Lab 5: Independence, Screening, and Normal Distribution

Felicia Liu

09/29/2020

- Due date: Friday, October 2nd at 11:59 PM.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file (you can easily view it on Gradescope via the provided link after submitting). If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

Instructions

Part 1 of this lab focuses on calculating probability of independent events.

Section I: Breakout Problems on Independence

Blood Type

From Baldi and Moore (3E question 10.28, 4E question 10.30) All human blood can be one of the following types: O, A, B, or AB, but the distribution of blood types varies a bit among different populations of people.
Here are the distributions of blood types for a randomly chosen person in China and in the United States:

Check out the following table:

	O	*A*	*B*	*AB*
China	0.35	0.27	0.26	0.12
U.S.	0.45	0.40	0.11	0.04

Choose an American person and a Chinese person at random, independently of each other.

1. [1 point] What is the probability that both have type O blood?

P(both type O blood) = 0.35 * 0.45 = 0.1575 = 15.75%

2. [1 point] What is the probability that both have the same blood type?

P(both have same blood type) = (0.35 * 0.45) + (0.27 * 0.4) + (0.26 * 0.11) + (0.12 * 0.04) = 0.1575 + 0.108 + 0.0286 + 0.0048 = 0.2989 = 29.89%

3. [1 point] From Baldi and Moore (3E question 10.29, 4E not present)

Universal blood donors.
People with type O-negative blood are universal donors.
That is, any patient can receive a transfusion of type O-negative blood.
Only 7.2 % of the American population have O-negative blood.
If 10 people appear at random to give blood, what is the probability that at least 1 of them is a universal donor?

P(at least 1 is a universal donor) = 1 - P(no one is a universal donor) = 1 - (1 - 0.072)ˆ10 = 1 - 0.928ˆ10 = 1 - 0.4736742 = 0.5263 = 52.63%

From Baldi and Moore (3E question 10.46, 4E not present)

Mendelian inheritance.

Some traits of plants and animals depend on inheritance of a single gene.
This is called *Mendelian inheritance*, after Gregor Mendel (1822-1884). Each of us has an ABO blood type, which describes whether two characteristics called A and B are present.
Every human being has two blood type alleles (gene forms), one inherited from our mother and one from our father.
Each of these alleles can be A, B, or O. Which two we inherit determines our blood type.
Here is a table that shows what our blood type is for each combination of two alleles.

Alleles inherited	Blood type
A and A	A
A and B	AB
A and O	A
B and B	B
B and O	B
O and O	O

We inherit each of a parent's two alleles with probability .50, and we inherit independently from our mother and our father.

Punnett squares are used in genetics courses to organize this type of information. The alleles for one parent label the rows and for the other parent label the columns.

4. [1 point] Rachel and Jonathan both have alleles A and B. What blood types can their children have?

Their children can have blood types: AA, AB, and BB.

5. [1 point] Jasmine has alleles A and O. Tyrone has alleles B and O. What is the probability that a child of these parents has blood type O?

P(child has blood type O) $= 0.25 = 25\%$

6. [1 point] If Jasmine and Tyrone have three children, what is the probability that all three have blood type O? What is the probability that the first child has blood type O and the next two do not?

The next calculations assume that they do not have any twins or triplets, so that each child is independent.

what is the probability that all three have blood type O?

P(all three have blood type O) = 0.25^3 = 0.015625 = 1.56%

what is the probability that the first child has blood type O and the next two do not?

P(first is type O, next 2 not O) = 0.25 * 0.75 * 0.75 = 0.140625 = 14.06%

The Flu

7. [2 points] From Baldi and Moore (2E question 10.20 to 10.23)

The November 2009 Gallup-Healthways Well-Being Index survey asked a random sample of 28,606 American adults whether they had the flu on the day before the interview. Here are the results by age group:

age group	flu	no flu	total
18 to 29	88	2,486	2,574
30 to 44	132	5,162	5,294
45 to 64	276	11,733	12,009
65+	122	8,607	8,729
total	618	27,988	28,606

The events “adult is in a specified age group” and “adult has the flu” are called **independent** if the probability of flu does not vary across the age groups.

The conditional probabilities of flu among those age 65 and older is 122/8,729 and flu among those 18 to 29 years old is 88/2,574.

Calculate and compare the probability of flu among those age 65 and older and the probability of flu among those 18 to 29 years old.

The probability of flu among those 65+ is 122/8,729 = 1.40% and the probability of flu among those 18 to 29 years old is 88/2,574 = 3.42%. The probability of flu among those 65+ is lower than that of those 18 to 29 years old.

This shows that the events “age group” and “adult has the flu” are **not** independent. Another way to check for independence is to see if the overall probability of flu differs from the age-group stratum specific probabilities of flu. Do this, too.

The overall probability of flu is 618/28,606 = 2.16%, which is different from the age-group stratum specific probabilities of the flu.

Testing for HIV

From Baldi and Moore (question 10.14)

Enzyme immunoassay tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. Antibodies indicate the presence of the virus. The test is quite accurate but is not always correct. Here are approximate probabilities of positive and negative test results when the blood tested does and does not actually contain antibodies to HIV. [Hint: these are conditional probabilities, given HIV status.]

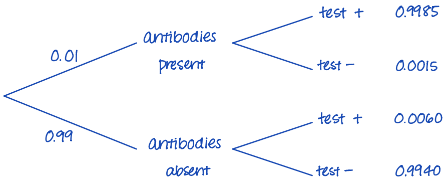
- P(test positive + | antibodies present) = 0.9985
- P(test positive + | antibodies absent) = 0.0060
- P(test negative – | antibodies present) = 0.0015
- P(test negative – | antibodies absent) = 0.9940

Suppose that 1% of a large population carries antibodies to HIV in its blood.

8. [1 point] Draw a tree diagram representing the HIV status of a person from this population (outcomes: antibodies present or absent) and the blood test result (outcomes: test positive or test negative).

(Use the code chunk below to include an image file of your drawing. To do so you need to delete the hashtag, upload the image to Datahub into the `src` directory and replace the file name with your file name. JPG or PNG will both work.)

```r
#Take off the '#' in the following code and replace the code with you file name
knitr::include_graphics("src/Lab05 Tree Diagram.pdf")
```



```
#Your code for range here
```

[TODO: YOUR ANSWER HERE]

9. [1 point] What is the probability that the test is positive for a randomly chosen person from this population?

If the target population is 1000, 10 will truly have antibodies and 990 will not. Of the 10 who will have antibodies, 9.985 will test positive (true positive) and of the 990, 5.94 will test positive (false positive). (9.985 + 5.94)/1000 = 0.015925 = 1.59%. So P(test is positive) = 1.59%.

From Baldi and Moore (10.16)

Continue your work and probability assumptions from Question 6.

10. [1 point] What is the probability that a person has the antibody, given that the test is positive? Explain in your own words what this means.

P(antibodies present|test positive +) = P(antibodies present & test positive)/P(test positive) = P(antibodies present & test positive)/[P(test positive & antibodies present) + P(test positive & antibodies absent)] = 0.009985/(0.009985+0.00594) = 0.009985/0.015925 = 0.62700 = 62.70%. This probability is the chance that someone actually has the antibody, given that their test was positive. It is the chance of a true positive out of all positive tests.

11. [1 point] Identify the test's sensitivity, specificity, and positive predictive value.

The sensitivity is P(test positive | antibodies present) = 99.85%. The specificity is P(test negative | antibodies absent) = 99.40%. The positive predicted value is P(antibodies present | test positive) = 62.70%.

According to the CDC's Behavioral Risk Factor Surveillance System (BRFSS) about 60% of American adults live a sedentary lifestyle.

Noting that random sampling gives us independent observations, you randomly select 10 adults from this population. Find the following probabilities:

12. [1 point] All 10 have a sedentary lifestyle

P(all 10 have a sedentary lifestyle) = 0.6ˆ10 = 0.60%

13. [1 point] At least one does not have a sedentary lifestyle

P(at least 1 does not have a sedentary lifestyle) = 1 - P(no one has sedentary lifestyle) = 1 - 0.4ˆ10 = 99.99%

Section 2: Normal Distribution

Part 2 of this lab introduces the normal distribution.

Notation reminder: Baldi and Moore use a compact notation for specifying that a population has a distribution that follows a normal curve with mean μ and standard deviation σ : N(μ , σ)

R Functions Reminder: `qnorm` takes a probability as its input and gives back a value on the distribution (aka a z-score if the distribution is N(0,1)). By default, it assumes the probability area you enter (as a decimal, not a percent) is the area below (or less than) the z score you need. The `pnorm` function takes a z value or an X value as its input and gives back a probability area.

Question 14. Z scores. (From Baldi and Moore, 3E question 11.27, 4E question 11.29) Use R to find the standardized value z that satisfies each of the following conditions. In each case, sketch a standard Normal curve with your value of z marked on the axis. You do not have to attach your diagrams.

14. [1 point] The probability is 0.8 that a randomly selected observation falls below z.

```
p14 <- qnorm(p = 0.8, mean = 0, sd = 1)
p14
```

```
## [1] 0.8416212
```

```
check_problem14()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You computed the correct z-score!"
##
## Problem 14
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

15. [1 point] The probability is 0.35 that a randomly selected observation falls above z.

```r
p15 <- qnorm(p = 0.65, mean = 0, sd = 1)
p15
```

```
## [1] 0.3853205
```

```r
check_problem15()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric!"
## [1] "Checkpoint 2 Passed: You computed the correct z-score!"
##
## Problem 15
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Full-Term Birth Weights (From Baldi and Moore, 3E question 11.31, 4E question 11.33)

For babies born at full term (37 to 39 completed weeks of gestation), the distribution of birth weight (in grams) is approximately normally distributed with a mean of 3350 grams and a standard deviation of 440 grams, N(3350,440).

16. [1 point] What is the 25th percentile of the birthweights for full term babies?

```r
p16 <- qnorm(p = 0.25, mean = 3350, sd = 440)
p16
```

```
## [1] 3053.225
```

```r
check_problem16()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct percentile!"
##
## Problem 16
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

17. [1 point] What is the 90th percentile of the birthweights for full term babies?

```r
p17 <- qnorm(p = 0.9, mean = 3350, sd = 440)
p17
```

```
## [1] 3913.883
```

```r
check_problem17()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct percentile!"
##
## Problem 17
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

18. [1 point] What is the range of the middle 50% of birthweights for full term babies?

```r
p18 <- qnorm(p = 0.75, mean = 3350, sd = 440) - qnorm(p = 0.25, mean = 3350, sd = 440)
p18
```

```
## [1] 593.551
```

```r
check_problem18()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct middle range of birthweights!"
##
## Problem 18
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Late Pre-Term Birth Weights (From Baldi and Moore, 3E question 11.32, 4E question 11.34)

How much of a difference do a couple of weeks make for birthweight? Late preterm babies are born with 35 to 37 weeks of completed gestation. The distribution of birth weight (in grams) or late preterm babies is approximately normally distributed with a mean of 2750 grams and a standard deviation of 560 grams, N(2750,560).

19. [1 point] What is the 25th percentile of the birthweights for late-preterm term babies?

```r
p19 <- qnorm(p = 0.25, mean = 2750, sd = 560)
p19
```

```
## [1] 2372.286
```

```r
check_problem19()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct percentile!"
##
## Problem 19
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

20. [1 point] What is the 90th percentile of the birthweights for late-preterm babies?

```r
p20 <- qnorm(p = 0.9, mean = 2750, sd = 560)
p20
```

```
## [1] 3467.669
```

```r
check_problem20()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct percentile!"
##
## Problem 20
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

21. [1 point] What is the range of the middle 50% of birthweights for late-preterm babies?

```
p21 <- qnorm(p = 0.75, mean = 2750, sd = 560) - qnorm(p = 0.25, mean = 2750, sd = 560)
p21
```

```
## [1] 755.4285
```

```
check_problem21()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct middle range of birthweights!"
##
## Problem 21
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

22. [1 point] Compare your answers to the parts of full term babies to late-preterm babies. What do you notice?

The 25th and 90th percentiles for the full term babies were higher than those of the late-preterm babies. The range of the middle 50% of full term birth weights was smaller than that of the late-preterm babies.

Question 5. Newborn Respirations. (From Baldi and Moore, 3E questions 11.15-11.17, 4E question 11.17-11.19)

23. [1 point] The respiratory rate per minute in newborns varies according to a distribution that is approximately Normal with mean 50 and standard deviation 5. The probability (convert to a percentage and round to two decimal places) that a randomly chosen newborn has a respiratory rate of 55 per minute or more is approximately:

```
p23 <- (1 - pnorm(q = 55, mean = 50, sd = 5)) * 100
p23
```

```
## [1] 15.86553
```

```
check_problem23()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
## [1] "Checkpoint 3 Passed: You calculated the correct percentage and rounded to 2 decimal places!"
##
## Problem 23
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

24. [1 point] The probability (convert to a percentage and round to two decimal places) that a randomly chosen newborn has a respiratory rate per minute between 40 and 55 is approximately:

```
p24 <- (pnorm(q = 55, mean = 50, sd = 5) - pnorm(q = 40, mean = 50, sd = 5)) * 100
p24
```

```
## [1] 81.85946
```

```r
check_problem24()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
## [1] "Checkpoint 3 Passed: You calculated the correct percentage and rounded to 2 decimal places!"
##
## Problem 24
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Drosophila (From Baldi and Moore, 3E questions 11.20 and 11.22, 4E question 11.23)

The common fruit fly, Drosophila melanogaster, is the most studied organism in genetic research because it is small, easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal with mean 0.800 millimeters (mm) and standard deviation 0.078 mm.

25. [**1 point**] Choose a male fruit fly at random. The probability (convert to a percentage and round to two decimal places) that the fly you choose has a thorax longer than 1 mm is about:

```r
p25 <- (1 - pnorm(q = 1, mean = 0.800, sd = 0.078)) * 100
p25
```

```
## [1] 0.5172149
```

```r
check_problem25()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
## [1] "Checkpoint 3 Passed: You calculated the correct percentage and rounded to 2 decimal places!"
##
## Problem 25
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Z scores. (From Baldi and Moore, 3E question 11.25, 4E question 11.27)

Use R's pnorm function to find the proportion of observations from a standard Normal distribution that fall in each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.You do not have to attach your diagrams.

26. [**1 point**] z $\leq$ -2.25

```r
p26 <- pnorm(q = -2.25, mean = 0, sd = 1)
p26
```

```
## [1] 0.01222447
```

```r
check_problem26()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
##
```

```
## Problem 26
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

27. [**1 point**] $z \geq -2.25$

```
p27 <- 1 - pnorm(q = -2.25, mean = 0, sd = 1)
p27
```

```
## [1] 0.9877755
```

```
check_problem27()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calcuated the correct probability!"
##
## Problem 27
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

28. [**1 point**] $z > 1.77$

```
p28 <- 1 - pnorm(q = 1.77, mean = 0, sd = 1)
p28
```

```
## [1] 0.03836357
```

```
check_problem28()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
##
## Problem 28
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

29. [**1 point**] $-2.25 < z < 1.77$

```
p29 <- pnorm(q = 1.77, mean = 0, sd = 1) - pnorm(q = -2.25, mean = 0, sd = 1)
p29
```

```
## [1] 0.949412
```

```
check_problem29()
```

```
## [1] "Checkpoint 1 Passed: You calculated a numeric value!"
## [1] "Checkpoint 2 Passed: You calculated the correct probability!"
##
## Problem 29
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##              Test Points_Possible       Type
## Problem 1  FAILED               0 autograded
## Problem 2  FAILED               0 autograded
## Problem 3  FAILED               0 autograded
## Problem 4  FAILED               0 autograded
## Problem 5  FAILED               0 autograded
## Problem 6  FAILED               0 autograded
## Problem 7  FAILED               0 autograded
## Problem 8  FAILED               0 autograded
## Problem 9  FAILED               0 autograded
## Problem 10 FAILED               0 autograded
## Problem 11 FAILED               0 autograded
## Problem 12 FAILED               0 autograded
## Problem 13 FAILED               0 autograded
## Problem 14 PASSED               1 autograded
## Problem 15 PASSED               1 autograded
## Problem 16 PASSED               1 autograded
## Problem 17 PASSED               1 autograded
## Problem 18 PASSED               1 autograded
## Problem 19 PASSED               1 autograded
## Problem 20 PASSED               1 autograded
## Problem 21 PASSED               1 autograded
## Problem 22 FAILED               0 autograded
## Problem 23 PASSED               1 autograded
## Problem 24 PASSED               1 autograded
## Problem 25 PASSED               1 autograded
## Problem 26 PASSED               1 autograded
## Problem 27 PASSED               1 autograded
## Problem 28 PASSED               1 autograded
## Problem 29 PASSED               1 autograded
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-fa20/lab/lab05; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.