ST3009: Statistical Methods for Computer Science

Final Assignment - Senán d'Art - 17329580

Q1

(a)

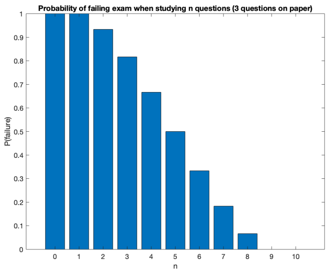
$\binom{10}{3} = 120$ possible combinations. There are 10 possible questions, we choose 3 of those 10.

(b)

$\frac{\binom{10-n}{3}}{\binom{10}{3}}$ where $0 <= n <= 7$. If $n > 7$, the the probability of none of the questions studied appearing = 0

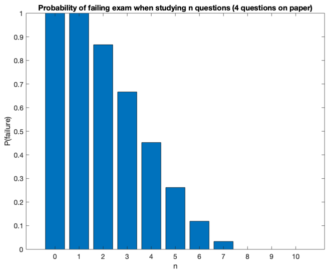
(c)

p(none)+p(exactly 1)
p(none) = $\frac{\binom{10-n}{3}}{\binom{10}{3}}$ where $0 <= n <= 7$, outside of this range $P = 0$
p(exactly 1) = $\frac{\binom{n}{1}*\binom{9-n}{2}}{\binom{10}{3}}$, where $1 <= n <= 8$, outside of this range $P = 0$



(d)

p(none)+p(exactly 1)
p(none) = $\frac{\binom{10-n}{4}}{\binom{10}{4}}$, where $0 <= n <= 6$, outside of this range $P = 0$
p(exactly 1) = $\frac{\binom{n}{1}*\binom{10-n}{3}}{\binom{10}{4}}$, where $1 <= n <= 7$, outside of this range $P = 0$



COMPARE WITH RESULTS FROM PART (C)
The chance of failing is much lower.
This is due to the additional question on the exam.
Do some maths on the percentage difference at each point.
Still need to answer 2 questions to pass but more likely that you will have studied at least 2.
Far less likely that will have studied none of 4 vs. none of 3.

(e)

The code generates 2 lists of numbers 0-10. It selects the first 3 from one and first n from the other. If there is an overlap of more than 1 ie. student passes exam, it returns 1, otherwise 0.

(f)

Call function from (e) N times and return the mean

Searching for range $[\mu - 2\sigma, \mu + 2\sigma]$

$$\mu \pm 2\frac{\sigma}{\sqrt{N}}$$

$var(X_i) = \sigma^2$

$$E[X_i] \pm 2\frac{\sqrt{var(X_i)}}{\sqrt{N}}$$

$$E[X_i] \pm 2\sqrt{\frac{var(X_i)}{N}}$$

Using (c):
$1 - E[X_i] = 0.1833, E[X_i] = 0.8167$ when $n = 7$.
$var(X_i) = \mu * (1 - \mu) = 0.8167 * 0.1833 = 0.1497$

Where $N = 1,000$:
$0.8167 \pm 2\sqrt{\frac{0.1497}{1000}}$
0.8167 ± 0.02447
$[0.79223, 0.84117]$

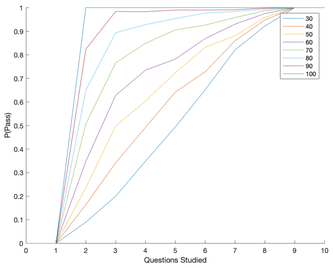
Where $N = 10,000$:
$0.8167 \pm 2\sqrt{\frac{0.1497}{10000}}$
0.8167 ± 0.00773
$[0.80897, 0.82443]$

(g)

Change this a bit
The simulation generates a new set of topics studied by a student and a set of topics to appear on the exam for each N. It then verifies if the student has passed the exam and uses the series of values to calculate the mean, Y.
For each iteration of the simulation, the resulting Y is compared to the confidence interval and if it lands within the interval it is added to a counter. When all X simulations have been run, this counter is divided by X to find what percentage was within the confidence interval.
I chose an X of $1,000$. The reason for this is that we are already using values of $N = 1,000, N = 10,000$ resulting in a total number of iterations of $1,000,000$ and $10,000,000$ respectively. This number is more than large enough to provide an accurate result. Running time was also unreasonable when $X > 1,000$ which did factor into the decision.
In the case of $N = 1,000$ the simulation resulted in an accuracy of 95.4%.
In the case of $N = 10,000$ the simulation resulted in an accuracy of 94.9%.
These results seem very reasonable as both are quite close to 95%. With a different seed for the random number generator the result could be slightly different but both results are acceptably close to 95%.

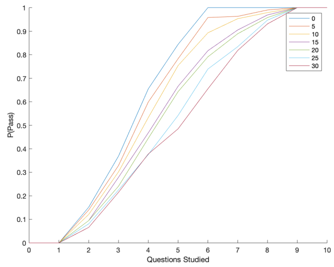
(h)

Case A - More likely to appear after being on last exam:
The student could modify their approach as follows:
The first 3 topics studied will be the ones on the previous exam. After this, they will be chosen at random from all remaining topics. This approach should heavily increase the likelihood of passing depending on how predictable the exam is.
The simulation was modified as follows:
A 'past exam' was generated, representing the previous year's exam. Based on how predicatable this year's exam is, the questions from the previous exam were more likely to be selected this year. The student's approach to the exam was to first study the previous exam and then all other topics.



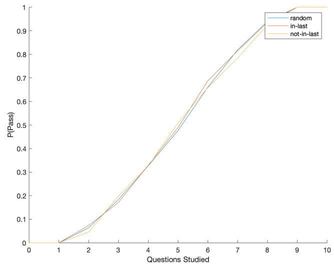
Here the legend refers to the % probability that a question will appear on this year's exam, given that it appeared last year. The simulation was run for $N = 1,000$ for all values. As can be clearly seen in this chart: a more predictable exam will result in a much higher chance of passing. In the case of a completely random exam, each question has a 30% chance of appearing on the paper. Here we can see what happens when we chance that probability from 30-100% in intervals of 10%.

Case B - Less likely to appear after being on the last exam:
In this case the student would take a different approach:
The last 3 questions studied would be those in last year's exam. All others would be chosen at random.
The code for this operates in a similar manner to Case A where percentages go from $0 \rightarrow 30\%$ in intervals of 5%.



Again the legend refers to the probability (in %) that a question will appear on this year's exam, given that it was on last year's exam. In this case, the chance of passing was not increased as drastically as in the previous case where a question is more likely to appear 2 exams in a row. It does reduce the number of topics required to be studied from 9 to 6 if the exam is perfectly predictable and the student wants to guarantee that they will pass.

Case C - The exam is not predictable but the student assumes it is:
Here we apply all 3 strategies of studying: random selection, prioritising questions that **were** on last year's exam, prioritising questions that **were not** on last year's exam. The results should be the same for all 3 as all questions have the same probability of appearing but we will run the simulation anyway.

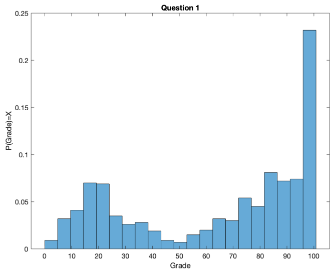


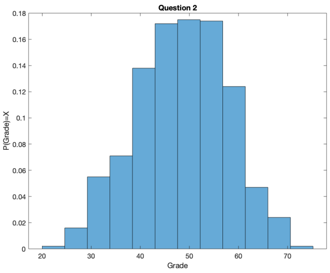
Here the legend refers to the student's approach when studying topics. As can be seen in the graph, there is no advantage to any of the study methods when the exam is randomised. While the results are not exactly the same, they are close enough to be considered equal for the sample size. This also means that there is no disadvantage to the student for assuming the exam is predictable, when it is not.

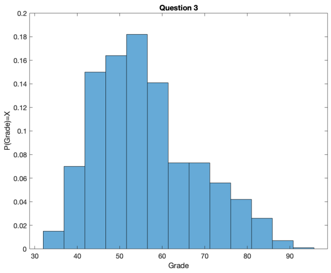
Q2

Dataset: `# id:0.332:0.5-0.524:2-0.308:2-0`

(a)







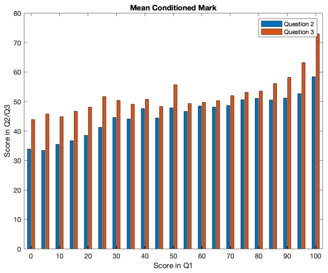
When comparing each of these questions, we can see a clear trend with each question. In the case of the first question, a large number of students got full marks. There is a cluster at each end with students trending towards doing very well or quite poorly. The question could have a relatively simple answer in a topic that would be easy to answer well if the student studied the topic but would be very hard to pass if they had not studied the topic.

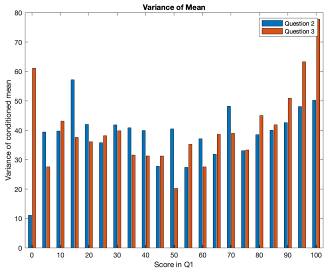
The second question has a more central distribution with few students failing (<40%) but none doing exceptionally well. Most scores are clustered around the 50% mark. This seems to show that it was relatively easy to pass but very hard to do well. It also implies that most students were prepared for this question.

Question 3 was reasonable,
allowed students to excel
Looks like a well balanced question

(b)

The following two graphs illustrate the conditioned mean and variance for Q2 & Q3 based on results in Q1.





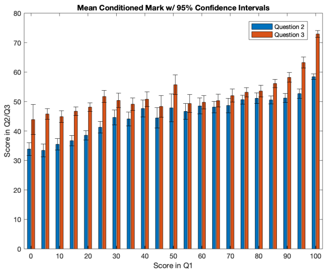
Here the mean follows a general upward trend, where students who performed well in the first question generally also did well in the subsequent 2. However the variance is very high.

I did not use binning to group the data. In this case all scores are in groups of 5% however they are not binned as all results in the dataset were multiples of 5%. It could be useful to bin the results in batches of 10% as these are generally what grade lines are based on eg. the difference between a 2nd and a 1st is $10\%+$, more granular measurements of grades are not pertinent to the final grade.
In the case of real-world data where results are not multiples of 5%, binning would be important, as it would mostly likely be a waste of time to compare exact scores, in this case I would again use bins of 5%.

(c)

using CLT

$$\sigma = \sqrt{variance}$$
$$\mu \pm 1.96 * \frac{\sigma}{\sqrt{N}}$$



Q2 is easier in general.

x axis is q1 mark.
y axis is mean for q2 & 3.
error bars for all.

(d)

Appendix

Section 1: Code for Q1 (c).

Section 2: Code for Q1 (d).

Section 3: Code for Q1 (e).

Section 4: Code for Q1 (f).

Section 5: Code for Q1 (g).

Section 6: Code for Q1 (h) A.

Section 7: Code for Q1 (h) B.

Section 8: Code for Q2 (a).

Section 9: Code for Q2 (b).

Section 10: Code for Q2 (c).