

# Homework #1

Winter 2020, STATS 509

Dino Bektesevic

## Problem 1

- Mutually exclusive implies  $P(A \cup B) = 1$  and we know probabilities are additive, i.e.  $P(A) + P(B) = P(A \cup B) \leq 1$  and problem tells us this doesn't hold.
- First determine  $P(B^C)$  as  $P(B) + P(B^C) = 1 \rightarrow P(B^C) = 1 - 0.45 = 0.55$  and then, in a similar manner as previous problem, conclude it's not possible since  $P(A) + P(B) = P(A \cup B) > 1$
- Same as above, determine  $P(A^C) = 0.4$  and note that  $P(A^C \cup B) = 0.95 < 1$  so they could be mutually exclusive.
- No reason they could not be independent.
- No because they are not mutually exclusive as shown in first-subproblem.

## Problem 2

- Problem states that the events are overlapping sets ( $P(A \cap B) = 0.06$ )
- From 2nd axiom of probability:  $P(A \cup B) = P(A) + P(B) - P(A \cap B) \rightarrow P(B) = P(A \cup B) - P(A) + P(A \cap B) = 0.2$
- Since we know  $P(A) = 0.3$  and have calculated  $P(B) = 0.2$  we can verify  $P(A)P(B) = 0.06 = P(A \cap B)$  and conclude that yes, they could be independent.

## Problem 3

We can start with 2nd axiom of probabilities again  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  and notice that because  $Z$  takes only integer values we have  $P(Z < 1 \cap Z \geq 0) = P(Z = 0)$  which is exactly what we are looking for. Additional information we have is that  $P(Z < 1 \cup Z \geq 0) = 1$  since the union spans the entire  $\mathbb{Z}$ , i.e. the domain of the random variable  $Z$ .

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\P(Z < 1 \cup Z \geq 0) &= P(Z < 1) + P(Z \geq 0) - P(Z = 0) \\P(Z = 0) &= P(Z < 1) + P(Z \geq 0) - P(Z < 1 \cup Z \geq 0) \\P(Z = 0) &= a + b - 1\end{aligned}$$

## Problem 4

a. We have a system of equations:

$$\begin{aligned} P(A \cup B) &= r = P(A) + P(B) - P(A \cap B) \\ P(A \cup C) &= s = P(A) + P(C) - P(A \cap C) \\ P(A \cup B \cup C) &\leq 1 \geq P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) - P(A \cap B \cap C) \end{aligned}$$

Multiplying the first two with a -1 and adding them we have:

$$\begin{aligned} -r - s &= -2P(A) - P(B) + P(A \cap B) + P(A \cap C) \\ 1 &\geq P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

Maximizing the  $P(A)$  is assuming probabilities add up to 1, since we are looking for extrema that's what we do. Adding what's left together:

$$\begin{aligned} 1 - r - s &= -P(A) - P(B \cap C) + P(A \cap B \cap C) \\ P(A) &= r + s + P(A \cap B \cap C) - P(B \cap C) - 1 \end{aligned}$$

b. Following above, we have:  $P(A) = r + s + t - 1 - P(B \cap C)$ .

Mathematically I was unable to further resolve the problem. Attending office hours a proverb was shared with me about a dog that barks but isn't, but I didn't understand it - refer to professor Thomas for details. However, it makes sense that since  $P(A) \leq P(A \cup B)$  and  $P(A) \leq P(A \cup C)$ , by implication the maximal probability for both sections of the problem is limited to  $\min(s, r)$ . For a) part of the problem there is nothing that really limits our lower bound so we must say  $0 \leq P(A) \leq \min(s, r)$  while in the b) part of the problem we can say  $t \leq P(A) \leq \min(s, r)$  since, by implication again,  $P(A) \geq P(A \cap B \cap C)$ .

## Problem 5

a. b. c.

$t_1$	$t_2$	$\text{avg}(Y_T)$	$\text{avg}(Y_C)$	$\text{avg}(Y_T) - \text{avg}(Y_C)$
1	2	6	4.5	1.5
1	3	4	6	-2
1	4	5.5	3.5	2
2	1	6	4.5	1.5
2	3	8	5	3
2	4	9.5	2.5	7
3	1	4	6	-2
3	2	8	5	3
3	4	7.5	4	3.5
4	1	5.5	3.5	2
4	2	9.5	2.5	7
4	3	7.5	4	3.5

- e. The average of the differences of hypothetical experiments is the mean value of the last column, i.e. 2.5.
- f. The average of given  $\langle Y_T \rangle = 6.75$  and the average of given  $\langle Y_C \rangle = 4.25$ , this their difference is again 2.5
- g. The average differences are the same. This is the basis on which medical randomized trials are based - it is expected that the difference of averages of the two groups match the averages of individuals.

## Problem 6

- a. The probability of selecting a non-diseased person out of a group of  $n$  people,  $k$  of which are diseased, on our first pick is  $1 - k/n$ , on the second pick (assuming we picked a healthy individual)  $1 - k/(n - 1)$

$$P = \left(1 - \frac{k}{n}\right) \left(1 - \frac{k}{n-1}\right) \dots \left(1 - \frac{k}{n-m}\right)$$

where we chose a total of  $m$  times, i.e. people. Reducing:

$$\begin{aligned} P &= \left(\frac{n-k}{n}\right) \left(\frac{n-k-1}{n-1}\right) \dots \left(\frac{n-k-m}{n-m}\right) \\ &= \frac{(n-k)!(n-m)!}{n!(n-k-m)!} \end{aligned}$$

Probability that our pick contains a diseased person, i.e. the probability that a pick of  $m$  random people from  $n$  sample of people containing  $k$  diseased people is, is the complement of the above probability:

$$P = 1 - \frac{(n-k)!(n-m)!}{n!(n-k-m)!}$$

- b. c.

```
def smallestSampleSizeWithP(k, n, Pthresh=0.95):
    i, P = 0, 1
    while True:
        P *= (1 - k / (n - i))
        i += 1
        if 1 - P >= Pthresh:
            return i, 1-P

print("m = {0}; at least 1 diseased P = {1}".format(*smallestSampleSizeWithP(20, 1000)))
print("m = {0}; at least 1 diseased P = {1}".format(*smallestSampleSizeWithP(200, 10000)))

>>> m = 138; at least 1 diseased P = 0.9502553198035257
>>> m = 148;; at least 1 diseased P = 0.9508276450580865
```

- d. If we were to redo the first problem with replacement the probability of selecting a healthy individual would just be  $(1 - k/n)$  for each pick, rendering the probability of selecting only healthy people after  $m$  picks  $P = (1 - k/n)^m$ . This is implied by the hint. Putting in values from part c, and using the complement to recover the probability that we picked at least 1 diseased person:

$$\begin{aligned} P &= 1 - \left(1 - \frac{k}{n}\right)^m \\ 0.95 &= 1 - \left(1 - \frac{200}{10000}\right)^m \\ m &= 148.284 \end{aligned}$$

which is reasonably close to our numerically solved answer in c, keeping in mind that  $m$  actually has to be integer. The problem here is reducing the binomial distribution to that probability, which can be written down as

$$P = \binom{n}{k} p^k (1-p)^{n-k}$$

setting  $n = k = m$  we can rewrite:

$$P = \binom{m}{m} p^m (1-p)^{m-m} = p^m$$

and since from earlier we know that success probability of each trial pick is  $p = 1 - n/k$  we write the probability of choosing  $m$  consecutive non-deseased individuals as:

$$P = \left(1 - \frac{k}{n}\right)^m$$

the complement of which gives us the expression we started this problem from. While I do not have good motivation for setting the population to the size of the sample, while designating them all to be diseased (a given eventuality in current circumstances) there is motivation to use sampling with replacement in circumstances where our population is very large. The probability of picking the same person twice tends towards 0 as the population size tends towards infinity.

## Problem 7

Considering the following cases:

- a.  $A = \{X=1\}$
- b.  $B = \{X=2\}$
- c.  $C = \{X \leq 3\}$
- d.  $D = \{X \text{ is even}\}$
- e.  $E = \{1 < X < 5\}$
- f.  $F = C \cup D$
- g.  $G = C \cap D$

and Bernoulli, Uniform and Binomial distributions with the following expressions, in order:

$$f(x) = p^x(1-p)^{1-x} \text{ for } x = 0, 1 \text{ and } 0 \text{ elsewhere}$$

$$f(x) = \frac{1}{N} \text{ at } x_i \text{ and } 0 \text{ elsewhere}$$

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ for } x = 0, \dots, n \text{ and } 0 \text{ elsewhere}$$

we have that for specific cases of:

- a. Bernoulli with parameter  $p = 0.4$ , the following results for each case:

- (a) 0.4
- (b) 0
- (c) 1
- (d) 0.6 (when treating 0 as even)
- (e) 0
- (f) 1
- (g) 0.6

- b. Discrete uniform with parameter  $N = 9$ , the following results for each case:

- (a) 1/9
- (b) 1/9
- (c) 3/9
- (d) 4/9
- (e) 3/9
- (f) 6/9
- (g) 1/9

- c. Binomial with parameter  $n = 2$  and  $p = 0.4$ , the following results for each case:

- (a) 0.48
- (b) 0.16
- (c) 1
- (d) 0.52
- (e) 0.16
- (f) 1
- (g) 0.52