Homework 2 Answers

EPI 525

These are the **some** numeric/short answers to the homework. Often, these answers are insufficient for your own work or solutions. I just wanted to give you a part of the answer to help guide you in the right direction.

Book exercises

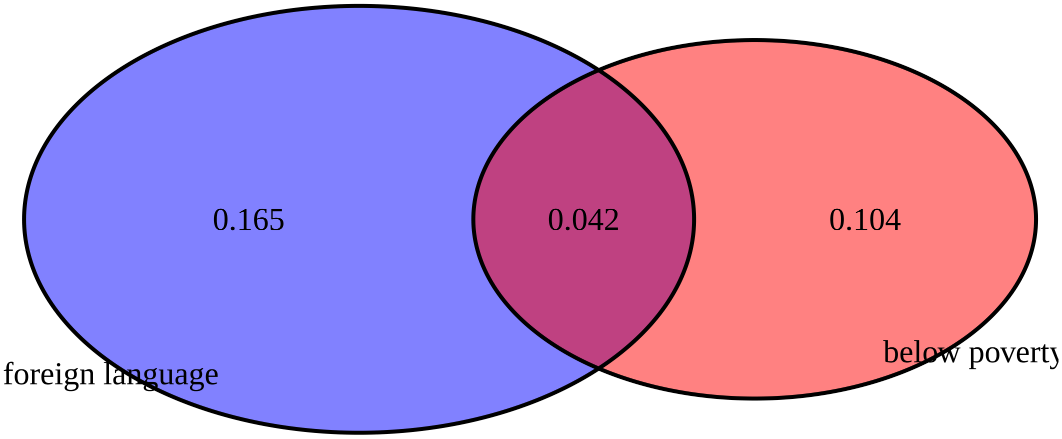
2.6 Poverty and language

The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.

a

Not disjoint

b



```
(polygon[GRID.polygon.1], polygon[GRID.polygon.2], polygon[GRID.polygon.3], polygon[GRID.poly
```

c

10.4%

d

31.1%

e

68.9%

f

Not independent

2.8 School absences

a

0.32

b

0.57

c

0.68

d

0.1024

e

0.4624

2.10 Health coverage, frequencies

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends.

The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.

```r
#table. not required
(covg.tab <- data.frame(
  health_status = c("No","Yes","total"),
  Excellent = c(459,4198,4657),
  Very_good = c(727,6245,6972),
  Good = c(854,4821,5675),
  Fair = c(385,1634,2019),
  Poor = c(99,578,677),
  Total = c(2524,17476,20000)
)) %>% gt()
```

health_status	Excellent	Very_good	Good	Fair	Poor	Total
No	459	727	854	385	99	2524
Yes	4198	6245	4821	1634	578	17476
total	4657	6972	5675	2019	677	20000

a

If one individual is drawn at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

```
(eh.no_covg <- (459/20000))
```

```
[1] 0.02295
```

The probability the respondent drawn at random has excellent health and doesn't have health coverage is 0.02295.

b

If one individual is drawn at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

Let A = excellent health and B = doesn't have health coverage.

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

```
a <- 4657/20000; b <- 2524/20000; c <- 459/20000
(eh_or_no.hc <- a+b-c)
```

```
[1] 0.3361
```

The probability that the respondent drawn at random has excellent health OR doesn't have health coverage is 0.3361.

2.14 Health coverage, relative frequencies

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) conditional on whether or not they have health insurance.

```r
(covg.tab <- data.frame(
  health_status = c("No","Yes","total"),
  Excellent = c(0.0230, 0.2099,0.2329),
  Very_good = c(0.0364, 0.3123,0.3486),
  Good = c(0.0427,0.2410,0.2838),
  Fair = c(0.0192,0.0817,0.1009),
  Poor = c(0.0050,0.0289,0.0338),
  Total = c(0.1262,0.8738,1.0000)
)) %>% gt()
```

health_status	Excellent	Very_good	Good	Fair	Poor	Total
No	0.0230	0.0364	0.0427	0.0192	0.0050	0.1262
Yes	0.2099	0.3123	0.2410	0.0817	0.0289	0.8738
total	0.2329	0.3486	0.2838	0.1009	0.0338	1.0000

a

Are being in excellent health and having health coverage mutually exclusive?
::: blue No. 0.2099 is the proportion of respondents that have health coverage and excellent health, which is not 0. :::

b

What is the probability that a randomly chosen individual has excellent health?
::: blue The probability that a randomly chosen individual has excellent health is 0.2329. This value is shown in the table. :::

c

What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

$$P(A|B) = \frac{P(A \; and \; B)}{P(B)}$$

```r
eh <- 0.2329; hc <- 0.8738; eh_and_hc <- 0.2099
(eh_give_hc <- eh_and_hc/hc)
```

```
[1] 0.2402152
```

The probability that a randomly chosen individual has excellent health given that he has health coverage is 0.2402152.

d

What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

```r
no_hc <- 0.1262; eh_and_no_hc <- 0.0230
(eh_give_no_hc <- eh_and_no_hc/no_hc)
```

```
[1] 0.1822504
```

The probability that a randomly chosen individual has excellent health given that he doesn't have health coverage is 0.1822504.

e

Do having excellent health and having health coverage appear to be independent?
::: blue No, because the probability that a person has excellent health varies between the two health coverage categories (24% vs 18%). That is, knowing something about someone's health coverage provides useful information in predicting whether the person has excellent health, which means the variables are not independent. :::

2.18 Predisposition for thrombosis

A genetic test is used to determine if people have a predisposition for thrombosis, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition.

a

What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

Let the events below be defined as follows:

- D = has disease (predisposition for thrombosis),

- D^c = absence of disease (no predisposition for thrombosis),
- T^+ = test positive for predisposition for thrombosis, and
- T^- = test negative for predisposition for thrombosis.

We are given the probabilities below, and can calculate the probabilities of their complements:

$$\begin{aligned} P(D) &= 0.03,\ P(D^c) = 0.97 \tag{1}\\ P(T^+|D) &= 0.99,\ P(T^-|D) = 1 - P(T^+|D) = 0.01 \tag{2}\\ P(T^-|D^c) &= 0.98,\ P(T^+|D^c) = 1 - P(T^+|D) = 0.02 \tag{3}\end{aligned}$$

Then using Bayes' Rule, we have

$$\begin{aligned} P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)} &= \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|D^c)P(D^c)} \tag{4}\\ &= \frac{(0.99)(0.03)}{(0.99)(0.03) + (0.02)(0.97)} \tag{5}\end{aligned}$$

```
#set each variable:
dis <- 0.03 #prob of having disease (predisposition)
dis.c <- (1-dis) #prob of not having predisposition
```

```
pos.giv.d <- 0.99 #prob of positive test given disease
neg.giv.dc <- 0.98 #prob of negative test given not having disease
(pos.giv.dc <- 1-neg.giv.dc) #prob of positive test given not having disease
```

```
[1] 0.02
```

```
(neg.giv.d <- 1-pos.giv.d)  #prob of negative test given having disease
```

```
[1] 0.01
```

```
#find the probability of having the disease given a positive test
(dis.giv.pos <- (pos.giv.d*dis)/((dis*pos.giv.d)+(dis.c*pos.giv.dc)))
```

```
[1] 0.604888
```

The probability that a randomly selected person who tests positive for thrombosis predisposition by the test actually has a predisposition for thrombosis is 0.605.

b

What is the probability that a randomly selected person who tests negative for the predisposition by the test actually does not have the predisposition?

Using Bayes' Rule, we have

$$P(D^c|T^-) = \frac{P(D^c\ and\ T^-)}{P(T^-)} = \frac{P(T^-|D^c)P(D^c)}{P(T^-|D)P(D) + P(T^-|D^c)P(D^c)} \tag{6}$$

$$= \frac{(0.98)(0.97)}{(0.01)(0.03) + (0.98)(0.97)} \tag{7}$$

```
#find the probability of not having the disease given testing negative for the predisposit
(nodis.giv.neg <- (neg.giv.dc*dis.c) / (dis*neg.giv.d + dis.c*neg.giv.dc))
```

```
[1] 0.9996845
```

The probability that a randomly selected person who tests negative for thrombosis predisposition by the test actually does not have thrombosis predisposition is 0.9997.

2.24 Breast cancer and age

The strongest risk factor for breast cancer is age; as a woman gets older, her risk of developing breast cancer increases. The following table shows the average percentage of American women in each age group who develop breast cancer, according to statistics from the National Cancer Institute. For example, approximately 3.56% of women in their 60's get breast cancer. A mammogram typically identifies a breast cancer about 85% of the time, and is correct 95% of the time when a woman does not have breast cancer.

a

Calculate the PPV for each age group. Describe any trend(s) you see in the PPV values as prevalence changes. Explain the reason for the trend(s) in language that someone who has not taken a statistics course would understand

Let the events below be defined as follows:

- D = has breast cancer,
- D^c = absence of breast cancer,
- T^+ = test positive for breast cancer, and
- T^- = test negative for breast cancer.

Let P(D30) = prevalence of breast cancer at ages 30-40, P(D40) at ages 40-50, and so on.

We are given the probabilities below, and can calculate the probabilities of their complements:

$$sensitivity: \; P(T^+|D) = 0.85, \; P(T^-|D) = 1 - P(T^+|D) = 0.15 \tag{8}$$

$$specificity: \; P(T^-|D^c) = 0.95, \; P(T^+|D^c) = 1 - P(T^+|D) = 0.05 \tag{9}$$

Then using Bayes' Rule, we have

$$PPV = P(D|T^+) = \frac{P(D \; and \; T^+)}{P(T^+)} = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|D^c)P(D^c)}$$

Define variables in R:

```
#define variables
d30 <- 0.0044 #30-40
d40 <- 0.0147 #40-50
d50 <- 0.0238 #50-60
d60 <- 0.0356 #60-70
```

```
d70 <- 0.0382 #70-80
pos.g.d <- 0.85 #prob of positive test given disease
neg.g.dc <- 0.95 #prob of negative test given not having disease
(pos.g.dc <- 1-neg.g.dc) #prob of positive test given not having disease.
```

```
[1] 0.05
```

```
(neg.g.d <- 1-pos.g.d)
```

```
[1] 0.15
```

PPV calculations in R for different age groups:

```
# Below we take advantage of R's capability of doing vector-wise calculations

# Create a vector of the prevalences
prevalence <- c(d30, d40, d50, d60, d70)

# Compute the PPV's using the vector of prevalences
(PPV <- round((pos.g.d*prevalence) / (prevalence*pos.g.d + (1-prevalence)*pos.g.dc), digit
```

```
[1] 0.070 0.202 0.293 0.386 0.403
```

The resulting PPV's are the following:

```
age_group <- c("30 - 40", "40 - 50", "50 - 60", "60 - 70", "70 - 80")

data.frame(age_group, prevalence, PPV) %>%
  gt()
```

age_group	prevalence	PPV
30 - 40	0.0044	0.070
40 - 50	0.0147	0.202
50 - 60	0.0238	0.293
60 - 70	0.0356	0.386
70 - 80	0.0382	0.403

From the table we see that as the prevalence of breast cancer increases in the older age groups, the positive predictive value (PPV) also increases. Thus, if more women in a population have breast cancer, the probability that a positive test correctly identifies breast cancer increases.

Mathematically this can be explained by looking at the Bayes' Rule equation. As the prevalence increases, the term (prev x sensitivity) will increase in both the numerator and denominator. However, (1-prev) will decrease in the denominator. Therefore, overall, as long as the sensitivity and specificity are held constant, as the prevalence increases the number of true positives increases and the number of false positives decreases. PPV increases when the number of true positives increases.

b

Suppose that two new mammogram imaging technologies have been developed which can improve the PPV associated with mammograms; one improves sensitivity to 99% (but specificity remains at 95%), while the other improves specificity to 99% (while sensitivity remains at 85%). Which technology offers a higher increase in PPV? Explain why.

The technology that raises the specificity to 99% offers a higher increase. Since prevalence of the disease only ranges between less than 1% to at most 4%, the sensitivity is being weighted (multiplied by) a small amount $(P(T^+|D)P(D))$ and thus changes in the sensitivity do not change the overall PPV very much.

In contrast, the term involving the specificity is being weighted (multiplied by) a lot when the prevalence is low, and thus changes in the specificity have a greater effect on the PPV. Furthermore, as the specificity increases, the likelihood of false positive goes down. Thus, with the relatively large decrease in the denominator, the PPV will increase more than compared to when there is a very small increase in both numerator and denominator (increase in sensitivity).

Intuitively, since prevalence of the disease is very small, there will be very few people with a true positive result, and comparatively many more people with a false positive true. While increasing the sensitivity of the test will increase the number of people with a true positive result, the number cannot increase much since so few women have breast cancer to begin with. In contrast, there are many more women without breast cancer, and so even a small increase in the specificity can lead to a relatively larger change (decrease) in the number of false positives. This decrease in the denominator will increase in the PPV.

Calculations of new PPV with new imagaing technologies:

```
# improves sensitivity to 99% (but specificity remains at 95%)

pos.g.d <- 0.99 #prob of positive test given disease
neg.g.dc <- 0.95 #prob of negative test given not having disease
```

```
(pos.g.dc <- 1-neg.g.dc) #prob of positive test given not having disease.
```

```
[1] 0.05
```

```
(neg.g.d <- 1-pos.g.d)
```

```
[1] 0.01
```

```
(PPV_newSens99 <- round((pos.g.d*prevalence) / (prevalence*pos.g.d + (1-prevalence)*pos.g.
```

```
[1] 0.080 0.228 0.326 0.422 0.440
```

```
# improves specificity to 99% (while sensitivity remains at 85%)

pos.g.d <- 0.85 #prob of positive test given disease
neg.g.dc <- 0.99 #prob of negative test given not having disease
(pos.g.dc <- 1-neg.g.dc) #prob of positive test given not having disease.
```

```
[1] 0.01
```

```
(neg.g.d <- 1-pos.g.d)
```

```
[1] 0.15
```

```
(PPV_newSpec99 <- round((pos.g.d*prevalence) / (prevalence*pos.g.d + (1-prevalence)*pos.g.
```

```
[1] 0.273 0.559 0.675 0.758 0.771
```

The table below compares the results, where we see that indeed the PPV increased more with the higher specificity vs. the higher sensitivity.

```
data.frame(age_group, prevalence, PPV, PPV_newSens99, PPV_newSpec99) %>%
  gt()
```

age_group	prevalence	PPV	PPV_newSens99	PPV_newSpec99
30 - 40	0.0044	0.070	0.080	0.273
40 - 50	0.0147	0.202	0.228	0.559
50 - 60	0.0238	0.293	0.326	0.675
60 - 70	0.0356	0.386	0.422	0.758
70 - 80	0.0382	0.403	0.440	0.771

Non-book exercise

Suppose a patient has abdominal pain and their clinician suspects that they either have disease 1, disease 2, or no disease, where the probability of having abdominal pain if they have disease 1 is 0.80, the probability of having abdominal pain if they have disease 2 is 0.90, and the probability of having abdominal pain if they have no disease is 0.01. Based on the patient's medical history, the probability of having disease 1 is 0.009, having disease 2 is 0.001, and having no disease is 0.99. What is the probability the patient has disease 2 given that they have abdominal pain?

Let A = event of having abdominal pain, D_1 = event has disease 1, D_2 = event has disease 2, and D_0 = event has no disease. The sample space is $S = \{D_1, D_2, D_0\}$.

$$P(A|D_1) = 0.80 \qquad P(D_1) = 0.009 \tag{10}$$

$$P(A|D_2) = 0.90 \qquad P(D_2) = 0.001 \tag{11}$$

$$P(A|D_0) = 0.01 \qquad P(D_0) = 0.99 \tag{12}$$

Thus, applying Bayes' Rule, we have

$$P(D_2|A) = \frac{P(A|D_2)P(D_2)}{P(A|D_1)P(D_1) + P(A|D_2)P(D_2) + P(A|D_0)P(D_0)} \tag{13}$$

$$= \frac{(0.90)(0.001)}{(0.80)(0.009) + (0.90)(0.001) + (0.01)(0.99)} \tag{14}$$

Using R to calculate the probability:

```
#name variables:
a.g.d1 <- 0.80 # a given d1
a.g.d2 <- 0.90 # a given d2
a.g.d0 <- 0.01 # a given d0
```

```
d1 <- 0.009
d2 <- 0.001
d0 <- 0.99
```

```
#find the probability the patient has disease 2 given the have abdominal pain:
(PDA <- (a.g.d2*d2) / (a.g.d2*d2 + a.g.d1*d1 + a.g.d0*d0))
```

```
[1] 0.05
```

The probability the patient has disease 2 given they have abdominal pain is 0.05.