Assignment 1 working and answers

Student B217754

2022-10-16

SDA ASSIGNMENT 1

For question answers, call question_[question number]_[question part]. e.g.: question_1_a

QUESTION 1 a)

As the question states that mic failures follow a Poisson distribution, we are to assume that they are independent. This allows us to use the PMF. The rate of mic failures is 1 every 250 minutes. Rate of failures per 50 minute class is 50/250 or 1/5.

```r
lambda<-50/250
```

Here, dpois gives probability that 'mic failures' occurs 0 times in 50 minutes, multiplying this by 100 gives % of classes with no mic failures.

```r
zero_failures<-0
(dpois(zero_failures,lambda))*100
```

```
## [1] 81.87308
```

Answer: 81.87%

QUESTION 1 b)

The probability of observing at least 18 classes without failure is the same as having at most 2 classes with mic failures. We calculate lambda according to the appropriate quantity of space/time

```r
lesson_total_mins<-20*50 #(20 lessons*50 mins)=time unit for this question
```

Average number of mic failures (lambda) for this time period would therefore be

```r
expected_failures_in_20<-(lesson_total_mins/250)
```

Here, ppois can tell us the cumulative probability of getting at most 2 mic failures

```
up_to_2_failures<-2
ppois(up_to_2_failures,expected_failures_in_20)
```

```
## [1] 0.2381033
```

Answer: 0.238 probability that at least 18 classes will have no mic failures.

QUESTION 1 c)

No, without further information, we would not be able to say for certain that the microphone issues had been resolved (nor could we say that the issues were still present). There is a chance that the issues were resolved as no mic failures had been observed, however the probabilities of a failure occurring support the observation of '0 failures'. Each class of 50 minutes is an independent event with an equal chance of a mic failure occurring. Although we theoretically expected about 4 classes (1-(20*81.87%)) to have microphone failures according to the answer from Q1)a)… The probability of 0 failures in 20 lessons is given by:

```
dpois(zero_failures,expected_failures_in_20)
```

```
## [1] 0.01831564
```

…this is a 1.83% chance of no mic failures. Although this is a low chance, the probability states that it is possible to not observe any failures in a given set of 20 lessons. Since we do not have information on future occurrences of mic failures, we cannot say that there will be no mic failures in the future. However, the chances of this occurring are low.

QUESTION 2 a)

To quantify the difference between the expected and observed sequence compositions, we can perform the chi squared hypothesis test for goodness of fit.

```
# data for this question
ave_comp<- c(0.25,0.25,0.25,0.25)
seq1_comp<- c(0.28,0.22,0.26,0.24)
seq2_comp<- c(0.27,0.23,0.26,0.24)
```

The H0: there is no significant difference between the expected and observed compositions, H1: there is a significant difference. We use 3 degrees of freedom (n-1), we use single-tailed assessment to see if there is any difference between expected and observed values. Chi squared tests require counts rather than percentages, so we convert % compositions to base counts.

```
ave_bases<-600*ave_comp #expected composition of bases
seq1_bases<-600*seq1_comp #observed composition of bases
```

to find our chi squared statistic, we take the sum of differences to the observed values divided by the expected values

```
chi_a<- sum((seq1_bases-ave_bases)^2/ave_bases)
chi_fit <- pchisq(chi_a,df=3,lower.tail = FALSE)
```

p value= 0.187, meaning a 18.7% chance of getting our observation when chance alone is the explanation for deviations from the expected value. So we cannot reject H0, the compositions are not significantly different. We assumed that the null hypothesis is true and that the chi-squared distribution is followed.

QUESTION 2 b)

The same process as above

```r
ave_bases_b<-2700*ave_comp
seq2_bases<-2700*seq2_comp
chi_b<- sum((seq2_bases-ave_bases_b)^2/ave_bases_b)
chi_fit_b<- pchisq(chi_b,df=3,lower.tail = FALSE)
```

p value= 0.0129, meaning a 1.29% chance of the composition in Sequence 2 occurring under H0. This leads to a rejection of H0, meaning that there is a significant difference between the expected composition and the observed one in Sequence 2. Although the composition is closer to the expected random sequence composition, the larger sample size of bases gives more power to the chi squared hypothesis test, so existing differences are revealed and a different conclusion is reached.

QUESTION 2 c)

H0: the frequencies of the bases are the same as a random sequence, H1: the frequencies are not the same as a random sequnece. To obtain a p-value<0.05, the chi squared test statistic would have to be

```r
chi_c<-qchisq(0.049,df=3, lower.tail = FALSE)
```

…7.86 or more. This test statistic represents a very small difference to the expected values of 25% for each base.

QUESTION 3 a)

False Discovery Rate (FDR) is given by: *False Positives/(False Positives + True Positives) or FP/(FP+TP)* FP+TP= 5995, so to find the FDR for these values, we can use the stated True Positive Rate (TPR) of 0.78 at alpha=0.1 to estimate the number of FP

```r
pred.inter<-5995
TPR<-0.78
TP<- TPR*pred.inter
FP<-pred.inter-TP
FDR<-FP/pred.inter
```

FDR is therefore estimated to be 0.22*100= 22%. This estimation could be improved if we knew the 'real' TP value rather than using the TPR, which is only confident at alpha=0.01. However, since this is a practical experiment, we cannot know the 'real' values for TP. If we had a complete list of p-values for each test, then we could estimate a more accurate FDR by eliminating the FP from the distribution of results using Storey's q-value, however, these are not given here.

QUESTION 3 b)

If total number of protein pairs is 28 SARS_Cov-2 proteins * 19804 encoded human proteins

```r
total_pairs<-28*19804
```

. . . .554512 pairs, and the total number of pairs predicted to have no interaction would be

```
pred.no_inter.<-total_pairs-pred.inter
```

. . . 548517 pairs.

The values for TPR and the estimated TP can be used to solve for False Negatives (FN), because TPR =TP/(TP+FN) rearranges to FN=(TP/TPR)-TP

```
FN<-(TP/TPR)-TP
```

. . . giving 1318.1 pairs which are FN and really positive. Adding this to the number of predicted interacting pairs gives the 'real' number of interactions:

```
real.inter<-FN+TP
```

. . . 5995 pairs 'really' interact. This is the same as the predicted number of interactions through the bioinformatic method in the question.

Since we do not have a complete list of the p-values from each test, we cannot use Storey's q-value to reach a more accurate FDR. The BH menthod for reaching an FDR would give a more conservative FDR:

```
BH_pred.inter.<-total_pairs*0.01
BH_FDR<-BH_pred.inter./real.inter
```

. . . This gives an FDR of 0.9250*100=92.5%, which gives leaves 450 TP interactions. This reduces the number of findings by quantifying the number of FP to be far higher, but doesn't improve the FDR.

QUESTION 3 c)

To improve the strategy in finding biomedically relevant interacting proteins, we could amend our method in different avenues.

On the practical side, we could filter the list of human encoded proteins to only include proteins encoded in tissues known to be affected by SARS-Cov-2 such as respiratory organs, specifically mucosal cells in the bronchial surface. Reducing the list of candidate proteins would give a smaller sample size and thus a correspondingly lower FDR at alpha=0.01.

Reducing the sample size may adversely affect statistical test power/ TDR, however it may be worth exploring to find a middle ground that trades off sensitivity and specificity. The new list of proteins may also express a larger effect size in these tissues due to SARS-Cov-2's mode of transmission.

On the statistical side, we could gather p-values for each hypothesis test run to enable the use of Storey's q-values and ultimately reduce the FDR, resulting in a more accurate number of interacting protein pairs.