BM20A9301 Statistics – Exercise set 4

To be done by 29.1.–2.2.2024

Text in blue or red is not part of the problem or its solution. It's there as extra information to help you learn.

Exercise 1 (Joint distribution)**.** Let S be the set of all students in class, $|S| = 60$, and $H \subseteq S$ is the set of students who have done their homework already, $|H| = 5$. Let's interview two students at random and let

$$X_1 = \begin{cases} 1, & \text{if the first student} \in H \\ 0, & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{if the second student} \in H \\ 0, & \text{otherwise} \end{cases}$$

(a) Assuming that you pick the students with replacement (=the second student is selected from all the students), calculate the joint distribution of (X_1, X_2).

(b) Assuming that you pick the students without replacement (=the second student is selected from the students except the first student), calculate the joint distribution of (X_1, X_2).

(c) Calculate the marginal distributions in both cases.

(d) Are X_1 and X_2 independent in (a)? In (b)?

Solution.

(a) Let's calculate $P(X_1 = 1, X_2 = 1)$ first. This event happens if the first student selected randomly among 60 has done homework, and if also the second student selected randomly among the 60 has done the homework. We have $P(X_1 = 1, X_2 = 1) = P(X_1 = 1) \cdot P(X_2 = 1|X_1 = 1)$ by the general product rule. But since we do the sampling with replacement (with bad luck we could select the first student again), we have $P(X_2 = 1|X_1 = 1) = P(X_2 = 1)$. In words: "after asking the first student, they'll go back into the pool of all students, so the first selection doesn't affect the second one."

Hence we have (by equal probability for each student) that $P(X_1 = 1, X_2 = 1) = 5/60 \cdot 5/60$. The probability of $X_1 = 1, X_2 = 0$ is calculated similarly, with $5/60$ probability for the 1st student to have done homework and $55/60$ for the second to not have done it. By independence $P(X_1 = 1, X_2 = 0) = 5/60 \cdot 55/60$. Similarly we see

		X_2	
		0	1
X_1	0	$55/60 \cdot 55/60$	$55/60 \cdot 5/60$
	1	$5/60 \cdot 55/60$	$5/60 \cdot 5/60$

(b) This time we guarantee that the first student will not be picked when we pick the second student. By the general product rule $P(X_1 = 1, X_2 = 1) = P(X_1 = 1) \cdot P(X_2 = 1|X_1 = 1)$. Again, $P(X_1 = 1) = {}^{5}/_{60}$. But now on the condition that $X_1 = 1$ (=the first student did do homework), we have the conditional probability $P(X_2 = 1|X_1 = 1) = {}^{4}/_{59}$ (there are 59 students left, and only 4 of them have done homework). Also $P(X_2 = 0|X_1 = 1) = {}^{55}/_{59}$. In a similar way we see that in this case

		X_2	
		0	1
X_1	0	${}^{55}/_{60} \cdot {}^{54}/_{59}$	${}^{55}/_{60} \cdot {}^{5}/_{59}$
	1	${}^{5}/_{60} \cdot {}^{55}/_{59}$	${}^{5}/_{60} \cdot {}^{4}/_{59}$

(c) The marginal distributions are calculated as follows

		X_2		
		0	1	Sum
X_1	0	${}^{55}/_{60} \cdot {}^{55}/_{60}$	${}^{55}/_{60} \cdot {}^{5}/_{60}$	${}^{55}/_{60}$
	1	${}^{5}/_{60} \cdot {}^{55}/_{60}$	${}^{5}/_{60} \cdot {}^{5}/_{60}$	${}^{5}/_{60}$
	Sum	${}^{55}/_{60}$	${}^{5}/_{60}$	1

		X_2		
		0	1	Sum
X_1	0	${}^{55}/_{60} \cdot {}^{54}/_{59}$	${}^{55}/_{60} \cdot {}^{5}/_{59}$	${}^{55}/_{60}$
	1	${}^{5}/_{60} \cdot {}^{55}/_{59}$	${}^{5}/_{60} \cdot {}^{4}/_{59}$	${}^{5}/_{60}$
	Sum	${}^{55}/_{60}$	${}^{5}/_{60}$	1

Hence, case (a)'s marginals:

j	0	1
$P(X_1 = j)$	${}^{55}/_{60}$	${}^{5}/_{60}$

j	0	1
$P(X_2 = j)$	${}^{55}/_{60}$	${}^{5}/_{60}$

and case (b)'s marginals:

j	0	1
$P(X_1 = j)$	${}^{55}/_{60}$	${}^{5}/_{60}$

j	0	1
$P(X_2 = j)$	${}^{55}/_{60}$	${}^{5}/_{60}$

(d) X_1 and X_2 are independent in (a) because the joint distribution $P(X_1 = j, X_2 = k)$ is the product of the marginals $P(X_1 = j) \cdot P(X_2 = k)$. In (b) they are not independent because $P(X_1 = j, X_2 = k) \neq P(X_1 = j) \cdot P(X_2 = k)$.

Exercise 2 (Expected value & variance definition)**.** Calculate the expected values and variances of the random variables A, B, C and D with the following distributions:

k	1	2
$P(A = k)$	0.9	0.1

k	1	2
$P(B = k)$	0.1	0.9

k	6	7
$P(C = k)$	0.9	0.1

k	1	2	3	4
$P(D = k)$	0.59	0.15	0.21	0.05

Solution. Let's calculate the expected value first. For a discrete r.v. X we have $E(X) = \sum_x xP(X = x)$ where we sum over all the possible values that X can have. Then

$$E(A) = 1 \cdot 0.9 + 2 \cdot 0.1 = 1.1$$

$$E(B) = 1 \cdot 0.1 + 2 \cdot 0.9 = 1.9$$

$$E(C) = 6 \cdot 0.9 + 7 \cdot 0.1 = 6.1$$

$$E(D) = 1 \cdot 0.59 + 2 \cdot 0.15 + 3 \cdot 0.21 + 4 \cdot 0.05 = 1.72$$

The variance is the expected value of the squared deviation from the expected value of the original random variable, $\mathrm{Var}(X) = \sum_x (x - E(X))^2 P(X = x)$. Hence

$$\mathrm{Var}(A) = (1 - 1.1)^2 \cdot 0.9 + (2 - 1.1)^2 \cdot 0.1 = 0.09$$

$$\mathrm{Var}(B) = (1 - 1.9)^2 \cdot 0.1 + (2 - 1.9)^2 \cdot 0.9 = 0.09$$

$$\mathrm{Var}(C) = (6 - 6.1)^2 \cdot 0.9 + (7 - 6.1)^2 \cdot 0.1 = 0.09$$

$$\mathrm{Var}(D) = (1-1.72)^2 \cdot 0.59 + (2-1.72)^2 \cdot 0.15 + (3-1.72)^2 \cdot 0.21 + (4-1.72)^2 \cdot 0.05 = 0.9216$$

Exercise 3 (Expected value usage)**.** In Finland, you can buy a lottery ticket called "Ässä–arpa". It costs 4 euros. Every batch has 3 000 000 printed tickets. The amount (in euros) and total number (N) of prize categories are as follows:

Euros	N
100 000	5
2 000	40
1 000	160
500	1 000
30	16 000
20	80 000
10	180 000
5	240 000
4	250 000

What is the expected **net** win in the case of buying one ticket?

Solution. Let W be a r.v. that tells us how much we win (between 0 to 250 000). Then the net win is $W - 4$ because the ticket costs 4 euros. Recall that for expected values $E(W - 4) = E(W) - 4$. To calculate $E(W)$ we need to know what is the probability to get various prizes. Since no extra information is given it is ok to assume that getting any lottery ticket is just as likely as any other. There are

3 000 000 tickets. So we must divide the numbers in the table above by the total quantity to get the probabilities.

The r.v. is discrete, so we get

$$\begin{aligned} E(W) &= \frac{100\,000}{3\,000\,000}\cdot 5 + \frac{2\,000}{3\,000\,000}\cdot 40 + \frac{1\,000}{3\,000\,000}\cdot 160 \\ &+ \frac{500}{3\,000\,000}\cdot 1\,000 + \frac{30}{3\,000\,000}\cdot 16\,000 + \frac{20}{3\,000\,000}\cdot 80\,000 \\ &+ \frac{10}{3\,000\,000}\cdot 180\,000 + \frac{5}{3\,000\,000}\cdot 240\,000 + \frac{4}{3\,000\,000}\cdot 250\,000 \\ &= \frac{7\,320\,000}{3\,000\,000} \end{aligned}$$

Then

$$E(W-4) = E(W) - 4 = \frac{7\,320\,000}{3\,000\,000} - 4 = -\frac{4\,680\,000}{3\,000\,000} = -\frac{39}{25} = -1.56$$

In other words, on average we would lose 1.56 euros for each ticket bought.

Exercise 4 (Simulating a probability)**.** Using Excel, RStudio, Python or whatever you want, estimate the following probability:

We roll three dice. What is the probablity the the largest number is at least 3 higher than the lowest number. For example if the rolls are $(2, 5, 4)$ then $5 - 2 \geq 3$ satisfies this, but in the case $(2, 4, 4)$ it is not true since $4 - 2 = 2 < 3$.

Recall the frequentist interpretation of probability: that probability of an event, is the number of times the event happens divided by the total number of times the random experiment is repeated when the number of repetitions grows very large.

Extra challenge: can you calculate the probability by hand?

Solution. We will explain it using *R*.

1. Let's roll 3 dice 100 000 times:

```
> n <- 100000
> x <- replicate(3, sample(c(1,2,3,4,5,6), n, replace=TRUE))
```

2. Define the r.v. D that measures the difference of the max and min:

```
> D <- pmax(x[,1], x[,2], x[,3]) - pmin(x[,1], x[,2], x[,3])
```

3. Make a PDF of the different walues of D:

```
> f <- table(D)
> f <- f/sum(f)
```

4. Check that `f` looks like expected.

```
> f
D
      0       1       2       3       4       5
0.02831 0.14001 0.22204 0.25123 0.22176 0.13665
```

It is the estimated probability density function for the discrete r.v. D. So $P(D \geq 3) = f(3) + f(4) + f(5)$ which will be given by $0.25123 + 0.22176 + 0.13665 = 0.60964 \approx 61.0\%$.

Solution to extra challenge. Let's list all $6^3 = 216$ possible rolls of three dice, calculate D after them, and let's circle those that satisfy $D \geq 3$ i.e. $D = 3$, $D = 4$ or $D = 5$.

111:0
112:1
113:2
[✓ 114:3]
[✓ 115:4]
[✓ 116:5]
121:1
122:1
123:2
[✓ 124:3]
[✓ 125:4]
[✓ 126:5]
131:2
132:2
133:2
[✓ 134:3]
[✓ 135:4]
[✓ 136:5]
[✓ 141:3]
[✓ 142:3]
[✓ 143:3]
[✓ 144:3]
[✓ 145:4]
[✓ 146:5]
[✓ 151:4]
[✓ 152:4]
[✓ 153:4]
[✓ 154:4]
[✓ 155:4]
[✓ 156:5]
[✓ 161:5]
[✓ 162:5]
[✓ 163:5]
[✓ 164:5]
[✓ 165:5]
[✓ 166:5]
211:1
212:1
213:2
[✓ 214:3]
[✓ 215:4]
[✓ 216:5]
221:1
222:0
223:1
224:2
[✓ 225:3]
[✓ 226:4]
231:2
232:1
233:1
234:2
[✓ 235:3]
[✓ 236:4]
[✓ 241:3]
242:2
243:2
244:2
[✓ 245:3]
[✓ 246:4]
[✓ 251:4]
[✓ 252:3]
[✓ 253:3]
[✓ 254:3]
[✓ 255:3]
[✓ 256:4]
[✓ 261:5]
[✓ 262:4]
[✓ 263:4]
[✓ 264:4]
[✓ 265:4]
[✓ 266:4]
311:2
312:2
313:2
[✓ 314:3]
[✓ 315:4]
[✓ 316:5]
321:2
322:1
323:1
324:2
[✓ 325:3]
[✓ 326:4]
331:2
332:1
333:0
334:1
335:2
[✓ 336:3]
[✓ 341:3]
342:2
343:1
344:1
345:2
[✓ 346:3]
[✓ 351:4]
[✓ 352:3]
353:2
354:2
355:2
[✓ 356:3]
[✓ 361:5]
[✓ 362:4]
[✓ 363:3]
[✓ 364:3]
[✓ 365:3]
[✓ 366:3]
[✓ 411:3]
[✓ 412:3]
[✓ 413:3]
[✓ 414:3]
[✓ 415:4]
[✓ 416:5]
[✓ 421:3]
422:2
423:2
424:2
[✓ 425:3]
[✓ 426:4]
[✓ 431:3]
432:2
433:1
434:1
435:2
[✓ 436:3]
[✓ 441:3]
442:2
443:1
444:0
445:1
446:2
[✓ 451:4]
[✓ 452:3]
453:2
454:1
455:1
456:2
[✓ 461:5]
[✓ 462:4]
[✓ 463:3]
464:2
465:2
466:2
[✓ 511:4]
[✓ 512:4]
[✓ 513:4]
[✓ 514:4]
[✓ 515:4]
[✓ 516:5]
[✓ 521:4]
[✓ 522:3]
[✓ 523:3]
[✓ 524:3]
[✓ 525:3]
[✓ 526:4]
[✓ 531:4]
[✓ 532:3]
533:2
534:2
535:2
[✓ 536:3]
[✓ 541:4]
[✓ 542:3]
543:2
544:1
545:1
546:2
[✓ 551:4]
[✓ 552:3]
[✓ 553:2]
554:1
555:0
556:1
[✓ 561:5]
[✓ 562:4]
[✓ 563:3]
564:2
565:1
566:1
[✓ 611:5]
[✓ 612:5]
[✓ 613:5]
[✓ 614:5]
[✓ 615:5]
[✓ 616:5]
[✓ 621:5]
[✓ 622:4]
[✓ 623:4]
[✓ 624:4]
[✓ 625:4]
[✓ 626:4]
[✓ 631:5]
[✓ 632:4]
[✓ 633:3]
[✓ 634:3]
[✓ 635:3]
[✓ 636:3]
[✓ 641:5]
[✓ 642:4]
[✓ 643:3]
644:2
645:2
646:2
[✓ 651:5]
[✓ 652:4]
[✓ 653:3]
654:2
655:1
656:1
[✓ 661:5]
[✓ 662:4]
[✓ 663:3]
664:2
665:1
666:0

There are 132 circled outcomes and 216 in total. So the probability is $P(D \geq 3) = 132/216 = 0.6111\ldots \approx 61.1\%$.

We can also calculate the distribution of D by seeing how many times each value occurs:

d	0	1	2	3	4	5
$P(D = d)$	$^6/_{216}$	$^{30}/_{216}$	$^{48}/_{216}$	$^{54}/_{216}$	$^{48}/_{216}$	$^{30}/_{216}$

Exercise 5 (Distributions)**.** Which distribution will model the following random variables' distribution? For example, if you think that X has the binomial distribution with parameters $n = 5$ and $p = 0.3$ write $X \sim \mathit{bin}(5, 0.3)$. Be ready to explain why. Then use a formula, Excel, distribution tables or RStudio to answer the question. **Hint:** Recall that the slides from lecture 03 listed functions in R which calculate the PDF or CDF of various distributions.

(a) W is the waiting time for the next train in minutes. You know that the train will come at the earliest in 5 minutes but no later than after 10 minutes. That's all you know. What is the probability the bus will arrive before 8 minutes pass?

(b) You maintain a server. Based on its specifications it can handle a maximum of 100 requests every second. Every second, what is the probability that it will crash? The requests R are coming in at around 70 a second on average ($E(R) = 70$) and are independent from each other.

(c) You roll a die 60 times and are interested in how many times you roll a six. What is the probability that you roll a six exactly 10 times? What about 10 times or less?

Solution.

(a) We have no reason to believe that the train would come more likely in any part of the interval $[5, 10]$ than any other of similar size. This means that it has the uniform distribution. Then $W \sim U(5, 10)$. The location where $W \leq 8$ is the subinterval $[5, 8] \subseteq [5, 10]$. The probability is the quotient of their size, namely $P(W \leq 8) = \frac{8-5}{10-5} = 3/5 = 60\%$.

(b) The requests are independent of each other and have an average rate $\lambda = 70$, and we are interested in their number. The Poisson distribution gives this, namely $R \sim Poisson(70)$ and so $f_R(r) = \frac{1}{r!}70^r e^{-70}$. The answer to the question is $P(R > 100)$. We have $P(R > 100) = 1 - P(R \leq 100) = 1 - F_R(100)$ where F_R is the cumulative distribution function. It has no simple formula, so we have to look it from tables or use `1-ppois(100, lambda=70)` in R, which gives

$$P(R > 100) = 1 - F_R(100) \approx 0.0003 = 0.03\%.$$

Extra info: That was the answer for the next second. If this rate of requests lasts an hour, one could ask what's the probaility of crashing then. An hour has 3600 seconds, each second there is a 0.03% chance of crashing, and since the requests are independent the chance of crashing between different minutes are also independent. Then we can use the binary distribution to answer that: $n = 3600$, $p = 0.0003$. Then $k = 0$ means no crashes so `1-dbinom(0, size=3600, p=0.0003)` gives the answer. It is $\approx 66\%$.

(c) Here we repeat an experiment 60 times, and each expermiment is independent of every other one. So it's the binary distribution. The number of repetitions is $n = 60$, and the probability of a success is $p = 1/6$. If N is the number of sizes rolled, we have $N \sim Bin(60, 1/6)$. Then

$$P(N = 10) = \binom{n}{10} p^{10}(1-p)^{n-10} = \binom{60}{10} \frac{1}{6^{10}} \cdot \frac{5^{50}}{6^{50}} \approx 0.137 = 13.7\%.$$

For 10 times or less we can use the cumulative distribution function $F_N(10)$ (which doesn't have a simple formula), or the sum of the probability density functions[1] $f_N(0) + f_N(1) + \cdots + f_N(10)$. Let's use R to calculate $F_N(10)$, using the function `pbinom(10,n=60,p=1/6)`. We get

$$P(N \leq 10) = F_N(10) \approx 0.583 = 58.3\%.$$

[1] Some call this the probability mass function when the r.v. is discrete.

Exercise 6 (Normal distribution)**.** To become a member of the high IQ score association Mensa, you need to have an IQ higher than 98% of the population. The population's mean IQ is defined to be 100. If you score more than 130 in the IQ test you will be offered membership. Assuming that the population's IQ scores are normally distributed, what is the standard deviation?

Solution. Let X represent the IQ of a random person selected from the population. Then we have $X \sim N(\mu, \sigma)$ with $\mu = 100$ and σ unknown. Instead we know that a random person has only a 2% chance of having higher IQ than 130, in other words: $P(X > 130) = 0.02$. If

$$Z = \frac{X - \mu}{\sigma}$$

then $Z \sim N(0, 1)$ and the event $\{X > 130\}$ is the same event as $\{Z > (130-100)/\sigma\}$. Hence $P(Z > 30/\sigma) = 0.02$. This implies $P(Z \leq 30/\sigma) = 1 - P(Z > 30/sigma) = 0.98$. By the definition of the cumulative distribution function Φ for the standard Gaussian, $\Phi(30/\sigma) = 0.98$. Hence $30/\sigma = \Phi^{-1}(0.98)$ so

$$\sigma = \frac{30}{\Phi^{-1}(0.98)} \approx 14.61$$

where $\Phi^{-1}(0.98)$ was calculated using `qnorm(0.98)` in R.