Stat101 Homework Report for Unit 2

Comparison among Sections

Overview

The report will compare the homework outcomes from 3 sections in 3 topics.

There are totally 150 students who submitted the answers. 150 of them did all the homeworks. A summary of the number of students and counts of the missing homeworks are displayed in Table 1.

Since different homeworks have different full scores, the percentages correct are used to compare among homeworks. The average correct percents for the 9 homeworks are shown in Table 2.

Figure 1 shows the temporal change of students' performance, and Figure 2 compares the histograms of those homeworks.

	Total	Missing.Topic06	Missing.Topic07	Missing.Topic08
A	50	0	0	0
B	50	0	0	0
C	50	0	0	0

Table 1: The number of students and counts of missing homeworks by section. The first column is the number of students, and the rest columns are the number of missing homeworks.

	Topic06	Topic07	Topic08
A	82.40	90.80	71.55
B	75.44	89.68	71.09
C	81.52	88.48	69.91

Table 2: Average correct percentages by section and topic.

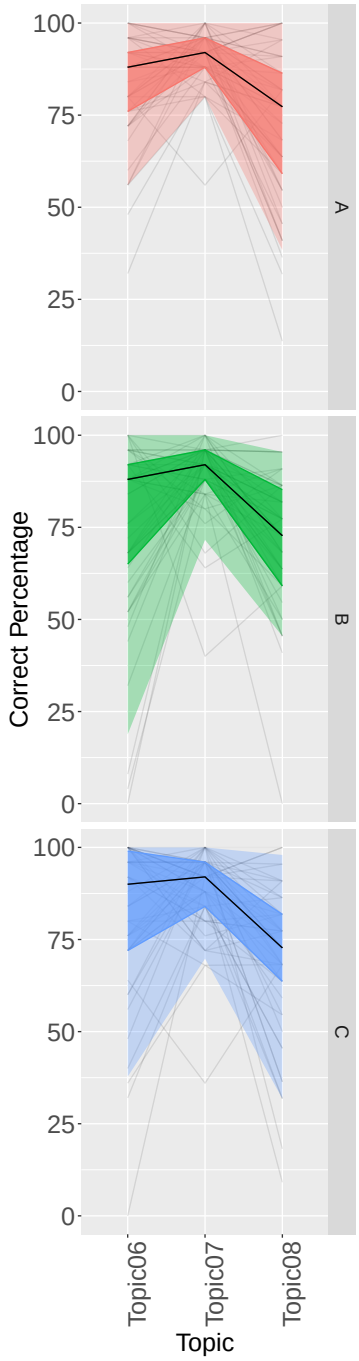

Figure 1: Line plots of the correct percentages by section. The black lines give the medians in each topic; the dark colored areas show the interquartile ranges (25%-75%); and the light colored areas are the 5%-95% bands. The light grey lines are the real correct percentage by student.

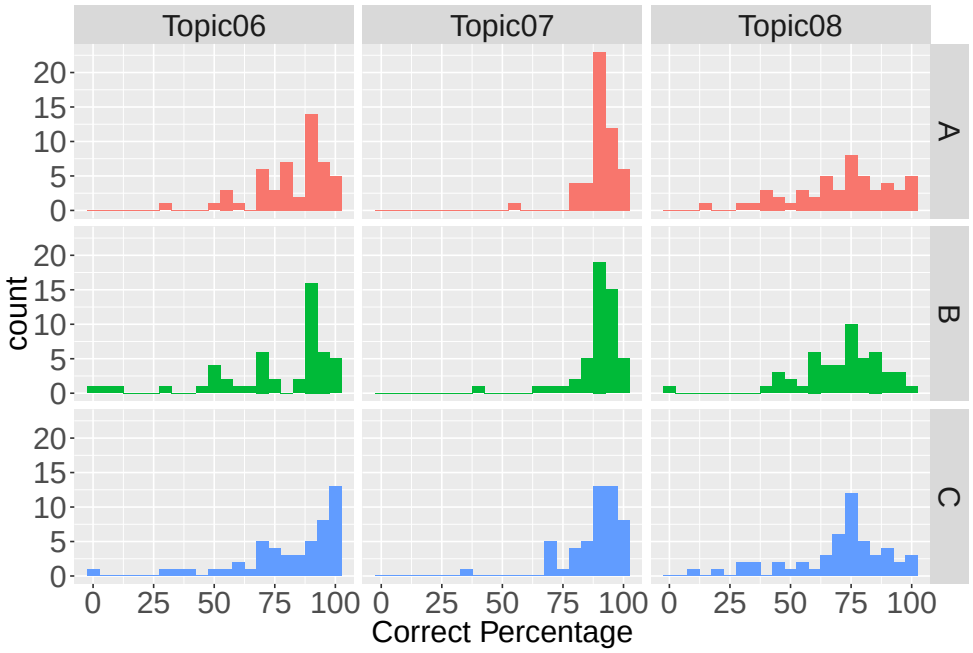

Figure 2: Histograms of the percentages correct by section and topic.

Factor effects and comparisons

To analyze the effects of sections and topics on students' performance, we consider the generalized linear mixed model, with the response (homework score), and the predictors (section and topic).

Let Y_{ijk} denote the score that the kth student in Section j gets for Topic i. $i = 1, ...,3$ topics; $j = 1, ...,3$ sections; $k = 1, ..., n_j$ students.

Consider that most homework questions give 1 point for 1 correct answer, for example, a multiple choice is usually 1 point; a matching question with 6 items is usually 6 points. Hence we assume that the full point is equal to the number of questions in one homework, and the students can only get either 0 or 1 point for each question. Assume that the correct probabilities of the questions in one homework are homogeneous, denoted by $p_{ijk} \in (0, 1)$, then we have

$$Y_{ijk} \sim Binomial(N_i, p_{ijk})$$

where N_i is the full point (i.e., the total number of questions) of the ith homework.

A mixed effects model is considered since we have multiple measurements on each student, while the students may have different ability to study. The fixed effects are topic (t), section (s), and their interaction (ts). The random effect is student (u).

Note that 150 students are divided into 3 sections. One issue is that if we believe σ_u^2 are equal between the sections. In this study we assume that the variances of random effects are not equal, i.e., $u_{jk} \sim N(0, \sigma_j^2)$. Hence we have the following equation,

$$g(p_{ijk}) = \mu + t_i + s_j + ts_{ij} + u_{jk}$$

where $g()$ is the link function. By default the software R will set $t_1 = 0$, $s_1 = 0$ and $ts_{ij} = 0, \forall i, j = 1$ as the identifiability constraints.

The result of the model is as below. Figure 3 compares the fitted values with the response. Figure 4 shows the estimated correct probability.

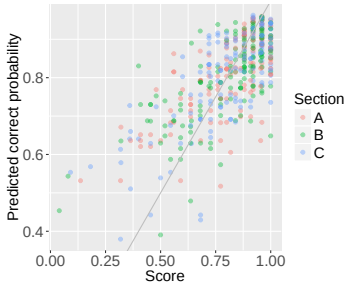

Figure 3: Scatterplot between the scaled scores and the predicted correct probability. The predicted probability is continuous in (0,1). But the true scores are discrete.

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Score ~ Topic * Section + (1 | Student)
##    Data: score
## Weights: FullPoints
##
##      AIC      BIC   logLik deviance df.resid
##   2712.2   2753.2  -1346.1   2692.2      437
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7382 -0.8069  0.1895  1.1884  3.6828
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  Student (Intercept) 0.3558   0.5965
## Number of obs: 447, groups:  Student, 150
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.6089     0.1137  14.147  < 2e-16 ***
## TopicTopic07            0.7632     0.1242   6.143 8.10e-10 ***
## TopicTopic08           -0.6439     0.1017  -6.330 2.45e-10 ***
## SectionB               -0.3244     0.1582  -2.050  0.04032 *
## SectionC                0.1045     0.1627   0.642  0.52090
## TopicTopic07:SectionB   0.2147     0.1709   1.256  0.20896
## TopicTopic08:SectionB   0.3851     0.1420   2.713  0.00667 **
## TopicTopic07:SectionC  -0.3064     0.1726  -1.775  0.07592 .
## TopicTopic08:SectionC  -0.1649     0.1457  -1.131  0.25791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) TpcT07 TpcT08 SectnB SectnC TT07:SB TT08:SB TT07:SC
## TopicTopc07 -0.400
## TopicTopc08 -0.495  0.449
## SectionB    -0.719  0.287  0.356
## SectionC    -0.699  0.279  0.346  0.502
## TpcTpc07:SB  0.291 -0.727 -0.326 -0.387 -0.203
## TpcTpc08:SB  0.355 -0.321 -0.717 -0.472 -0.248  0.433
## TpcTpc07:SC  0.288 -0.720 -0.323 -0.207 -0.420  0.523   0.231
## TpcTpc08:SC  0.345 -0.313 -0.698 -0.248 -0.507  0.228   0.500   0.471
```

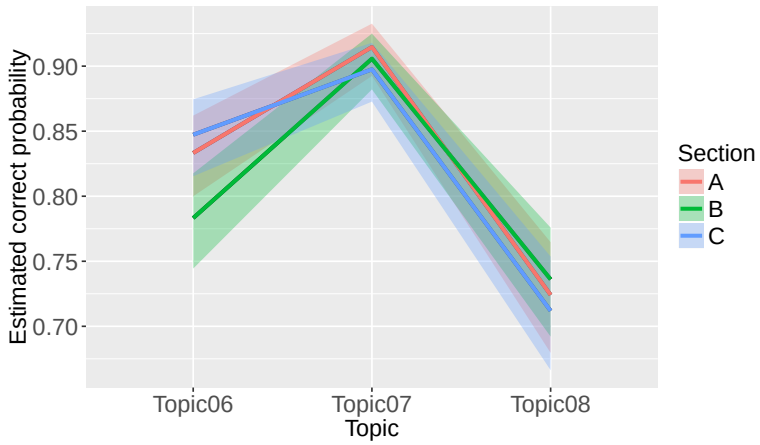

Figure 4: The estimated correct probability by section. On average, section A has the highest correct probability, and B has the lowest.

From the model we see that some sections are better than other sections, and some topics are easier to understand than other topics. So the next question is that are the differences significant between topics and between sections?

We run the multiple test for comparison between the topics:

1. $H_0: \beta_{t_i} = 0 \quad \text{vs} \quad H_a: \beta_{t_i} \neq 0$
   for $\forall\, i \in \{2, 3, \cdots, 3\}$.

2. $H_0: \beta_{t_i} - \beta_{t_j} = 0 \quad \text{vs} \quad H_a: \beta_{t_i} - \beta_{t_j} \neq 0$
   for $\forall\, i \neq j,\ i, j \in \{2, 3, \cdots, 3\}$.

and get the the adjusted p-values by Bonferroni method in Table 3.

	Topico6	Topico7
Topico7	0.0000	
Topico8	0.0000	0.0000

Table 3: P-values of the multiple comparison between topics

Then run the multiple test pairwisely between the sections:

1. $H_0: \beta_{s_i} = 0 \quad \text{vs} \quad H_a: \beta_{s_i} \neq 0$
   for $\forall\, i \in \{2, 3, \cdots, 3\}$.

2. $H_0: \beta_{s_i} - \beta_{s_j} = 0 \quad \text{vs} \quad H_a: \beta_{s_i} - \beta_{s_j} \neq 0$
   for $\forall\, i \neq j,\ i, j \in \{2, 3, \cdots, 3\}$.

and get the the adjusted p-values by Bonferroni method in Table 4.

	A	B
B	0.1227	
C	1.0000	0.0230

Table 4: P-values of the multiple comparison between sections

Clustering

During a semester of 16 weeks, students may not keep the same pace in studying. Some students are smart and working very hard all the time. Some students start full of energy, and gradually lose their passions. Some students do not take the course seriously until they are challenged by some difficult content.

We are interested in finding some featured behaviors along with time, by clustering the students into groups. The result of hierarchical clustering on the scaled scores is shown in Figure 5. Up to down it seperates the students in 1 to 9 groups. It suggests the best number of clusters is 2.

Note that the euclidean distance of the correct percentages is not a reasonable distance measure, because it emphasizes more on the score level, less on the temporal pattern. For example, student A gets 100, 95, 90 for the first three homeworks; student B gets 70, 65, 60; and student C gets 60, 65, 70. Student A and B have the same decreasing pattern, but by the euclidean distance, B and C are closer.

Similar as using the correlation distance, we scaled the correct percentages within the records of each student. Revisit the example above, the scaled scores for student A and B are 1, 0, -1; for student C are -1, 0, 1. Then A and B can be grouped together.

To find the better number of clusters, first we consider two simple criteria. One is the within group sum of squares (SSE), the other is the proportion of between group sum of squares (SSR) over the total sum of squares (SSR+SSE), i.e., the R-square. Figure 6 gives the scree plots for SSE and R-squares for the models with 2 to 9 clusters. The lines connect the real values from the criteria, and the dots are the bootstrap simulation. The plots suggest 3 clusters.

At the meanwhile, we consider another criterion - the optimum average silhouette width. As seen in Figure 7, it suggests the best number of clusters is 3.

Since 3 clusters is a reasonable choice, the mean scaled scores of each cluster on each topic is displayed in Figure 8.

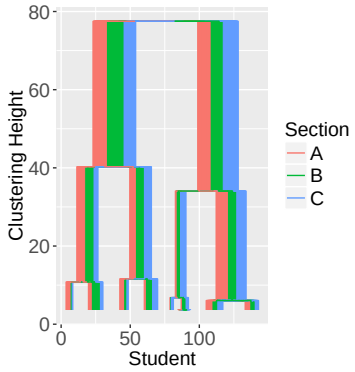

Figure 5: Tree plot of hierarchical clustering, from one to nine clusters.

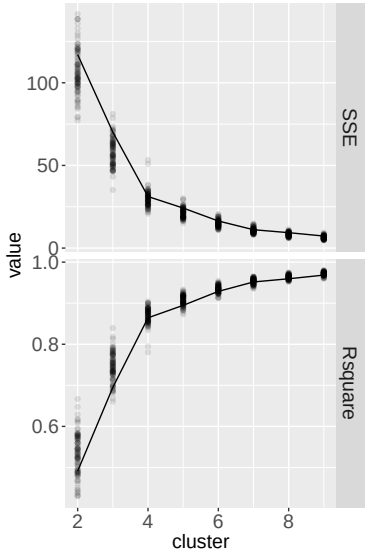

Figure 6: Two criteria to refer the number of clusters. The points are results from bootstrap simulation, and the line connects the real values from the criteria.

	1	3	2
A	39.47	28.21	32.88
B	36.84	38.46	28.77
C	23.68	33.33	38.36

Table 5: Contingency table in Column% between sections and clusters.

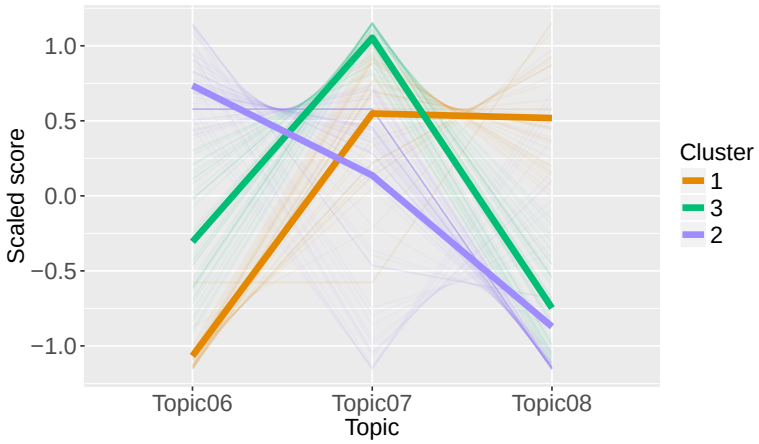

Figure 8: Trend of the 3 clusters. The thick lines connect the means of the clusters; the light thin lines show the scaled scores by student.

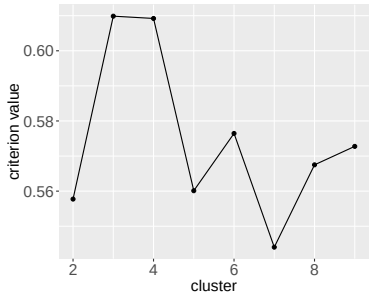

Figure 7: Using the optimum average silhouette width criterion to find the number of clusters. The larger value, the better model.

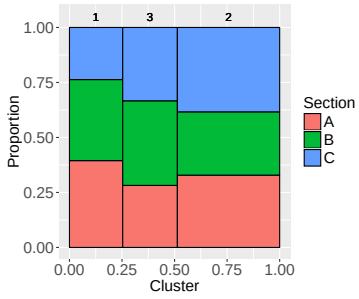

Figure 9: Mosaic plot by section and cluster.

Students

A	B	C
3	2	6

Table 6: Number of students that have at least a half of the homework scores in the bottom 20%.

Acknowledgement

This report is generated by Xiaoyue Cheng, Dianne Cook, Lindsay Rutter, and Amy Froelich, using R-3.1.2 with packages knitr, xtable and ggplot2.