BE601 - Data Analytics I

Seminar 2

Uyen Vu ([uyen.vu@phdstudent.hhs.se](mailto:uyen.vu@phdstudent.hhs.se))

13/08/2022

# Test Exercise

3.17 - p.114 A department store manager has monitored the number of complaints received per week about poor service. The probabilities for numbers of complaints in a week, established by this review, are shown in the following table. Let A be the event “there will be at least one complaint in a week” and B the event “there will be fewer than ten complaints in a week.”

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of complaints | 0 | 1 to 3 | 4 to 6 | 7 to 9 | 10 to 12 | More than 12 |
| Probability | 0.14 | 0.39 | 0.23 | 0.15 | 0.06 | 0.03 |

1. Find the probability of A.
2. Find the probability of B.
3. Find the probability of the complement of A.
4. Find the probability of the union of A and B.
5. Find the probability of the intersection of A and B.
6. Are A and B mutually exclusive?
7. Are A and B collectively exhaustive?

P\_A <- 1 - 0.14  
paste(c("a) P(A) is",P\_A), collapse = " ")

## [1] "a) P(A) is 0.86"

P\_B <- 1 - 0.06 - 0.03  
paste(c("b) P(B) is", P\_B), collapse = " ")

## [1] "b) P(B) is 0.91"

P\_NotA <- 0.14  
paste(c("c) P(NotA) is",P\_NotA), collapse = " ")

## [1] "c) P(NotA) is 0.14"

P\_A\_intersect\_B <- P\_B - P\_NotA  
P\_A\_union\_B <- P\_A + P\_B - P\_A\_intersect\_B  
paste(c("d) P(A\_union\_B) is",P\_A\_union\_B), collapse = " ")

## [1] "d) P(A\_union\_B) is 1"

paste(c("e) P(A\_intersect\_B) is",P\_A\_intersect\_B), collapse = " ")

## [1] "e) P(A\_intersect\_B) is 0.77"

P\_A\_intersect\_B == 0

## [1] FALSE

paste(c("f) Because P(A\_intersect\_B) is not equal to 0", "A and B are not mutually exclusive"), collapse = ", ")

## [1] "f) Because P(A\_intersect\_B) is not equal to 0, A and B are not mutually exclusive"

P\_A + P\_B == 1

## [1] FALSE

paste(c("g) Because P(A) + P(B) is not equal to 1", "A and B are not collectively exhaustive"), collapse = ", ")

## [1] "g) Because P(A) + P(B) is not equal to 1, A and B are not collectively exhaustive"

3.18 - page 115 A corporation receives a particular part in shipments of 100. Research indicated the probabilities shown in the accompanying table for numbers of defective parts in a shipment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number | 0 | 1 | 2 | 3 | >3 defective |
| Probability | 0.29 | 0.36 | 0.22 | 0.1 | 0.03 |

1. What is the probability that there will be fewer than three defective parts in a shipment?
2. What is the probability that there will be more than one defective part in a shipment?
3. The five probabilities in the table sum to 1. Why must this be so?

E1 <- "E1: Fewer than 3 defective parts in a shipment"  
E2 <- "E2: More than one defective part in a shipment"  
E1

## [1] "E1: Fewer than 3 defective parts in a shipment"

E2

## [1] "E2: More than one defective part in a shipment"

P\_E1 <- 0.29 + 0.36 + 0.22  
P\_E2 <- 0.29 + 0.36  
paste(c("P(E1) = ", P\_E1), collapse = " ")

## [1] "P(E1) = 0.87"

paste(c("P(E2) = ", P\_E2), collapse = " ")

## [1] "P(E2) = 0.65"

# Discussion exercises

## 46 - page 125

A conference began at noon with two parallel sessions. The session on portfolio management was attended by 40% of the delegates, while the session on chartism was attended by 50%. The evening session consisted of a talk titled “Is the Random Walk Dead?” This was attended by 80% of all delegates.

1. If attendance at the portfolio management session and attendance at the chartism session are mutually exclusive, what is the probability that a randomly chosen delegate attended at least one of these sessions?
2. If attendance at the portfolio management session and attendance at the evening session are statistically independent, what is the probability that a randomly chosen delegate attended at least one of these sessions?
3. Of those attending the chartism session, 75% also attended the evening session. What is the probability that a randomly chosen delegate attended at least one of these two sessions?

E1 <- "E1: Delegates attend portfolio management session"  
E2 <- "E2: Delegates attend chartism session"  
E3 <- "E3: Delegates attend evening session"  
P\_E1 <- 0.4  
P\_E2 <- 0.5  
P\_E3 <- 0.8  
#a) assume that E1 & E2 are mutually exclusive  
P\_E1\_union\_E2 <- P\_E1 + P\_E2  
paste(c("P(E1\_union\_E2) is", P\_E1\_union\_E2), collapse = " ")

## [1] "P(E1\_union\_E2) is 0.9"

#b) assume that E1 & E3 are statistically independent  
P\_E1\_intersect\_E3 <- P\_E1 \* P\_E3 #statistically independent  
E4 <- "E4: Delegates attend atleast one of 2 sessions E1 & E3"  
P\_E4 <- P\_E1 + P\_E3 - P\_E1\_intersect\_E3  
paste(c("P(E4) is", P\_E4), collapse = " ")

## [1] "P(E4) is 0.88"

#c) conditional probability  
P\_E3\_on\_E2 <- 0.75  
P\_E3\_intersect\_E2 <- P\_E3\_on\_E2 \* P\_E2 #multiplicative rule  
E5 <- "E5: Delegates attend atleast one of 2 sessions E2 & E3"  
P\_E5 <- P\_E2 + P\_E3 - P\_E3\_intersect\_E2  
paste(c("P(E5) is", P\_E5), collapse = " ")

## [1] "P(E5) is 0.925"

## 51 - page 126

An editor may use all, some, or none of three possible strategies to enhance the sales of a book:

1. An expensive pre-publication promotion
2. An expensive cover design
3. A bonus for sales representatives who meet predetermined sales levels

In the past, these three strategies have been applied simultaneously to only 2% of the company’s books. Twenty percent of the books have had expensive cover designs, and, of these, 80% have had expensive pre-publication promotion. A rival editor learns that a new book is to have both an expensive pre-publication promotion and an expensive cover design and now wants to know how likely it is that a bonus scheme for sales representatives will be introduced. Compute the probability of interest to the rival editor.

E1 <- "E1: Book has an expensive pre-publication promotion"  
E2 <- "E2: Book has an expensive cover design"  
E3 <- "E3: Book is associated with the bonus scheme for sales representatives"  
E4 <- "E4: Book has both expensive pre-publication promotion and an expensive cover design"  
  
P\_E1\_E2\_E3\_intersect <- 0.02  
P\_E2 <- 0.2  
P\_E1\_on\_E2 <- 0.8  
P\_E1\_intersect\_E2 <- P\_E1\_on\_E2 \* P\_E2 #multiplication rule  
P\_E4 <- P\_E1\_intersect\_E2  
P\_E3\_intersect\_E4 <- P\_E1\_E2\_E3\_intersect  
P\_E3\_on\_E4 <- P\_E3\_intersect\_E4 / P\_E4  
paste(c("The probability of interest to the rival editor is", P\_E3\_on\_E4), collapse = " ")

## [1] "The probability of interest to the rival editor is 0.125"

## E6

We define the events:

d1: item being defective

d2: item not being defective

c1: item being classified as defective

c2: item being classified as not\_defective

p\_d1 <- 0.05  
p\_d2 <- 0.95  
p\_c2\_on\_d2 <- 0.98  
p\_c1\_on\_d1 <- 0.95  
#a) Find p\_c1  
p\_c1\_intersect\_d1 <- p\_c1\_on\_d1 \* p\_d1  
p\_c2\_intersect\_d1 <- p\_d1 - p\_c1\_intersect\_d1  
p\_c2\_intersect\_d2 <- p\_c2\_on\_d2 \* p\_d2  
p\_c1\_intersect\_d2 <- p\_d2 - p\_c2\_intersect\_d2  
p\_c1\_intersect\_d1

## [1] 0.0475

p\_c2\_intersect\_d1

## [1] 0.0025

p\_c2\_intersect\_d2

## [1] 0.931

p\_c1\_intersect\_d2

## [1] 0.019

p\_c1 <- p\_c1\_intersect\_d1 + p\_c1\_intersect\_d2  
p\_c1

## [1] 0.0665

#b) Find p\_d1\_on\_c1  
p\_d1\_on\_c1 <- p\_c1\_intersect\_d1 / p\_c1  
p\_d1\_on\_c1

## [1] 0.7142857

We can calculate the marginal probability of each event by using two-way table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **D1** | **D2** | **Totals** |
| **C1** | 0.95\*0.05=0.0475 | 0.02\*0.95=0.019 | 0.0475+0.019=0.0665 |
| **C2** | 0.05\*0.05=0.0025 | 0.98\*0.95=0.931 | 0.0025+0.0931=0.9335 |
| **Totals** | 0.0475+0.0025=0.05 | 0.019+0.931=0.95 | 1 |

#Additional recommended exercises ## 40 A mail-order firm considers three possible events in filling an order: A: The wrong item is sent. B: The item is lost in transit. C: The item is damaged in transit. Assume that A is independent of both B and C and that B and C are mutually exclusive. The individual event probabilities are P(A) = 0.02, P(B) = 0.01, and P(C) = 0.04. Find the probability that at least one of these foul-ups occurs for a randomly chosen order.

1- 0.96 \* 0.99 \* 0.98

## [1] 0.068608

## 3.99

In a campus restaurant it was found that 35% of all customers order vegetarian meals and that 50% of all customers are students. Further, 25% of all customers who are students order vegetarian meals. a. What is the probability that a randomly chosen customer both is a student and orders a vegetarian meal? b. If a randomly chosen customer orders a vegetarian meal, what is the probability that the customer is a student? c. What is the probability that a randomly chosen customer both does not order a vegetarian meal and is not a student? d. Are the events “customer orders a vegetarian meal” and “customer is a student” independent? e. Are the events “customer orders a vegetarian meal” and “customer is a student” mutually exclusive? f. Are the events “customer orders a vegetarian meal” and “customer is a student” collectively exhaustive?

S1: Customer is student | S1: Customer is not student V1: Customer eats vegetarian | V2: Customer does not eat vegetarian

#35% of all customers order vegetarian meals and 50% of all customers are students.  
p\_v1 <- 0.35  
p\_v2 <- 1 - p\_v1  
p\_s1 <- 0.5  
p\_s2 <- 1 - p\_s1  
#25% of all customers who are students order vegetarian meals  
p\_v1\_on\_s1 <- 0.25  
p\_v2\_on\_s1 <- 1 - p\_v1\_on\_s1  
#a) randomly chosen customer both is a student and eats vegetarian: p\_s1\_intersect\_v1  
p\_s1\_intersect\_v1 <- p\_v1\_on\_s1 \* p\_s1  
p\_s1\_intersect\_v1

## [1] 0.125

#b) p\_s1\_on\_v1  
p\_s1\_on\_v1 <- p\_s1\_intersect\_v1 / p\_v1  
p\_s1\_on\_v1

## [1] 0.3571429

#c) p\_s2\_intersect\_v2 (we can draw a 2-way table)  
p\_s1\_intersect\_v2 <- p\_v2\_on\_s1 \* p\_s1  
p\_s2\_intersect\_v2 <- p\_v2 - p\_s1\_intersect\_v2  
p\_s2\_intersect\_v2

## [1] 0.275

#d) independent events?  
p\_v1 \* p\_s1 == p\_s1\_intersect\_v1 #FALSE so not independent

## [1] FALSE

#e) mutually exclusive?   
p\_s1\_intersect\_v1 == 0 #FALSE so not mutually exclusive

## [1] FALSE

#f) collectively exhaustive?  
p\_s1 + p\_v1 == 1 #FALSE so not collectively exhaustive

## [1] FALSE

## 3.102

A large corporation organized a ballot for all its workers on a new bonus plan. It was found that 65% of all night-shift workers favored the plan and that 40% of all female workers favored the plan. Also, 50% of all employees are night-shift workers and 30% of all employees are women. Finally, 20% of all night-shift workers are women. a. What is the probability that a randomly chosen employee is a woman in favor of the plan? b. What is the probability that a randomly chosen employee is either a woman or a night-shift worker (or both)? c. Is employee gender independent of whether the night shift is worked? d. What is the probability that a female employee is a night-shift worker? e. If 50% of all male employees favor the plan, what is the probability that a randomly chosen employee both does not work the night shift and does not favor the plan?

Define the events:

G1: worker is female | G2: worker is male

N1: worker works night-shift | N2: worker does not work night-shift

F1: worker favors the plan | F2: worker does not favor the plan

#65% of all night\_shift workers favored the plan  
p\_f1\_on\_n1 <- 0.65  
#40% of all female workers favored the plan  
p\_f1\_on\_g1 <- 0.4  
#50% of all employees are night-shift workers  
p\_n1 <- 0.5  
p\_n2 <- 1 - p\_n1 #complementary  
#30% of all employees are women  
p\_g1 <- 0.3  
p\_g2 <- 1 - p\_g1  
#20% of all night-shift workers are women  
p\_g1\_on\_n1 <- 0.2  
#a) a randomly chosen employee is a woman in favor of the plan: p\_g1\_intersect\_f1  
p\_g1\_intersect\_f1 <- p\_f1\_on\_g1 \* p\_g1 #multiplication rule  
p\_g1\_intersect\_f1

## [1] 0.12

#b) a randomly chosen employee is either a woman or a nigh-shift worker (or both): p\_g1\_intersect\_n1 & p\_g1\_union\_n1  
p\_g1\_intersect\_n1 <- p\_g1\_on\_n1 \* p\_n1  
p\_g1\_intersect\_n1

## [1] 0.1

p\_g1\_union\_n1 <- p\_g1 + p\_n1 - p\_g1\_intersect\_n1  
p\_g1\_union\_n1

## [1] 0.7

#c) employee gender independent of night shift is worked?  
p\_g1\_intersect\_n1 == p\_g1 \* p\_n1 #not independent

## [1] FALSE

#d) female employee is a night-shift worker: p\_n1\_on\_g1  
p\_n1\_on\_g1 <- p\_g1\_intersect\_n1/p\_g1  
p\_n1\_on\_g1

## [1] 0.3333333

#50% of all male employees favor the plan  
p\_f1\_on\_g2 <- 0.5  
#e) employee both does work the night shift and does not favor the plan: p\_n2\_intersect\_f2  
p\_g2\_intersect\_f1 <- p\_f1\_on\_g2 \* p\_g2  
p\_f1 <- p\_g1\_intersect\_f1 + p\_g2\_intersect\_f1  
p\_n1\_intersect\_f1 <- p\_f1\_on\_n1 \* p\_n1  
p\_n2\_intersect\_f1 <- p\_f1 - p\_n1\_intersect\_f1  
p\_n2\_intersect\_f2 <- p\_n2 - p\_n2\_intersect\_f1  
p\_n2\_intersect\_f2

## [1] 0.355

## 3.101

f1: employee is female | f2: employee is male g: employee had graduate degree | u: employee had undergraduate degree | h: employee had high school degree

#80% of the employees are men & 20% of the employees are women  
p\_f1 <- 0.2  
p\_f2 <- 0.8  
#men: 10% had graduate, 30% had undergraduate, 60% had high school  
p\_g\_on\_f2 <- 0.1  
p\_u\_on\_f2 <- 0.3  
p\_h\_on\_f2 <- 0.6  
#women: 15% had graduate, 40% had undergraduate, 45% had high school  
p\_g\_on\_f1 <- 0.15  
p\_u\_on\_f1 <- 0.4  
p\_h\_on\_f1 <- 0.45  
#a) p\_f2\_intersect\_h  
p\_f2\_intersect\_h <- p\_h\_on\_f2 \* p\_f2  
p\_f2\_intersect\_h

## [1] 0.48

#b) p\_g  
p\_g\_intersect\_f1 <- p\_g\_on\_f1 \* p\_f1  
p\_g\_intersect\_f2 <- p\_g\_on\_f2 \* p\_f2  
p\_g <- p\_g\_intersect\_f1 + p\_g\_intersect\_f2  
p\_g

## [1] 0.11

#c) p\_f2\_on\_g  
p\_f2\_on\_g <- p\_g\_intersect\_f2 / p\_g  
p\_f2\_on\_g

## [1] 0.7272727

#d) independent events?  
p\_f1 \* p\_g == p\_g\_intersect\_f1

## [1] FALSE

#e) p\_f1\_on\_k (k is the union of u & h)  
p\_k\_on\_f1 <- p\_u\_on\_f1 + p\_h\_on\_f1  
p\_k\_on\_f2 <- p\_u\_on\_f2 + p\_h\_on\_f2  
p\_f1\_on\_k\_nominator <- p\_k\_on\_f1 \* p\_f1  
p\_f1\_on\_k\_denominator <- p\_k\_on\_f1 \* p\_f1 + p\_k\_on\_f2 \* p\_f2  
p\_f1\_on\_k <- p\_f1\_on\_k\_nominator / p\_f1\_on\_k\_denominator  
p\_f1\_on\_k

## [1] 0.1910112

## E7

s1: The email is spam | s2: The email is not spam A: The email contains the word “Act” B: The email contains the word “Buy” C: The email contains the word “Call”

#There are 25 spam email in 100 emails  
p\_s1 <- 0.25  
p\_s2 <- 0.75  
#In 25 spam email, we have 20 with "Act", 15 with "Buy" and 10 with "Call"  
p\_a\_on\_s1 <- 20/25  
p\_b\_on\_s1 <- 15/25  
p\_c\_on\_s1 <- 10/25  
#In 75 NOT\_spam email, we have 3 with "Act", 6 with "Buy" and 9 with "Call"  
p\_a\_on\_s2 <- 3/75  
p\_b\_on\_s2 <- 6/75  
p\_c\_on\_s2 <- 9/75  
#a) we are looking for: p\_s1\_on\_a  
p\_s1\_intersect\_a <- p\_a\_on\_s1 \* p\_s1  
p\_a <- p\_a\_on\_s1 \* p\_s1 + p\_a\_on\_s2 \* p\_s2  
p\_s1\_on\_a <- p\_s1\_intersect\_a / p\_a  
p\_s1\_on\_a

## [1] 0.8695652

## Because p\_s1\_on\_a < 0.9, this email should not be classified as spam  
#b) we are looking for: p\_s1\_on\_m (m is the intersection of a and c)  
p\_s2\_intersect\_a <- p\_a\_on\_s2 \* p\_s2  
p\_a <- p\_s1\_intersect\_a + p\_s2\_intersect\_a  
p\_s1\_intersect\_c <- p\_c\_on\_s1 \* p\_s1  
p\_s2\_intersect\_c <- p\_c\_on\_s2 \* p\_s2  
p\_c <- p\_s1\_intersect\_c + p\_s2\_intersect\_c  
p\_a\_intersect\_c <- p\_a \* p\_c # a & c are statistically independent  
p\_m <- p\_a\_intersect\_c  
p\_s1\_intersect\_m <- p\_m \* p\_s1  
p\_s1\_on\_m <- p\_s1\_intersect\_m / p\_m  
p\_s1\_on\_m

## [1] 0.25

#c) we are looking for: p\_s1\_on\_n (n is the intersection of a, b and c)  
p\_s1\_intersect\_b <- p\_b\_on\_s1 \* p\_s1  
p\_s2\_intersect\_b <- p\_b\_on\_s2 \* p\_s2  
p\_b <- p\_s1\_intersect\_b + p\_s2\_intersect\_b  
p\_a\_intersect\_b\_intersect\_c <- p\_a \* p\_b \* p\_c  
p\_n <- p\_a\_intersect\_b\_intersect\_c  
p\_s1\_intersect\_n <- p\_n \* p\_s1  
p\_s1\_on\_n <- p\_s1\_intersect\_n / p\_n  
p\_s1\_on\_n

## [1] 0.25