Lab 4: Screening and Precision Public Health

Screening

We will talk about a very important use of conditional probability in public health and medicine, which is the idea of tools that screen for health outcomes. There are many examples of this, including mammograms for detection of breast cancer, the prostate specific antigen (PSA) for detection of prostate cancer, as well as tests for exposure to infectious diseases, and so forth. We will consider two types of events: i) whether the subject truly has the health condition of interest (let D denote the disease of interest, and D' its complement), and ii I) whether a test was positive or not for this outcome (let $+$ denote a positive test and $-$ denote its complement, a negative test). There are several statistics that are used to evaluate the performance of a test, some of which are derivable from each other:

- Sensitivity: $P(+ \mid D)$ or the probability of test being positive if one has the disease.
- Specificity: $P(- \mid D')$ or the probability of test being negative given one does not have the disease.
- Positive predictive value (PPV): $P(D \mid +)$ or the probability of having the disease if an individual tests positive.
- Negative predictive value (NPV): $P(D' \mid -)$ or the probability of not having the disease if an individual tests negative.

Consider the following situation: Assume the total number of subjects is 10,000 and that, $P(D) = 0.05$, $P(+ \mid D) = 0.95$, $P(- \mid D') = 0.95$. This set up implies that the disease is rare, but that a very accurate test exists (i.e., equally high sensitivity and specificity).

1. Fill in the following two-way table with the absolute frequencies using the information provided in the problem. (You can fill it in simply by typing the frequencies in the nine empty cells.):

	D	D'	Total
+			
-			
Total			

2. Calculate the PPV using the table.
3. Re-do the two-way table and re-calculate the PPV, this time assuming that $P(D) = 0.02$. (Note that $P(+ \mid D) = 0.95$, $P(- \mid D') = 0.95$, as with the previous question.)
4. Explain why the sensitivity is so high, but the PPV is low for the first calculation and even lower for the second calculation.

Precision public health

One of the goals of public health research is to group people by risk factors or demographic variables so that decision-makers can predict, with actionable accuracy, which groups are at high and low risk of an adverse health outcome. In this set of questions, we consider stratified two-way tables, which are two-way tables specific to levels of a third grouping variable. Here, the adverse health outcome is coronary heart disease (CHD), which we represent by D. We study two categorical risk factors, smoking (defined by S for smoking and S' for no smoking) and age (defined by A for older age and A' for younger age). First,read in the aggregated data set. The last column (n) is the number of individuals in each group.

```r
library(dplyr)
library(tidyverse)
chd_dat<- read_csv("Data/Lab5_CHD.csv")

chd_dat
```

```
## # A tibble: 8 x 4
##   Age   Smoking CHD       n
##   <chr> <chr>   <chr> <int>
## 1 young yes     yes      60
## 2 young yes     no      240
## 3 young no      yes     105
## 4 young no      no      595
## 5 old   yes     yes     180
## 6 old   yes     no      120
## 7 old   no      yes     210
## 8 old   no      no      490
```

From this table:

4. Calculate the following probabilities. Convert your answer to percentages.:

- $P(D \mid A', S)$
- $P(D \mid A', S')$
- $P(D \mid A, S)$
- $P(D \mid A, S')$

If you prefer, you can do these calculations by hand based on `chd_dat`. Some students might wish to use R commands to calculate these probabilities. There are **many** ways to do this. You could use `dplyr` functions to perform the calculations. Alternatively, here are some new functions for those of you interested in learning more R. (Note that these new functions won't be tested, they are for your information only, and to expose you to more of the R language!). You could consider using the `uncount()` function to expand the data based upon the numbers in each group (i.e., `n`) and assign the expanded data frame to a new name.Then, you can use the `tabyl` function from the `janitor` package to create stratified two-way tables, and the relevant `adorn_` functions from `janitor` to convert the frequencies to percentages.

```r
# write code here if you wish to use R to calculate the answer (optional)
```

5. What do these numbers imply about smoking and the risk of CHD?

6. Calculate the marginal probability of CHD. This can be written as $P(D)$.

```r
# write code here if you wish to use R to calculate the answer (optional)
```

7. Calculate the conditional probabilities The $P(D \mid A')$ and $P(D \mid A)$

```r
# write code here if you wish to use R to calculate the answer (optional)
```

8. If you had an intervention that could eliminate the risk of smoking on CHD, which group (defined by age) would see the biggest change in CHD from this intervention?