Biostatistics Homework 3

By 𝕃umi (12112618)

```
library(ggplot2)
```

1 Rolling Die Again and Again

1.1)

```
dice_a <- c(1,2,3,4)
dice_b <- c(1,2,3,4)
possible_outcomes <- expand.grid(dice_a, dice_b)
colnames(possible_outcomes) <- c("dice_a", "dice_b")
possible_outcomes$Abs_dif <- abs(possible_outcomes$dice_a - possible_outcomes$dice_b)
print(possible_outcomes)
```

```
##    dice_a dice_b Abs_dif
## 1       1      1       0
## 2       2      1       1
## 3       3      1       2
## 4       4      1       3
## 5       1      2       1
## 6       2      2       0
## 7       3      2       1
## 8       4      2       2
## 9       1      3       2
## 10      2      3       1
## 11      3      3       0
## 12      4      3       1
## 13      1      4       3
## 14      2      4       2
## 15      3      4       1
## 16      4      4       0
```

We first print all the possible outcomes of rolling the two die simultaneously. Then we compute the absolute of the difference between the two numbers on the dice. We can see that there are four possible outputs for our random variable X, 0, 1, 2 and 3. We then count their occurrences.

```
p0 <- sum(possible_outcomes$Abs_dif == 0)
p1 <- sum(possible_outcomes$Abs_dif == 1)
p2 <- sum(possible_outcomes$Abs_dif == 2)
p3 <- sum(possible_outcomes$Abs_dif == 3)
prob <- data.frame(p0,p1,p2,p3)
colnames(prob) <- c("0", "1", "2", "3")
prob
```

```
##   0 1 2 3
## 1 4 6 4 2
```

So we can write the PMF for our random variable X.

$$P(X=x)=\begin{cases} \frac{1}{4} & \text{, if } x=0 \\ \frac{3}{8} & \text{, if } x=1 \\ \frac{1}{4} & \text{, if } x=2 \\ \frac{1}{8} & \text{, if } x=3 \\ 0 & \text{, otherwise} \end{cases}$$

The expected value, according to definition, is

$$E(X)=\sum xP_X(x)=0\times\frac{1}{4}+1\times\frac{3}{8}+2\times\frac{1}{4}+3\times\frac{1}{8}=\frac{5}{4}$$

The variance, according to deduction, is

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2)-E^2(X)=\sum x^2P_X(x)-(\sum xP_X(x))^2 \\ &= 0^2\times\frac{1}{4}+1^2\times\frac{3}{8}+2^2\times\frac{1}{4}+3^2\times\frac{1}{8}-(\frac{5}{4})^2 \\ &= \frac{15}{16} \end{aligned}$$

1.2)

The PMF for the function is,

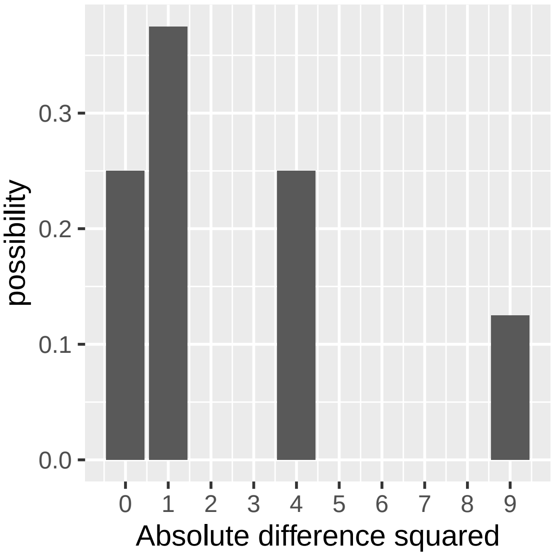
$$P(X^2=x^2)=\begin{cases} \frac{1}{4} & \text{, if } x^2=0 \\ \frac{3}{8} & \text{, if } x^2=1 \\ \frac{1}{4} & \text{, if } x^2=4 \\ \frac{1}{8} & \text{, if } x^2=9 \\ 0 & \text{, otherwise} \end{cases}$$

As shown in 1.1) the expected value of X^2 is

$$E(X^2)=\sum x^2P_X(x)=0^2\times\frac{1}{4}+1^2\times\frac{3}{8}+2^2\times\frac{1}{4}+3^2\times\frac{1}{8}=\frac{5}{2}$$

The plot for the PMF is,

```r
prob <- as.data.frame(t(prob))
colnames(prob) <- "frequency"
sum <- sum(prob[,1])
prob$possibility <- prob$frequency / sum
prob$x <- row.names(prob)
prob$x_squared <- as.numeric(prob$x) ^ 2
ggplot(data = prob,
       mapping = aes(x_squared, possibility)) +
  geom_col() +
  labs(x = "Absolute difference squared") +
  scale_x_continuous(n.breaks = 9)
```



2 Han.M.M and Li.L Playing Golf

2.1)

My instinct tells me that the score of the final 8 rounds follows a binomial distribution with $n = 8$ and p.

So the PMF for the final score of Han.M.M will be (Z indicates all integers)

$$p_H(h) = \binom{n}{h} p^h (1-p)^{n-h} = \binom{8}{h} p^h (1-p)^{8-h}$$
$$p_H(h) = \begin{cases} \binom{8}{h} p^h (1-p)^{8-h} & , h \in \{[0,8] \cap Z\} \\ 0 & , \text{otherwise} \end{cases}$$

The expected value and variance for a binomial distribution is simply

$$E(H) = np = 8p\text{Var}(H) = np(1-p) = 8p(1-p)$$

Oops, we don't need the variance but anyways.

2.2)

Similarly, we can see that the score of the final 8 rounds for Li.L also follows a binomial distribution with $n = 8$ and $1 - p$.

So the PMF for the final score of Li.L will be

$$p_L(l) = \binom{n}{l} p^{n-l} (1-p)^{l} = \binom{8}{l} p^{n-l} (1-p)^{l}$$
$$p_L(l) = \begin{cases} \binom{8}{l} p^{8-l} (1-p)^{l} & , l \in \{[0,8] \cap Z\} \\ 0 & , \text{otherwise} \end{cases}$$

The expected value will simply be

$$E(L) = n(1-p) = 8(1-p)$$

2.3)

For each round, the event of Han.M.M and Li.L winning are mutually exclusive. One winning means that the other one will definitely lose, so I am going to use Han.M.M's random variable to monitor the event.

If they end up with a draw after 8 rounds, Han.M.M will win another 5 rounds and Li.L another 3. So if Han.M.M wants to win, she will have to win more than 5 rounds. If Li.L wants to win, he will have to win more than 3 rounds, in other words, Han.M.M have to win less than 5 rounds.

Thus, suppose event A = H.M.M win, B = drawn, C = L.L win

$$\begin{aligned}
P(A) &= p_H(H > 5) = p_H(H = 6) + p_H(H = 7) + p_H(H = 8) \\
&= \binom{8}{6}p^6(1-p)^2 + \binom{8}{7}p^7(1-p)^1 + \binom{8}{8}p^8(1-p)^0 \\
P(B) &= p_H(H = 5) \\
&= \binom{8}{5}p^5(1-p)^3 \\
P(C) &= p_H(H < 5) = 1 - P(B) - P(A) \\
&= 1 - (\binom{8}{5}p^5(1-p)^3 + \binom{8}{6}p^6(1-p)^2 + \binom{8}{7}p^7(1-p)^1 + \binom{8}{8}p^8(1-p)^0)
\end{aligned}$$

2.4)

The random variables $X = \{\text{money H.M.M would get and}\}$ $Y = \{\text{money L.L would get}\}$ can be denoted as (with unit ¥)

$$X = \begin{cases} 50 & \text{, if H.M.M win} \\ 25 & \text{, if drawn} \\ 0 & \text{, if L.L win} \end{cases}$$

$$Y = 50 - X$$

Thus with the definition of expected values,

$$\begin{aligned}
E(X) &= 50 \times P(A) + 25 \times P(B) + 0 \times P(C) \\
&= 50 \times (\binom{8}{6}p^6(1-p)^2 + \binom{8}{7}p^7(1-p)^1 + \binom{8}{8}p^8(1-p)^0) + 25 \times \binom{8}{5}p^5(1-p)^3
\end{aligned}$$

Because $Y = 50 - X$ is a linear function, $E(Y) = 50 - E(X)$. Thus,

$$E(Y) = 50 - (50 \times (\binom{8}{6}p^6(1-p)^2 + \binom{8}{7}p^7(1-p)^1 + \binom{8}{8}p^8(1-p)^0) + 25 \times \binom{8}{5}p^5(1-p)^3)$$

3 Who is Correct?

3.1)

H.M.M's thoughts:

$$P_X(X = k) = \frac{\text{choose } k \text{ people from B type} \cdot \text{choose } (5-k) \text{ people from not-B type}}{\text{choose 5 people from 25 people}}$$

Thus mathematically, the PMF for H.M.M is:

$$P_X(X=k) = \begin{cases} \frac{\binom{5}{k}\binom{20}{5-k}}{\binom{25}{5}} & \text{, k = 0,1,...,5} \\ 0 & \text{, otherwise} \end{cases}$$

Which, based on my previous knowledge, is a hypergeometric distribution, and have a very large possibility of being correct. But never the less, let's look at L.L's thoughts, maybe he is right as well.

3.2)

L.L's thinks the chance of choosing blood type B from 25 person is an event with a fixed probability, which we know at a glance is wrong. The probability will change once you took a person with blood type B from the group. Never the less, here is his PMF.

$$P_X(X=k) = \begin{cases} \binom{5}{k}0.2^k 0.8^{5-k} & \text{, k = 0,1,2,...,5} \\ 0 & \text{, otherwise} \end{cases}$$

But, if we have an infinite amount of people, and what you know is the proportion of people having blood type B, then L.L is probably right.

3.3)

```r
calculations <- data.frame(c(0,1,2,3,4,5))
colnames(calculations) <- "k"
calculations$"Han.M.M_P(X=k)" <- dhyper(calculations$k, 5, 20, 5)
calculations$"Li.L_P(X=k)" <- dbinom(calculations$k, size = 5, prob = 0.2)
calculations
```

```
##   k Han.M.M_P(X=k) Li.L_P(X=k)
## 1 0   2.918125e-01     0.32768
## 2 1   4.559571e-01     0.40960
## 3 2   2.145680e-01     0.20480
## 4 3   3.576134e-02     0.05120
## 5 4   1.882176e-03     0.00640
## 6 5   1.882176e-05     0.00032
```

As we can see from the table, the two PMF give different results. The column Li.L_P(X=k) means the probability of $P_L(L=k)$, and the column Han.M.M_P(X=k) means the probability of $P_H(H=k)$. The difference doesn't seem to be very big when k is small, but as k gets larger, the difference is quite huge.

3.4)

As I have said in 3.2), I think Han.M.M have the correct answer because the number of possibility will change after you get a person with B blood type.

For example, consider the case of getting five people with blood type B. After you picked four people with blood type B from the class, there are 21 person left in the classroom, and only one of them have blood type B. Now, the possibility of getting a person with blood type B becomes:

$$P(\text{B type blood} \mid \text{getting 4 B type blood}) = \frac{1}{21} \approx 0.0476 \ll 0.2$$

So, brutally assigning 0.2 to all possibilities is wrong.

3.5)

We have 256 people, in which 50 have B type blood, when we choose 5 people, H.M.M's PMF is still correct, but I think as I predicted in 3.2), L.L's PMF results will not differ from H.M.M's in a very big amount. The probability of h taking any other value than $\{[0, 5] \cap Z\}$ are 0 (Z means all the integers).

$$P_H(H = h) = \frac{\binom{50}{h}\binom{206}{5-h}}{\binom{256}{5}} P_L(L = l) = \binom{5}{l}(\frac{50}{256})^l(1 - \frac{50}{256})^{5-l}$$

We can calculate the possibility using both methods again.

```r
more_people <- data.frame(c(0,1,2,3,4,5))
colnames(more_people) <- "k"
more_people$"Han.M.M_P(X=k)" <- dhyper(more_people$k, 50, 206, 5)
more_people$"Li.L_P(X=k)" <- dbinom(more_people$k, size = 5, prob = 50 / 256)
more_people
```

```
##   k Han.M.M_P(X=k)  Li.L_P(X=k)
## 1 0   0.3341671324 0.3373931611
## 2 1   0.4135731837 0.4094577198
## 3 2   0.1996560197 0.1987658834
## 4 3   0.0469778870 0.0482441465
## 5 4   0.0053852700 0.0058548721
## 6 5   0.0002405072 0.0002842171
```

As predicted, the difference between the two results are not that significance compared to the case with lesser students.

In this case, taking out five people from the total amount of 50 and 256 doesn't affect the general possibility of getting a B type blood person that much. The influence was drowned in the large amount of data.

4 Independence of Many Events

From the lecture, we know that if we want to check that some events are independent, we just need to calculate the probability of $P(\bigcap_{i=1}^{n} A_i)$ and see if it's the same as $\Pi_{i=1}^{n} P(A_i)$.

Thus we know that we are choosing 2,3,. . . ,n events and applying this equation to see if they are independent. Thus the number of times to check N, can be calculated as follows:

$$N = \sum_{i=2}^{n} \binom{n}{i}$$

I wanted to give up from this at the beginning, but I have an instinct that this can be simplified with a very easy trick. After a bit of thinking and research (mainly research), I found the answer. Because summing up all the possible ways of choosing items in a set is equivalent to dividing the set into two groups, so each item have only two possibilities of being in the group or not. Thus:

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

So, with a bit of fiddling, we can obtain:

$$N = 2^n - n - 1$$

5 Is This Really That Simple?

Because the signal sending are independent events, we can see that sending one character is just a Bernoulli Trial with the probability of success being $\frac{3}{7}$. Thus the event of sending 44 characters are just a binomial distribution with $n = 44$ and $p = \frac{3}{7}$.

Suppose the event of sending 22 characters successfully is A.

$$P(A) = \binom{44}{22}(\frac{3}{7})^{22}(\frac{4}{7})^{22}$$

With a bit of help from the computer, we know the answer is:

```
dbinom(22,44,3/7)
```

```
## [1] 0.07598709
```

So

$$P(A) \approx 0.07598709$$

6 Weird Disease

Because the sample of a large population of women shows result that 5% of them are positive for bacteriuria, we can assume that the chance of getting bacteriuria is 5%. Thus the chance of k people getting bacteriuria when we choose n people follows a binomial distribution.

6.1)

When we choose 5 people, suppose the random variable X is the number of people having bacteriuria. Then $X \sim \text{Bin}(n = 5, p = 0.05)$.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (\binom{5}{0}0.05^0 \times 0.95^5) \approx 0.2262190625$$

6.2)

When we choose 5 people, suppose the random variable X is the number of people having bacteriuria, then $X \sim \text{Bin}(n = 100, p = 0.05)$.

$$\begin{aligned} P(X \geq 3) &= 1 - (P(X = 0) + P(X = 1) + P(X = 2)) \\ &= 1 - (\binom{100}{0}0.05^0 \times 0.95^{100} + \binom{100}{1}0.05^1 \times 0.95^{99} + \binom{100}{2}0.05^2 \times 0.95^{98}) \\ &\approx 0.881737 \end{aligned}$$

6.3)

A simple "treeish" diagram of all the possible outcomes of the two tests. + indicates a positive result, $-$ indicates a negative result. The first bracket indicates the first test, the second bracket indicates the second. The number inside the open parenthesis indicates the possibility of each event.

$$\begin{cases} +(0.05) \begin{cases} +(0.2) \\ -(0.8) \end{cases} \\ -(0.95) \begin{cases} +(0.042) \\ -(0.958) \end{cases} \end{cases}$$

Thus, we can write the random variable for the two test, the number indicates the number of test positives.

$$X = \{0, 1, 2\}$$

Thus we can write the possibility mass function as:

$$P_X(X = k) = \begin{cases} 0.01 & \text{, k = 2} \\ 0.0799 & \text{, k = 1} \\ 0.9101 & \text{, k = 0} \\ 0 & \text{, otherwise} \end{cases}$$

6.4)

Expected value of X can be calculated with the definition,

$$E(X) = \sum_{k=0}^{2} kP_X(k) = 2 \times 0.01 + 1 \times 0.0799 + 0 \times 0.9101 = 0.0999$$

6.5)

The variance can be calculated with the small trick:

$$\text{Var}(X) = E(X^2) - E^2(X)$$

Thus we still need the $E(X^2)$, which can also be calculated from the definition:

$$E(X^2) = \sum_{k=0}^{2} k^2 P_X(k) = 4 \times 0.01 + 1 \times 0.0799 + 0 \times 0.9101 = 0.1199$$

$$\text{Var}(X) = 0.1199 - 0.0999^2 = 0.10992$$

7 Weird Disease

7.1)

We can see from the question that the episode of the disease follows a Poisson Distribution. Suppose the random variable X means the number of episodes in a certain time interval t. Then $X \sim P(\lambda t)$. When $t = 2, \lambda = 1.6$, we have:

$$\begin{aligned} P_X(X \geq 3) &= 1 - (P_X(X = 0) + P_X(X = 1) + P_X(X = 2)) \\ &= 1 - \sum_{i=0}^{2} \frac{(\lambda t)^i}{i!} e^{-\lambda t} \\ &= 1 - \sum_{i=0}^{2} \frac{(3.2)^i}{i!} e^{-3.2} \\ &\approx 0.62 \end{aligned}$$

7.2)

This time, we are having $X \sim P(\lambda t)$ with $t = 1, \lambda = 1.6$.

$$P_X(X = 0) = \frac{(\lambda)^0}{0!} e^{-\lambda} \approx 0.202$$

8 Lots of Poisson

8.1)

Because we already have a parameter λ for visitors per hour, thus, the PMF of A can be written as follows:

$$P(A = x) = \frac{(0.25\lambda)^x}{x!} e^{-0.25\lambda}, x > 0, x \in Z$$

8.2)

Similarly for t hours:

$$P(B = x) = \frac{(t\lambda)^x}{x!} e^{-t\lambda}, x > 0, x \in Z, t > 0$$

8.3)

Because we observed 5 visitors in the last 5 hours, we can assume that $\lambda = 5$. Thus:

$$P(C = x) = \frac{(5t)^x}{x!} e^{-5t}, x > 0, x \in Z, t > 0$$

8.4)

[tick] They are continuous random variables [tick] They all have the same type of distribution

8.5)

$$P(T_1 = t_1) = 0$$

8.6)

[tick] No arrivals in the time interval $[0, t_1]$

8.7)

Suppose the random variable Y counts the numbers of visitors on the website.

$$P(T_1 > t_1) = P_Y(Y = 0) = \frac{(t_1\lambda)^0}{0!} e^{-t_1\lambda} = e^{-t_1\lambda}$$

8.8)

$$F_{T_1}(t_1) = P(T_1 \leq t_1) = 1 - P(T_1 > t_1) = 1 - e^{-t_1\lambda}$$

8.9)

$$f_{T_1}(t_1) = \frac{d}{dt_1}(1 - e^{-t_1\lambda}) = \lambda e^{-t_1\lambda}$$

Very good question!!! I like it.

9 Γ Distributions

9.1)

[tick] No more than $\alpha - 1$ in the time interval $[0, w]$

9.2)

Suppose the random variable X records the number of visits of a website in a given interval of time.

$$P(W > w) = \sum_{i=0}^{\alpha-1} P(X = i) = \sum_{i=0}^{\alpha-1} \frac{(\lambda w)^i}{i!} e^{-\lambda w}$$

9.3)

$$F_W(w) = P(W < w) = 1 - P(W \geq w) = 1 - \sum_{i=0}^{a-1} \frac{(\lambda w)^i}{i!} e^{-\lambda w}$$

9.4)

Let's take the derivative!

$$\begin{aligned}
f_W(w) &= \frac{d}{dw}(1 - \sum_{i=0}^{\alpha-1} \frac{(\lambda w)^i}{i!} e^{-\lambda w}) \\
&= -\frac{d}{dw}(\sum_{i=0}^{\alpha-1} \frac{(\lambda w)^i}{i!} e^{-\lambda w}) \\
&= -\sum_{i=0}^{\alpha-1} \frac{d}{dw}(\frac{(\lambda w)^i}{i!} e^{-\lambda w}) \\
&= \sum_{i=0}^{\alpha-1}(\frac{\lambda^{i+1}}{i!} w^i e^{-\lambda w}) - \sum_{i=1}^{\alpha-1}(\frac{\lambda^i}{(i-1)!} w^{i-1} e^{-\lambda w}) \\
&= \frac{\lambda^\alpha}{(\alpha - 1)!} w^{\alpha-1} e^{-\lambda w}
\end{aligned}$$