## **So sánh năng suất nghiên cứu khoa học các ngành**

***Vấn đề:***

* Ở độ tuổi nào người tham gia nghiên cứu nhiều nhất?
* Ở độ tuổi nào người nghiên cứu thường có năng suất nghiên cứu tốt nhất?
* Độ tuổi nghiên cứu của các ngành có gì khác biệt không?
* Do 1 bài có thể có nhiều người tham gia, số liệu được đánh giá trên cả số người, số bài ở từng độ tuổi và số bài trung bình của 1 người ở từng độ tuổi

Thống kê so sánh số lượng bài nghiên cứu trên csdl SSHPA của 4 nhóm chính bao gồm nhóm ngành kinh tế (eco), nhóm ngành giáo dục (edu), nhóm ngành y (med), và nhóm các ngành khác (others)

**Table 1. Tổng số bài nghiên cứu theo nhóm tuổi (agegrp\_articles.csv):**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **<25** | | **25-29** | | **30-34** | | **35-39** | | **40-44** | | **45-49** | | **50-54** | | **55-59** | | **>=60** | | **Total** |
| **eco** | 22 | 2% | 107 | 8% | 219 | 17% | 305 | 23% | 334 | 25% | 161 | 12% | 75 | 6% | 49 | 4% | 50 | 4% | 1,322 |
| **edu** | 7 | 2% | 27 | 7% | 79 | 20% | 146 | 37% | 71 | 18% | 39 | 10% | 16 | 4% | 8 | 2% | 4 | 1% | 397 |
| **med** | 45 | 6% | 112 | 15% | 133 | 18% | 85 | 12% | 92 | 13% | 133 | 18% | 36 | 5% | 39 | 5% | 53 | 7% | 728 |
| **other** | 22 | 2% | 97 | 8% | 259 | 20% | 300 | 24% | 246 | 19% | 144 | 11% | 68 | 5% | 73 | 6% | 65 | 5% | 1,274 |

**Table 2. Tổng số người nhiên cứu theo nhóm tuổi (agegrp\_authors.csv):**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **<25** | **25-29** | **30-34** | **35-39** | **40-44** | **45-49** | **50-54** | **55-59** | **>=60** |
| **eco** | 18 | 80 | 133 | 169 | 142 | 78 | 41 | 23 | 17 |
| **edu** | 5 | 20 | 46 | 67 | 50 | 26 | 13 | 7 | 4 |
| **med** | 18 | 54 | 60 | 47 | 56 | 49 | 19 | 15 | 13 |
| **others** | 16 | 69 | 157 | 170 | 135 | 90 | 52 | 41 | 32 |

Trên table 1 và table 2 ta có thể thấy số lượng tác giả và bài viết của ngành y phân bố tương đối đều bắt đầu từ độ tuổi trên 25 cho đến 50 tuổi trong khi đó các ngành khác độ tuổi nghiên cứu nhiều nhất tập trung ở khoảng từ 30 đến 45. Đa phần các ngành có độ tuổi nghiên cứu 35-39 (tuổi nghiên cứu sinh????)

|  |  |
| --- | --- |
|  | Đồ thị tương quan số bài nghiên cứu và độ tuổi các ngành |

|  |  |
| --- | --- |
|  | Đồ thị tương quan số tác giả và độ tuổi các ngành |

**Table 3. Số bài trung bình theo độ tuổi**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **<25** | **25-29** | **30-34** | **35-39** | **40-44** | **45-49** | **50-54** | **55-59** | **>=60** |
| **eco** | 1.22 | 1.34 | 1.65 | 1.80 | 2.35 | 2.06 | 1.83 | 2.13 | 2.94 |
| **edu** | 1.40 | 1.35 | 1.72 | 2.18 | 1.42 | 1.50 | 1.23 | 1.14 | 1.00 |
| **med** | 2.50 | 2.07 | 2.22 | 1.81 | 1.64 | 2.71 | 1.89 | 2.60 | 4.08 |
| **other** | 1.38 | 1.41 | 1.65 | 1.76 | 1.82 | 1.60 | 1.31 | 1.78 | 2.03 |

|  |  |
| --- | --- |
|  | Số bài trung bình theo độ tuổi tác giả từng ngành |

Số liệu độ tuổi trung bình các tác giả nhằm đánh giá xu hướng tuổi tác giả tham gia nghiên cứu.

|  |  |
| --- | --- |
|  | Độ tuổi trung bình của tác giả từng ngành theo năm |

Đồ thị độ tuổi trung bình tác giả từng ngành cho thấy độ tuổi trung bình nghiên cứu ở khoảng 40 tuổi. Độ tuổi trung bình có xu hướng giảm nhẹ về khoảng 37-38 tuổi ở các năm gần đây.

## **Sử dụng bayesian để phân tích dữ liệu số bài theo độ tuổi:**

* Mô hình

Oarticle ~ αage[αfield]

* Biến outcome Oarticle là số bài viết của các tác giả ở các độ tuổi và lĩnh vực khác
* Biến phụ thuộc α gồm độ tuổi và ngành được chia thành 2 lớp hierachy
  + Có 4 mức giá trị biến ngành (kinh tế, giáo dục, y tế, các ngành còn lại)
  + Có 9 mức giá trị độ tuổi từ 25-60 tuổi và hơn nữa chia thành các nhóm 5 năm. Độ tuổi ở các mức giá trị lại được nhóm ở từng ngành, do đó biến tuổi đã phân nhóm gồm có 36 mức giá trị (4 ngành x 9 mức tuổi)
* Kết quả mô hình trên bayesvl

**Code stan:**

|  |
| --- |
| data{  // Define variables in data  int<lower=1> Nobs; // Number of observations (an integer)  real article[Nobs]; // outcome variable  int Nagenum;  int<lower=1,upper=Nagenum> agenum[Nobs];  int agenum2fieldnum[Nagenum];  int Nfieldnum;  int<lower=1,upper=Nfieldnum> fieldnum[Nobs];  }  transformed data{  // Define transformed data  }  parameters{  // Define parameters to estimate  real<lower=0> sigma\_article;  real<lower=0> sigma\_agenum;  vector[Nagenum] u\_agenum;  real a0\_fieldnum;  real<lower=0> sigma\_fieldnum;  vector[Nfieldnum] u\_fieldnum;  }  transformed parameters{  // Transform parameters  real mu\_article[Nobs];  real a\_agenum[Nagenum];  vector[Nfieldnum] a\_fieldnum;  // Varying intercepts definition  for(k in 1:Nfieldnum) {  a\_fieldnum[k] = a0\_fieldnum + u\_fieldnum[k];  }  // Next level random intercepts  for(k in 1:Nagenum) {  a\_agenum[k] = a\_fieldnum[agenum2fieldnum[k]] + u\_agenum[k];  }  for (i in 1:Nobs) {  mu\_article[i] = a\_agenum[agenum[i]];  }  }  model{  // Priors  sigma\_agenum ~ normal(0,10);  u\_agenum ~ normal(0, sigma\_agenum);  a0\_fieldnum ~ normal(0,10);  sigma\_fieldnum ~ normal(0,10);  u\_fieldnum ~ normal(0, sigma\_fieldnum);  // Likelihoods  article ~ normal(mu\_article, sigma\_article);  }  generated quantities {  // simulate data from the posterior  real yrep\_article[Nobs];  // log-likelihood posterior  vector[Nobs] log\_lik\_article;  for (i in 1:num\_elements(yrep\_article)) {  yrep\_article[i] = normal\_rng(mu\_article[i], sigma\_article);  }  for (i in 1:Nobs) {  log\_lik\_article[i] = normal\_lpdf(article[i] | mu\_article[i], sigma\_article);  }  } |

**Kết quả mô phỏng:**

|  |
| --- |
| Model Info:  nodes: 3  arcs: 2  scores: NA  formula: article ~ a\_agenum[agenum]  Estimates:  Inference for Stan model: 360396df8dfc64b1ff2509fe87675927.  4 chains, each with iter=5000; warmup=2000; thin=1;  post-warmup draws per chain=3000, total post-warmup draws=12000.  mean se\_mean sd 2.5% 25% 50% 75% 97.5% n\_eff Rhat  a\_agenum[1] 2.07 0.01 0.48 1.11 1.74 2.08 2.40 3.01 1573 1.00  a\_agenum[2] 1.77 0.01 0.32 1.13 1.55 1.77 2.00 2.37 1533 1.00  a\_agenum[3] 1.91 0.00 0.26 1.41 1.74 1.91 2.08 2.40 9800 1.00  a\_agenum[4] 2.04 0.01 0.24 1.59 1.88 2.04 2.20 2.51 1139 1.00  a\_agenum[5] 2.62 0.00 0.25 2.12 2.45 2.62 2.79 3.11 4441 1.00  a\_agenum[6] 2.39 0.01 0.32 1.77 2.18 2.40 2.62 3.01 3031 1.00  a\_agenum[7] 1.99 0.00 0.39 1.22 1.75 1.99 2.26 2.75 11168 1.00  a\_agenum[8] 2.27 0.01 0.45 1.39 1.96 2.27 2.56 3.16 1560 1.00  a\_agenum[9] 2.39 0.02 0.50 1.45 2.04 2.37 2.72 3.37 791 1.01  a\_agenum[10] 2.24 0.01 0.56 1.15 1.86 2.24 2.59 3.36 1842 1.00  a\_agenum[11] 2.01 0.01 0.46 1.08 1.73 2.02 2.31 2.91 5140 1.00  a\_agenum[12] 1.98 0.02 0.40 1.22 1.72 1.97 2.23 2.82 274 1.01  a\_agenum[13] 2.38 0.02 0.35 1.71 2.15 2.38 2.62 3.05 375 1.01  a\_agenum[14] 1.92 0.01 0.38 1.17 1.67 1.92 2.18 2.64 2247 1.00  a\_agenum[15] 1.98 0.01 0.43 1.11 1.70 2.00 2.27 2.81 6974 1.00  a\_agenum[16] 2.03 0.01 0.50 0.98 1.69 2.04 2.37 2.98 6519 1.00  a\_agenum[17] 1.97 0.01 0.55 0.85 1.62 1.99 2.36 2.98 2336 1.00  a\_agenum[18] 2.20 0.01 0.56 1.05 1.86 2.19 2.56 3.29 10072 1.00  a\_agenum[19] 2.56 0.03 0.51 1.59 2.21 2.53 2.89 3.58 390 1.01  a\_agenum[20] 2.86 0.01 0.38 2.16 2.61 2.86 3.10 3.63 4226 1.00  a\_agenum[21] 3.24 0.01 0.39 2.49 2.97 3.24 3.50 4.02 3718 1.00  a\_agenum[22] 2.26 0.01 0.38 1.55 2.00 2.26 2.52 3.01 676 1.00  a\_agenum[23] 2.08 0.01 0.35 1.37 1.85 2.09 2.33 2.77 4480 1.00  a\_agenum[24] 3.31 0.01 0.43 2.48 3.00 3.30 3.60 4.19 2021 1.00  a\_agenum[25] 2.12 0.01 0.47 1.20 1.83 2.12 2.43 3.05 5216 1.00  a\_agenum[26] 2.67 0.01 0.50 1.74 2.33 2.64 2.99 3.72 3042 1.00  a\_agenum[27] 3.02 0.01 0.59 1.96 2.62 2.99 3.37 4.28 2823 1.00  a\_agenum[28] 2.26 0.01 0.47 1.33 1.96 2.26 2.56 3.20 8383 1.00  a\_agenum[29] 1.88 0.00 0.33 1.23 1.67 1.87 2.10 2.52 8219 1.00  a\_agenum[30] 1.92 0.00 0.24 1.45 1.76 1.92 2.08 2.41 7934 1.00  a\_agenum[31] 2.08 0.00 0.23 1.61 1.93 2.08 2.24 2.53 6191 1.00  a\_agenum[32] 2.14 0.01 0.26 1.66 1.96 2.14 2.32 2.66 1597 1.00  a\_agenum[33] 1.99 0.01 0.31 1.37 1.77 1.98 2.20 2.58 448 1.01  a\_agenum[34] 1.72 0.00 0.37 0.97 1.49 1.72 1.97 2.41 6220 1.00  a\_agenum[35] 2.35 0.04 0.44 1.14 2.09 2.37 2.64 3.14 133 1.02  a\_agenum[36] 2.22 0.01 0.42 1.44 1.93 2.22 2.51 3.06 1044 1.01  sigma\_agenum 0.55 0.00 0.14 0.31 0.45 0.54 0.64 0.85 2026 1.00  a\_fieldnum[1] 2.32 0.01 0.22 1.94 2.18 2.31 2.46 2.79 992 1.00  a\_fieldnum[2] 2.10 0.00 0.22 1.62 1.97 2.12 2.26 2.51 4999 1.00  a\_fieldnum[3] 2.32 0.01 0.21 1.94 2.18 2.31 2.45 2.76 982 1.00  a\_fieldnum[4] 2.23 0.00 0.20 1.83 2.09 2.23 2.35 2.63 3075 1.00  a0\_fieldnum 2.24 0.01 0.29 1.68 2.10 2.24 2.37 2.82 1172 1.00  sigma\_fieldnum 0.36 0.01 0.46 0.03 0.11 0.23 0.43 1.48 1584 1.00 |

Các hệ số a\_agenum là hệ số tuổi ở các ngành tương ứng với 36 mức giá trị

Hệ số a\_fieldnum là hệ số ngành với 4 mức giá trị



Fig 1.11



Fig 1.12 - Giá trị các hệ số

|  |  |
| --- | --- |
| Fig 1.13 năng suất theo độ tuổi ngành kinh tế | Fig 1.14 năng suất theo độ tuổi ngành giáo dục |
| Fig 1.15 năng suất theo độ tuổi ngành y | Fig 1.16 năng suất theo độ tuổi ngành khác |
|  |  |