số. | 2 |
| 27 | * Chỉnh sửa và bổ sung chương 12. * Hoàn thành và kiểm thử module 6. * Hoàn thành tài liệu hướng dẫn cho module 6. Bắt đầu viết tài liệu hướng dẫn cho module 7. | 1 |
| 28 | * Viết chương 13: Quản lý chất lượng. * Phát triển module 7: Hỗ trợ quản lý chất lượng. | 1 |
| 29 | * Chỉnh sửa và bổ sung chương 13. * Hoàn thành và kiểm thử module 7. * Hoàn thành tài liệu hướng dẫn cho module 7 và bắt đầu viết tài liệu hướng dẫn cho module 8. | 1 |
| 30 | * Viết chương 14: Phân tích dữ liệu nhiều chiều. * Phát triển module 8: Phân tích dữ liệu nhiều chiều. | 2 |
| 31 | * Viết tiếp chương 14: Phân tích dữ liệu nhiều chiều. * Nâng cấp module 8: Phân tích dữ liệu nhiều chiều. | 2 |
| 32 | * Chỉnh sửa và bổ sung chương 14. Kết thúc tập 2 bộ sách. * Hoàn thành và kiểm thử module 8. * Hoàn thành tài liệu hướng dẫn cho module 8. Bắt đầu viết tài liệu hướng dẫn cho module 9. | 1 |
| 33 | * Phát triển module 9: Bộ phân tích cú pháp dòng lệnh. * Duyệt lại tập II bộ sách, chỉnh sửa và bổ sung. | 1 |
| 34 | * Nâng cấp module 9: Shell và bộ phân tích cú pháp dòng lệnh. * Nâng cấp module 9: Giao diện người dùng. | 1 |
| 35 | * Nâng cấp module 9: Shell và bộ phân tích cú pháp dòng lệnh. * Nâng cấp module 9: Giao diện người dùng | 1 |
| 36 | * Hoàn thành và kiểm thử module 9. * Hoàn thành tài liệu hướng dẫn cho module 9. | 1 |
| 37 | * Tổng kiểm thử phần mềm và sách, phối hợp tương ứng module và chương. * Đóng gói sản phẩm, hoàn thành dự án. | 3 |
| **Tổng cộng** | | 55 |

Thiết kế chiếm tỉ trọng 20%, thi công chiếm tỉ trọng 48%, kiểm thử chiếm tỉ trọng 32%. Bước 1 – 6 tập trung vào thiết kế, bước 7 – 36 là sự phối hợp giữa thi công và kiểm thử, bước 37 tập trung vào kiểm thử. Mô hình phát triển là sự kết hợp giữa mô hình thác nước (phân tích thiết kế - lập trình – kiểm thử) và mô hình agile (lập trình và kiểm thử song hành từ bước 7 – 36) nhằm đảm bảo tính thống nhất giữa thiết kế – thi công và bảo đảm tốc độ lẫn hiệu quả của thi công.

**3.2. Dự toán kinh phí**

Chúng ta phân tích kinh phí dự án theo đầu người, với khối lượng công việc rất nhiều từ kiến trúc tổng thể đến nghiên cứu, viết sách, viết tài liệu, lập trình và kiểm thử, số lượng người tối thiểu phải là 10, phân bổ như sau:

* 1 kiến trúc sư kiêm quản trị dự án và quản trị hệ thống. Quản trị hệ thống là quản lý phiên bản, mã nguồn, tài liệu, cơ sở dữ liệu. Kiến trúc sư này ngoài việc cùng nhà khoa học hoạch định kiến trúc tổng thể còn theo dõi xuyên suốt quá trình phát triển và triển khai dự án, phát hiện kịp thời những lỗ hổng trong thi công.
* 1 nhà khoa học chuyên nghiên cứu, viết sách và viết báo kiêm đề xuất thuật toán. Bộ sách hai tập chủ yếu do nhà khoa học viết.
* 1 chuyên viên viết tài liệu. Ngoài việc viết tài liệu, chuyên viên này còn hỗ trợ nhà khoa học viết sách.
* Tối thiểu 2 chuyên viên kiểm thử. Hai chuyên viên này chỉ tập trung vào kiểm thử.
* Tối thiểu 5 lập trình viên xuất sắc. Bốn người trong đó tập trung vào lập trình, một người hỗ trợ hai người kiểm thử.

Giả sử mức lương tối thiểu cho lập trình viên xuất sắc là 8 triệu VNĐ / tháng, mức lương chuyên viên viết tài liệu và kiểm thử là 6 triệu VNĐ / tháng, mức lương nhà khoa học là 16 triệu / tháng, quản trị dự án là 24 triệu / tháng. Tổng lương cho 10 người là 98 triệu VNĐ / tháng. Do đó tổng giá trị dự án là 55 \* 98 triệu + 10 \* 10 triệu / máy tính = 5,49 tỷ VNĐ, với 10 máy tính khấu hao hết trong 55 tháng. Người ta thường bảo tinh nhuệ là tối ưu nhưng thực tế số lượng có thể bù vào tinh nhuệ nhưng tinh nhuệ luôn nhỉnh hơn do số lượng đến một lúc sẽ bão hòa, nếu hơn nữa sẽ phản tác dụng và số lượng, ví dụ tăng *k* lần, thì hiệu quả công việc luôn tăng với tỉ lệ nhỏ hơn *k* đáng kể cho dù có thêm yếu tố thời gian, không hẳn do chi phí quản lý tăng mà chủ yếu do trong quá trình giao thoa công việc sẽ phát sinh phản ứng phụ. Vì vậy nếu số lượng người giảm đi *k* lần thì giá trị dự án tăng ít nhất *k* lần và ngược lại.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Bước** | **Nội dung chi** | **Chi phí**  **(triệu)** | **Tỉ lệ** |
| 1 | 1 | Tổng thiết kế phối hợp chương sách và module phần mềm. Kết quả là bản tổng thiết kế. | 100 | 1,8% |
| 2 | 2 | Phác họa các chương sách. | 50 | 0,9% |
| 3 | 3 | Thiết kế phần mềm bao gồm: môi trường phát triển, các module, use case, giao diện, quản lý dự án, tiến độ, etc. | 200 | 3,6% |
| 4 | 4 | Xây dựng sơ đồ lớp. | 400 | 7,3% |
| 5 | 5 | * Viết phần tóm tắt (abstract) 14 chương sách. * Bắt đầu viết tài liệu hướng dẫn cho module 1. | 300 | 5,5% |
| 6 | 6 | Cài đặt toàn bộ môi trường phát triển phần mềm cho dự án bao gồm: công cụ lập trình, trình chủ quản lý phiên bản, quản lý kiểm thử, quản lý tiến độ. Hoạch định các gói (package) và assembly. | 50 | 0,9% |
| 7 | 7, 8 | Viết chương 1 và phát triển module 1. | 200 | 3,6% |
| 8 | 9, 10, 11, 12, 13 | Viết chương 2, 3, 4, 5, 6 và phát triển module 2. | 800 | 14,6% |
| 9 | 14, 15 | Viết chương 7 và phát triển một phần module 3. | 200 | 3,6% |
| 10 | 16, 17, 18 | Viết chương 8 và phát triển module 4. | 400 | 7,3% |
| 11 | 19, 20, 21 | Viết chương 9 và phát triển module 5. | 400 | 7,3% |
| 12 | 22, 23, 24 | Viết chương 10, 11 và nâng cấp module 3. | 500 | 9,1% |
| 13 | 25, 26, 27 | Viết chương 12 và phát triển module 6. | 500 | 9,1% |
| 14 | 28, 29 | Viết chương 13 và phát triển module 7. | 200 | 3,6% |
| 15 | 30, 31, 32 | Viết chương 14 và phát triển module 8 | 500 | 9,1% |
| 16 | 33, 34, 35, 36 | Phát triển module 9. | 400 | 7,3% |
| 17 | 37 | Tổng kiểm thử và đóng gói sản phẩm, hoàn thành dự án. | 290 | 5,4% |
|  |  | **Tổng cộng** | **5490** | 100% |

Thiết kế chiếm tỉ trọng 20% với kinh phí 1,1 tỷ.

Thi công chiếm tỉ trọng 48% với kinh phí 2,6 tỷ.

Kiểm thử chiếm tỉ trọng 32% với kinh phí 1,79 tỷ.

**3.3. Phân bổ kinh phí**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tháng**  (triệu) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | **Tổng** |
| 100(1) | 50(2) | 100(3) | 100(3) | 150(4) | 150(4) | 100(4) | 100(5) | 100(5) | 100(5) | 50(6) | 100(7) | **1200** |
| 1,8% | 0,9% | 1,8% | 1,8% | 2,8% | 2,7% | 1,8% | 1,9% | 1,8% | 1,8% | 0,9% | 1,8% | **21,8%** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Tháng**  (triệu) | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | **Tổng** |
| 100(7) | 100(8) | 100(8) | 100(8) | 100(8) | 100(8) | 100(8) | 100(8) | 100(8) | 100(9) | 100(9) | 100(10) | **1200** |
| 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 2% | 1,8% | 1,8% | 1,8% | **21,8%** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Tháng**  (triệu) | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | **Tổng** |
| 100(10) | 100(10) | 100(10) | 100(11) | 100(11) | 100(11) | 100(11) | 100(12) | 100(12) | 100(12) | 100(12) | 100(12) | **1200** |
| 1,8% | 1,8% | 1,9% | 1,8% | 1,8% | 1,8% | 1,9% | 1,8% | 1,8% | 1,8% | 1,8% | 1,9% | **21,8%** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Tháng**  (triệu) | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | **Tổng** |
| 100(13) | 100(13) | 100(13) | 100(13) | 100(13) | 100(14) | 100(14) | 100(15) | 100(15) | 100(15) | 100(15) | 100(15) | **1200** |
| 1,8% | 1,8% | 1,8% | 1,8% | 1,9% | 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 1,8% | 1,9% | **21,8%** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Tháng**  (triệu) | 49 | 50 | 51 | 52 | 53 | 54 | 55 |  |  |  |  |  | **Tổng** |
| 100(16) | 100(16) | 100(16) | 100(16) | 97(17) | 97(17) | 96(17) |  |  |  |  |  | **690** |
| 1,8% | 1,8% | 1,8% | 1,9% | 1,8% | 1,8% | 1,8% |  |  |  |  |  | **12,8%** |

*Ghi chú*: dòng 1 là tháng, dòng 2 là kinh phí, ví dụ 100(1) là 100 triệu tương ứng với nội dung chi 1 trong bảng dự toán kinh phí, dòng 3 là tỉ lệ phần trăm trên tổng kinh phí.

Kinh phí phân bổ trung bình 100 triệu mỗi tháng và 1,2 tỷ mỗi năm.

**Danh sách một số phần mềm thống kê**

1. *ADMB*: phần mềm nguồn mở, chuyên dụng mô hình hóa thống kê phi tuyến. Trang web <http://www.admb.be/>.
2. *AM*: phần mềm nguồn mở hướng đến xử lý những mẫu phức tạp, dữ liệu lớn, tập trung vào ước lượng mô hình hồi quy bằng phương pháp cực đại hàm khả năng (MLE) cho dữ liệu lớn. Trang web <http://am.air.org/>.
3. *Dap*: phần mềm nguồn mở, là gói hỗ trợ phân tích và hiển thị dữ liệu. Đây là gói phần mềm nhỏ gọn viết bằng C. Trang web <http://www.gnu.org/software/dap/>.
4. *Dataplot*: phần mềm nguồn mở, tập trung vào chức năng vẽ biểu đồ, thống kê mô tả. Trang web <http://www.itl.nist.gov/div898/software/dataplot/>.
5. *Fytik*: phần mềm nguồn mở, chuyên dụng hồi quy phi tuyến. Trang web <http://fityk.nieto.pl/>.
6. *Gnumeric*: là phần mềm bảng tính mở rộng chức năng của Excel, đặc biệt có chức năng mô phỏng Monte Carlo. Trang web <http://projects.gnome.org/gnumeric/>.
7. *Gretl*: phần mềm nguồn mở, chuyên dụng phân tích chuỗi thời gian, hồi quy và tính toán các chỉ số kinh tế học. Trang web <http://gretl.sourceforge.net/>.
8. *IRRISTAT*: phần mềm quản lý dữ liệu và cung cấp một số chức năng thống kê cơ bản, ban đầu được phát triển để phân tích dữ liệu nông học nhưng một số tính năng có thể áp dụng cho dữ liệu khác. Các tính năng bao gồm quản lý dữ liệu bằng bảng tính, phân tích phương sai, hồi quy, phân tích tương tác môi trường, phân tích mẫu, vẽ biểu đồ. Nhìn chung phần mềm này thiên về phân tích dữ liệu hơn thống kê. Trang web <http://www.irri.org/science/software/irristat.asp>.
9. *Minitab*: phần mềm thống kê thương mại, có nhiều tính năng, được sử dụng rộng rãi, giao diện thân thiện, phần thống kê mô tả rất chi tiết như R. Các tính năng bao gồm: quản lý dữ liệu; vẽ biểu đồ; các kiểm định: z-test, t-test, tỉ lệ mẫu; phân tích phương sai; thiết kế thí nghiệm; kiểm định phi tham số; quản lý chất lượng; hồi quy; phân tích reliability và survival; xử lý dữ liệu nhiều chiều như PCA, phân cụm, discriminant; phân tích chuỗi thời gian và dự báo. Đây là phần mềm rất đa dụng, kết xuất các phân tích thống kê rất chi tiết. Trang web <http://www.minitab.com>.
10. *MicrOsiris*: nguồn mở, chạy trên hệ điều hành Windows được phát triển tại đại học Michigan, có nhiều tính năng đặc sắc thiên về phân tích dữ liệu, khai khoáng dữ liệu và hồi quy. Trang web <http://www.microsiris.com/>.
11. *OpenEpi*: nguồn mở, chuyên dụng phân tích dịch tễ học. Trang web <http://www.openepi.com/>.
12. *OpenStat*: gói thống kê tổng quát do Bill Miller đại học Iowa State U phát triển, chạy trên nền Win95/98/NT hỗ trợ phân tích và xử lý dữ liệu có giao diện tương tự SPSS. Trang web <http://www.statprograms4u.com/OpenStatMain.htm>.
13. *PAST*: gói phần mềm dễ sử dụng hỗ trợ môn cổ sinh vật học, nhưng cung cấp khả đủ các tính năng thống kê như kiểm định (F, t, chi-square, hoán vị, Kolmogorov-Smirnov, Mann-Whitney, Shapiro-Wilk, Spearman's Rho and Kendall's Tau, Kruskal-Wallis), one-way ANOVA. Trang web <http://folk.uio.no/ohammer/past/>.
14. *R*: phần mềm nguồn mở, các tính năng bao gồm: xử lý dữ liệu, vẽ biểu đồ, kiểm định. Các tính năng có thể được thêm vào bằng cách cài đặt các gói riêng lẻ. Nhìn chung các tính năng cốt lõi của R không nhiều như Minitab và giao diện không thân thiện như SPSS nhưng điểm xuất sắc của R là cung cấp một ngôn ngữ lập trình cho phép các nhà thống kê thực hiện những phân tích phức tạp. Hiện R càng trở nên phổ biến như Linux trong lĩnh vực hệ điều hành, nhiều nhà phát triển đang tập trung vào và tiềm năng mở rộng của R. Trang web <http://www.r-project.org>.
15. *Regress+*: phần mềm chuyên dụng mô hình hóa dữ liệu và hồi quy. Chức năng hồi quy và curve fitting rất mạnh mẽ. Trang web <http://www.causascientia.org/software/Regress_plus.html>.
16. *SalStat-2*: chạy đa nền, hỗ trợ quản lý dữ liệu và tính toán thống kê như: thống kê mô tả, hàm xác suất, kiểm định chi-square test, kiểm định t, ANOVA một yếu tố, hồi quy, hệ số tương quan, kiểm định phi tham số, vẽ biểu đồ. Trang web <http://code.google.com/p/salstat-statistics-package-2/>.
17. *SCaViS*: nguồn mở, hỗ trợ hiển thị dữ liệu, mẫu và số ngẫu nhiên, mạng neuron, hồi quy, phân cụm. Đây là phần mềm tương đối nhỏ gọn, trang web <http://jwork.org/scavis/>.
18. *Simfit*: phần mềm nguồn mở, chuyên dụng mô phỏng, hiển thị dữ liệu và curve fitting. Điểm đặc sắc của phần mềm này là vẽ biểu đồ, hiển thị 2 chiều và 3 chiều. Trang web <http://www.simfit.manchester.ac.uk/>.
19. *SISA*: gói phần mềm hỗ trợ phân tích thống kê, tính năng khá hạn chế nhưng có điểm đặc sắc là cung cấp giao diện web hỗ trợ tương tác trực tuyến. Trang web <http://www.quantitativeskills.com/sisa/>.
20. *SOFA*: chương trình báo cáo, phân tích và thống kê dữ liệu, chạy trên Windows, Mac và Linux, đặc trưng là dễ sử dụng. Trang web <http://www.sofastatistics.com/home.php>.
21. *SPSS*: phần mềm thương mại lâu đời với phiên bản đầu tiên vào năm 1968, được phát triển bởi IBM, có nhiều chức năng hữu ích như thống kê mô tả, kiểm định có tham số và phi tham số, hồi quy, ANOVA và khai khoáng dữ liệu. Giao diện thân thiện nhất và được sử dụng rộng rãi nhất. Phiên bản chạy thử mới nhất 21.0.0 tải từ trang <http://www-01.ibm.com/software/analytics/spss>.
22. *SSP*: gói phần mềm rất nhỏ gọn cung cấp một số chức năng thống kê cơ bản như vẽ biểu đồ, so sánh trung bình mẫu và tỉ lệ mẫu, phân tích phương sai, kiểm định chi-square, mô hình hồi quy đơn giản.
23. *Statext*: phần mềm nguồn mở, tập trung vào kiểm định giả thuyết, kết xuất dưới dạng văn bản. Trang web <http://www.statext.com/>.
24. *Statist*: gói phần mềm nhỏ gọn, khả chuyển cung cấp những chức năng thống kê cơ bản như xử lý dữ liệu, biểu đồ, hồi quy và tương quan, kiểm định chi-square và t. Phần mềm viết bằng C, chạy trên Windows, Linux và Max. Trang web <http://wald.intevation.org/projects/statist/>.
25. *Tanagra*: đây là phần mềm khai khoáng dữ liệu, cung cấp vài chức năng thống kê như hồi quy, phân tích phương sai. Trang web <http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>.
26. *ViSta*: nguồn mở, phần mềm chuyên dụng hiển thị dữ liệu thống kê. Trang web <http://forrest.psych.unc.edu/research/>.
27. *WebStat*: phần mềm nguồn mở viết bằng Java, có chức năng đặc sắc là cung cấp giao diện web để phân tích dữ liệu. Trang web <http://www.statcrunch.com/> sử dụng WebStat.

Nguồn:

<http://en.wikipedia.org/wiki/List_of_statistical_packages>

<http://statpages.org/javasta2.html>

Tp. Hồ Chí Minh ngày 25 tháng 02 năm 2013

Chủ nhiệm dự án

Nguyễn Phước Lộc