Kyle Connall
Fady Youssef
Math 321 Lab2
Gonzaga University
12/10/2025

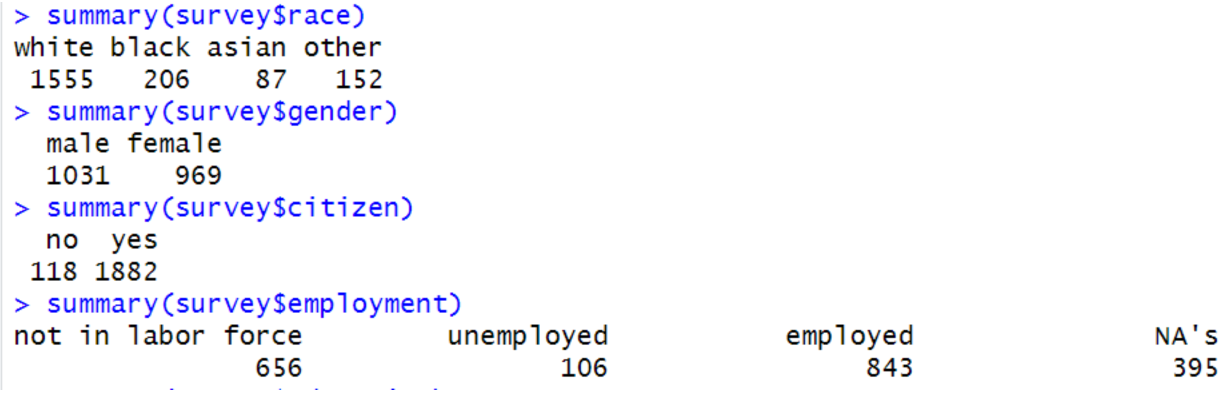
The data

We will use data from the 2012 US Census American Community Survey. Given that 2012 is not a multiple of 10, we know this is not the full census, like the ones conducted in 2010 and 2020. Instead, this is a sample of data, from 2000 people. There are 13 columns of data and we will focus on categorical data, including:

- “employment”, which gives employment status as “not in labor force”, “unemployed”, “employed”, or NA.
- “race”
- “gender”
- “edu” – education level, categorized as high school or lower, college, graduate, or NA.
- “disability”

Exercise 1: Using the summary command, create tables of summary information for race, gender, citizen, employment, and education level. (You will create 4 tables.)
Tables of data are great for exact values and we will need those throughout the lab. A nice bar graph can also get the big ideas across well. Try the command:

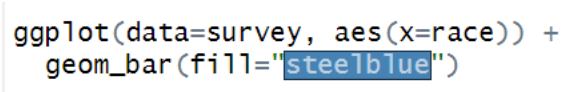
```
ggplot(data=survey, aes(x=employment)) +
        geom_bar(fill="steelblue")
```



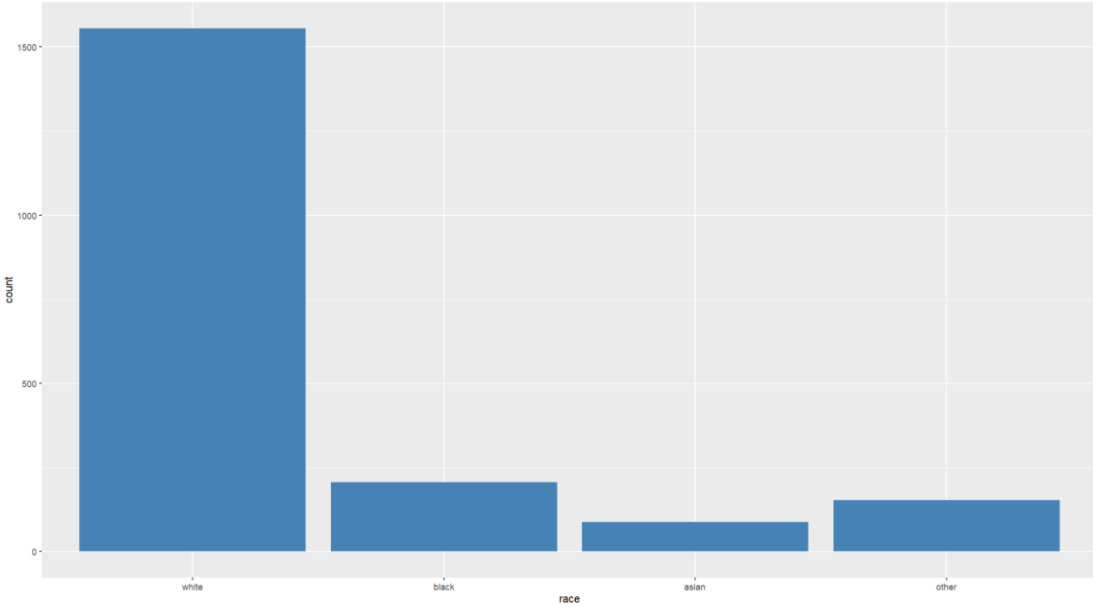
```
> summary(survey$race)
white black asian other 
 1555   206    87   152 
> summary(survey$gender)
  male female 
  1031    969 
> summary(survey$citizen)
  no  yes 
 118 1882 
> summary(survey$employment)
not in labor force         unemployed           employed               NA's 
               656                106                843                395 
```

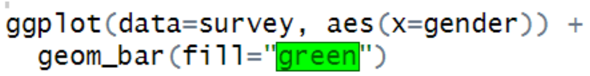
Exercise 2: Using ggplot(), create bar graphs to summarize the same data from Exercise 1: race, gender, citizen, employment, and education level. For each graph, *choose a different color*.

Note: If you avoid copy/paste and retype the command for ggplot() each time, it will help you to commit it to memory.

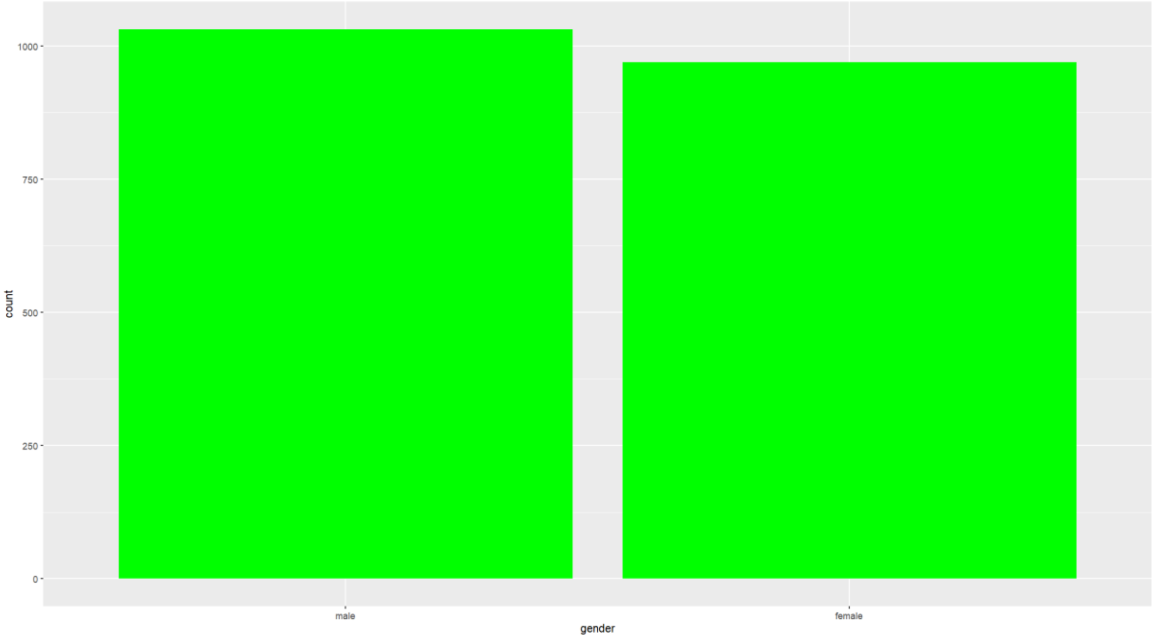


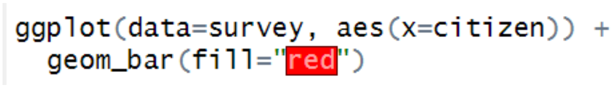
```
ggplot(data=survey, aes(x=race)) +
  geom_bar(fill="steelblue")
```



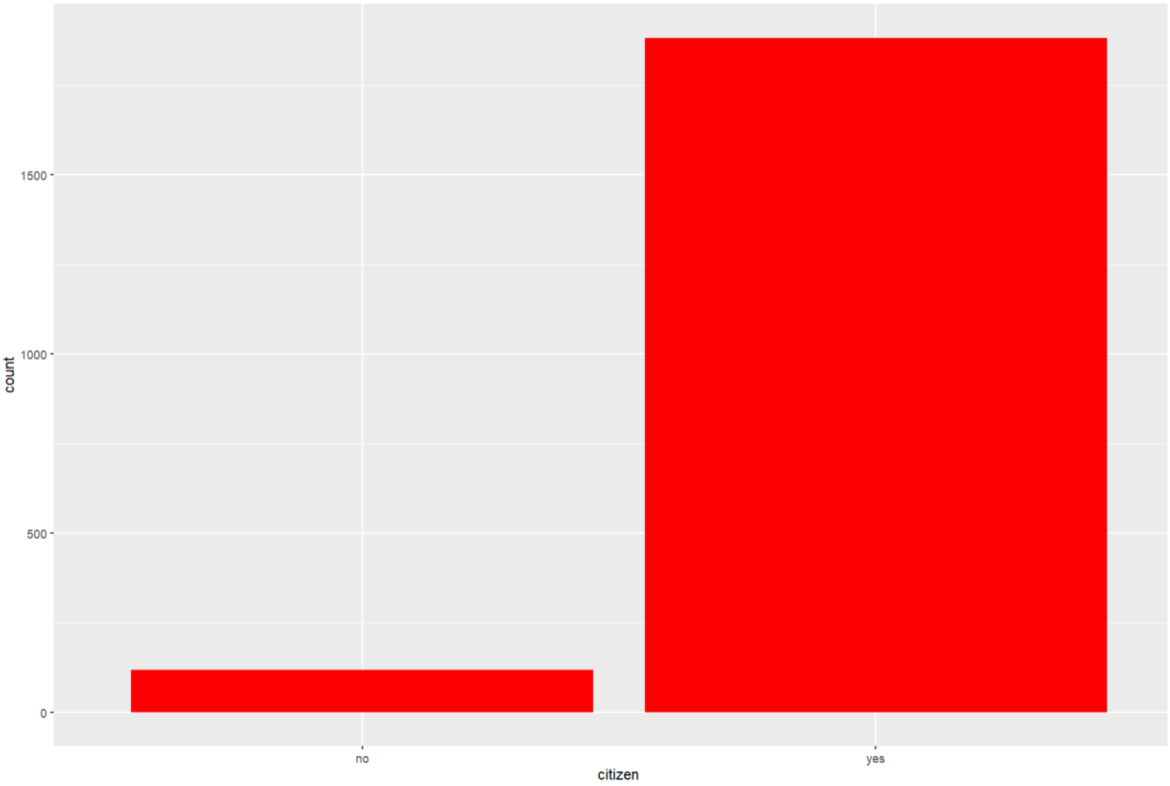


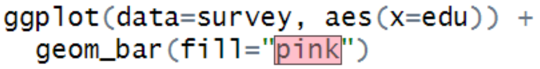
```
ggplot(data=survey, aes(x=gender)) +
  geom_bar(fill="green")
```



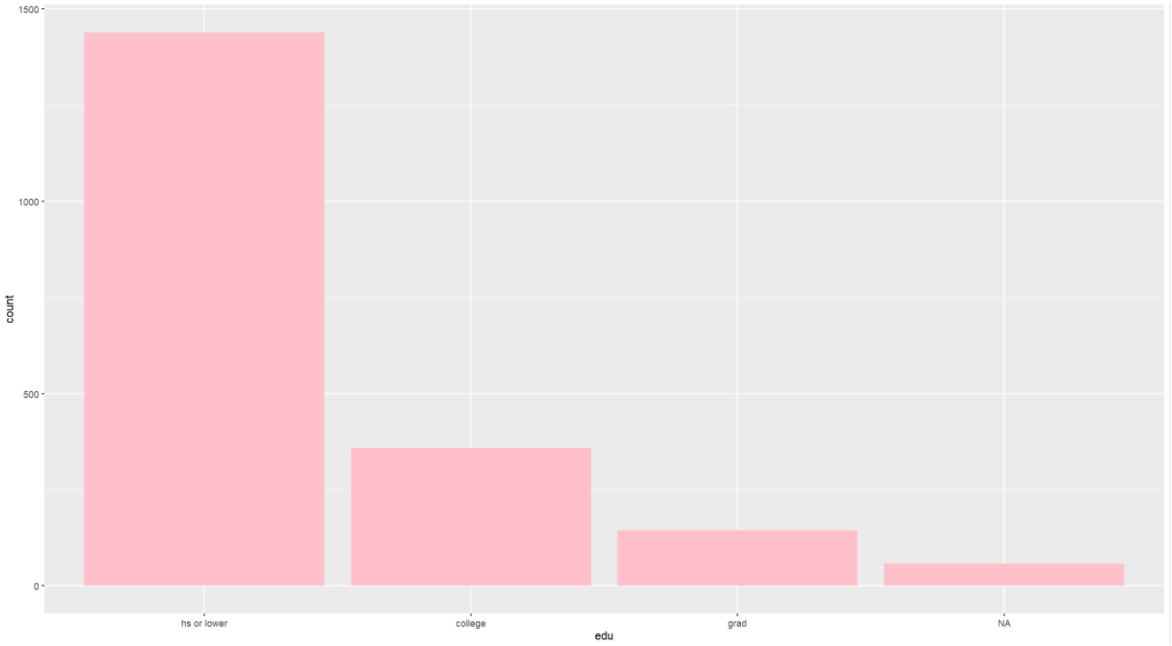


```
ggplot(data=survey, aes(x=citizen)) +
  geom_bar(fill="red")
```





```
ggplot(data=survey, aes(x=edu)) +
  geom_bar(fill="pink")
```



Exercise 3: Create a table of gender vs. employment and copy this into your lab write-up. Then answer the following probability questions, each time assuming you are choosing a randomly selected participant who has provided data (excluding NAs). Provide your answer as an un-simplified fraction AND as a decimal or percent accurate to 3 significant digits.

```
 -----------------------------------------------------
employed | not in labor force | unemployed
  female    | 373               373 |       47
-----------------------------------------------------
  male      | 470               283 |       59
-----------------------------------------------------
```

g) Probability of selecting someone who is employed.
   - P(Employed) = (843/1605) = (470+373)/1605 = .525

g) Probability of selecting someone who is unemployed.
   - P(!employed) = (106/1605) = (59+47)/1605 = .066

g) Probability of selecting someone who is not in the labor force.
   - P(!labor force) = (656/1605) = (283+373)/1605 = .409

g) Probability that the person you selected is a woman, given they are employed.
   - P(Woman | Employed) = 373/843 = .442

g) Probability that the person you selected is employed, given they are a woman.
   - P(Employed | Woman) = 373/(373+373+47) = 0.470

g) Probability that the person you selected is male, given they are not in the labor force.
   - P(Male | Not in the work force) = 283/656 = .431

g) Probability that the person you selected is not in the labor force, given they are male.
   - P(Not in labor force | Male) = 283/(383+59+470) = .349

Exercise 4: To answer the following probability questions, you will need to create additional tables. Copy relevant tables into your lab write-up. Provide your answer as an un-simplified fraction AND as a decimal or percent accurate to 3 significant digits.

```
gender_citizen_table          race_citizen_table
  Citizen| no yes               Citizen | No Yes
  male   | 51 980               White   | 19  1506
  female | 67 902               Black   | 12  194
                           Asian    | 30  57
                           Other    | 27  125
```

i) Probability that a randomly selected female is a citizen. i.e. P(citizen | female)
   -P(Citizen | Female) = 902/969 = 0.931

i) Probability that a randomly selected male is a citizen.
- P(Citizen | Male) = 980/1031 = 0.951

i) Are your answers for (a) and (b) pretty close to each other, in your opinion? Or farther part than you expected? Answer in a complete sentence.

- We thought that they were close to each other only being .02 difference. We thought this was expected because the overall population of men and women are relatively close, and the citizenship of men and women are close.

i) Probability that a randomly selected person is a female and a non-citizen.
- P (Female and Non-Citizen) = 67/2000 = 0.034

i) Probability that a randomly selected black person is a citizen.
- P(Citizen | Black) = 194/206 = 0.942

i) Probability that a randomly selected Asian person is a citizen.
- P(Citizen | Asian) = 57/87 = 0.655

i) P(person is white | person is not a citizen)
- P(White | non-Citizen) = 49/118 = 0.415

i) P(person is not a citizen | person is white)
- P(non-Citizen | White) = 49/1555= 0.031

i) In a few sentences, explain why the questions for (g) and (h) are quite different.
- These two are different because there are two different groups that are being compared. In Question “G” they have a sample space of 118, which is all the people that are non-citizens, and in question “H” they have a sample space of all the people that are white of which there are 1555. Since the total number of people in “G” is much different then “H” they have a large difference in the probability, despite sounding similar when reading the question.

Applying conditional probability to Covid-19 Testing

Based on a May 2020 journal article in BMJ by Watson, Whiting, and Brush (https://www.bmj.com/content/bmj/369/bmj.m1808.full.pdf), we will investigate conditional probability questions related to the RT-PCR testing for Covid-19. The article states “A systematic review of the accuracy of covid-19 tests reported false negative rates of between 2% and 29% (equating to sensitivity of 71-98%), based on negative RT-PCR tests which were positive on repeat testing" (p.1). Reasons for false negatives include when the sample was taken (relative to the progression of the illness) and the how the test was administered (throat swap, nasal, etc.) Meanwhile, false positive rates were much lower because the tests are looking for specific genetic material from the virus.

Exercise 5: Let's assume that you are in a community with 5000 people. The rate of Covid-19 in your community is 3.8%. According to the local lab, if a patient has Covid-19, there is a 74% chance they will get a correct positive result. If the patient does not have Covid-19, there is a 98.5% chance they will get the correct negative result. Answer the following and describe how you solved each problem.

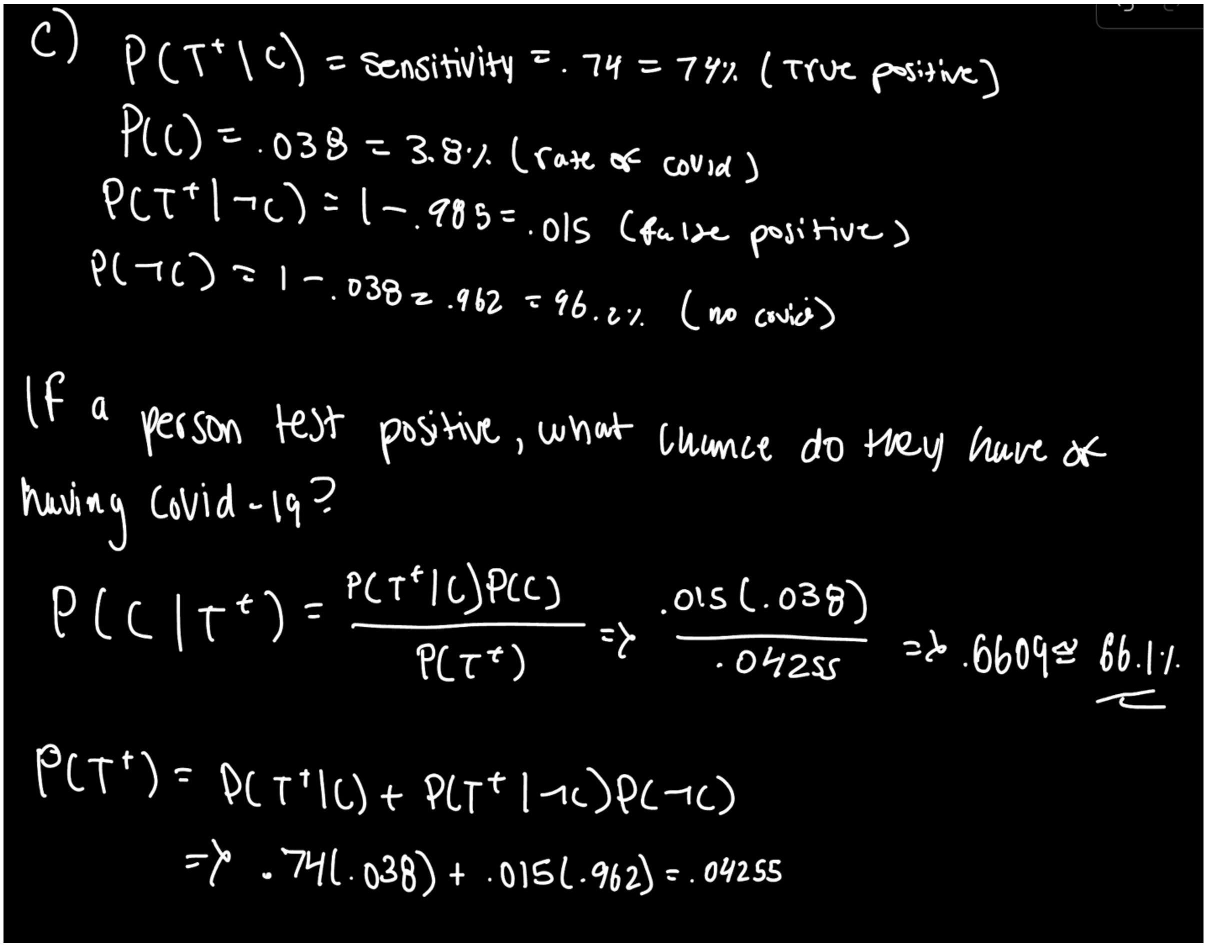
f) About how many people in this community will have Covid-19 and test positive? (round to the nearest whole number of people)
True Positive = P(Covid-19) * N * sensitivity
=> .038 * 5000 * .74 = 141 people
the percentage of people who have covid * total number of people * likelyhood correct postitive result

Sensitivity: .74
Specificity: not looking for negative results

f) If a person is selected at random, what's the chance they have Covid-19?
P(Covid-19) = 3.8% => .038
any random person in the community has the same rate of having Covid-19

Sensitivity: simple calc, not test related
Specificity: simple calc, not test related

f) If a person tests positive, what is the chance they have Covid-19?



c) $P(T^+ \mid c) = \text{Sensitivity} = .74 = 74\%$ (True positive)

$P(C) = .038 = 3.8\%$ (rate of covid)

$P(T^+ \mid \neg c) = 1 - .985 = .015$ (false positive)

$P(\neg c) = 1 - .038 = .962 = 96.2\%$ (no covid)

If a person test positive, what chance do they have of having Covid-19?

$$P(C \mid T^+) = \frac{P(T^+ \mid C)P(C)}{P(T^+)} \Rightarrow \frac{.015(.038)}{.04255} \Rightarrow .6609 \cong \underline{66.1\%}$$

$$P(T^+) = P(T^+ \mid C) + P(T^+ \mid \neg c)P(\neg c)$$

$$\Rightarrow .74(.038) + .015(.962) = .04255$$

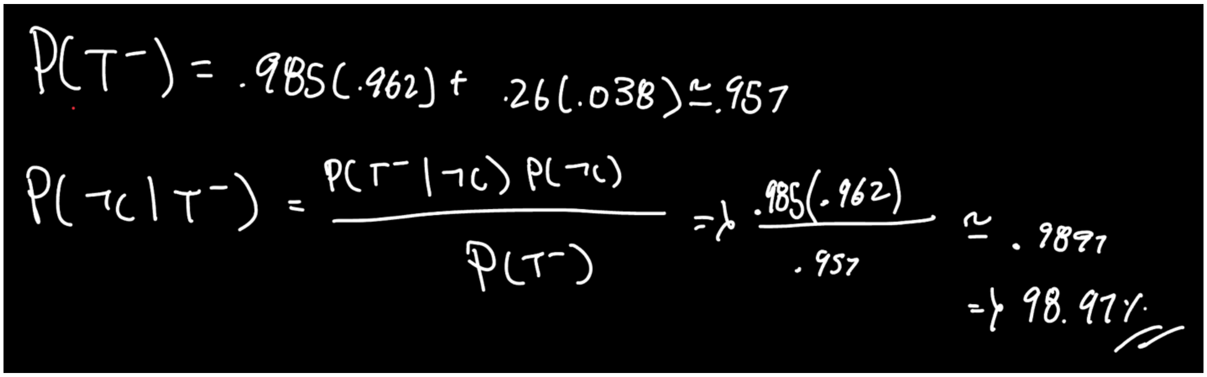
since false positives affect true positives, we use bayes theorem to account for it. This gives the true probability of having COVID-19 after a positive test, which in our case is 66.1%, not just the test’s 74% sensitivity.

Sensitivity: .74
Specificity: .985

f) If a person is selected at random, what’s the chance they do not have Covid-19?
P(!Covid-19) = 1 – .038 = .962 => 96.2%
any random person in the community has the same chance of not having Covid-19

Sensitivity: compliment, not test related
Specificity: compliment, not test related

f) If a person tests negative, what is the chance they do not have Covid-19?



$$P(T^-) = .985(.962) + .26(.038) \simeq .957$$

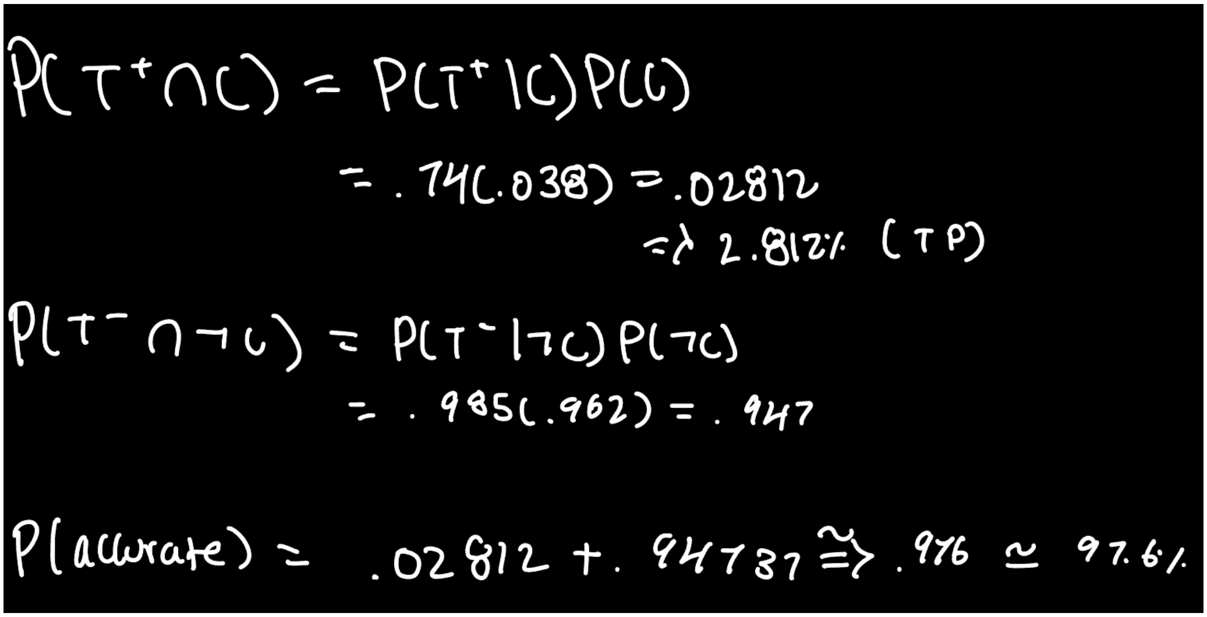
$$P(\neg c \mid T^-) = \frac{P(T^- \mid \neg c)\,P(\neg c)}{P(T^-)} \Rightarrow \frac{.985(.962)}{.957} \simeq .9897$$

$$\Rightarrow 98.97\%$$

Since false negatives exist, a negative test doesn’t guarantee someone doesn‘t have covid. Bayes’ Theorem finds the true chance of not having COVID by comparing true negatives to all negative tests, correcting for missed cases.

Sensitivity: .74
Specificity: .985

f) What percent of people get accurate test results?



$$P(T^+ \cap C) = P(T^+ \mid C)\,P(C)$$

$$= .74(.038) = .02812$$

$$\Rightarrow 2.812\% \text{ (TP)}$$

$$P(T^- \cap \neg c) = P(T^- \mid \neg c)\,P(\neg c)$$

$$= .985(.962) = .947$$

$$P(\text{accurate}) = .02812 + .94737 \overset{\sim}{\Rightarrow} .976 \simeq 97.6\%$$

Accurate test results include both true positives and true negatives. The total accuracy is found by adding the chances of correctly detecting COVID and correctly identifying those without it. This ensures the final percentage reflects all people who got the right test result.

Sensitivity: .74
Specificity: .985

Exercise 6: Look up the terms “sensitivity” and “specificity” as they relate to clinical tests. Explain what these terms mean, in language appropriate for a friend who is not yet taking statistics and is not in the medical field. Then go back to Exercise 5 and clearly identify the sensitivity and specificity of the test described.

Sensitivity: this tells how good a test is at finding people who have the condition. A test with high sensitivity correctly identifies most sick people, meaning it does not miss many real cases. However, it might still include some healthy people by mistake (false positives). for example, if a covid-19 test has 90% sensitivity, it will catch 90 out of 100 infected people, but it may still miss 10 who have covid (false negatives).

Specificity: this tells how good a test is at ruling out people who don’t have the condition. A test with high specificity correctly identifies most healthy people, meaning it rarely gives a false alarm (false positives). However, it might miss some actual cases (false negatives). for example, if a covid-19 test has 95% specificity, it correctly tells 95 out of 100 healthy people that they are negative, but it may incorrectly mark 5 healthy people as positive (false positives).

Code:

```
#library(tidyverse)
#library(openintro)
```

```
library(ggplot2)
survey <- acs12

----- 1. Task 1 (Summary Tables) -----
cat("Summary of Race:\n")
print(summary(survey$race))
cat("\nSummary of Gender:\n")
print(summary(survey$gender))
cat("\nSummary of Citizenship:\n")
print(summary(survey$citizen))
cat("\nSummary of Employment:\n")
print(summary(survey$employment))
cat("\nSummary of Education Level:\n")
print(summary(survey$edu))

----- 2. Task 2 (Bar Graphs) -----
ggplot(survey, aes(x = employment)) + geom_bar(fill = "steelblue") + ggtitle("Employment Distribution")
ggplot(survey, aes(x = gender)) + geom_bar(fill = "red") + ggtitle("Gender Distribution")
ggplot(survey, aes(x = race)) + geom_bar(fill = "lightgreen") + ggtitle("Race Distribution")
ggplot(survey, aes(x = citizen)) + geom_bar(fill = "purple") + ggtitle("Citizenship Status Distribution")
ggplot(survey, aes(x = edu)) + geom_bar(fill = "orange") + ggtitle("Education Level Distribution")

----- 3. Task 3 (Probability Tables) -----

cat("\nContingency Table: Gender vs Employment\n")
print(table(survey$gender, survey$employment))

cat("\nProbability Table (Proportions): Gender vs Employment\n")
print(prop.table(table(survey$gender, survey$employment)))

----- 4. Task 4 (Employment Probabilities) -----

P_employed <- sum(survey$employment == "employed", na.rm = TRUE) / nrow(survey)
P_unemployed <- sum(survey$employment == "unemployed", na.rm = TRUE) / nrow(survey)

cat("\nProbability of being employed:", P_employed, "\n")
cat("Probability of being unemployed:", P_unemployed, "\n")
```

```
----- 5. Task 5 (Conditional Probability: P(Woman | Employed)) -----
P_woman_given_employed <- prop.table(table(survey$gender, survey$employment), margin =
2)
cat("\nConditional Probability: P(Woman | Employed)\n") print(P_woman_given_employed)

----- 6. Task 6 (COVID-19 Testing Probability (Bayes' Theorem)) -----

#Given data
N <- 5000 # Total population P_covid <- 0.038 # Probability of having COVID-19
sensitivity <- 0.74 # True positive rate
specificity <- 0.985 # True negative rate

#Calculate values
covid_cases <- P_covid * N
true_pos <- covid_cases * sensitivity
false_pos <- (N - covid_cases) * (1 - specificity) #(N^c)

#Bayes' Theorem: P(COVID | Positive Test)
P_covid_given_positive <- true_pos / (true_pos + false_pos)
cat("\nProbability of having COVID given a positive test result:", P_covid_given_positive, "\n")
```