# Patel's Week 3 Dataset Intro to Probability

Harsh Patel

2024-10-11

```
library(readxl)
Chronic_Kidney_Disease_data <-
read_excel("C:/Users/hpate/Downloads/Chronic_Kidney_Dsease_data.xlsx")
str(Chronic_Kidney_Disease_data)
```

```
## tibble [1,659 × 54] (S3: tbl_df/tbl/data.frame)
##  $ PatientID                    : num [1:1659] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                          : num [1:1659] 71 34 80 40 43 22 41 72 21
49 ...
##  $ Gender                       : num [1:1659] 0 0 1 0 0 0 0 1 0 0 ...
##  $ Ethnicity                    : num [1:1659] 0 0 1 2 1 0 1 0 1 3 ...
##  $ SocioeconomicStatus          : num [1:1659] 0 1 0 0 1 0 0 1 0 0 ...
##  $ EducationLevel               : num [1:1659] 2 3 1 1 2 1 1 3 2 1 ...
##  $ BMI                          : num [1:1659] 31.1 29.7 37.4 31.3 23.7
...
##  $ Smoking                      : num [1:1659] 1 1 1 0 0 0 0 1 0 0 ...
##  $ AlcoholConsumption           : num [1:1659] 5.13 18.61 11.88 16.02 7.94
...
##  $ PhysicalActivity             : num [1:1659] 1.676 8.378 9.607 0.409
0.78 ...
##  $ DietQuality                  : num [1:1659] 0.24 6.5 2.1 6.96 3.1 ...
##  $ SleepQuality                 : num [1:1659] 4.08 7.65 4.39 6.28 4.02
...
##  $ FamilyHistoryKidneyDisease   : num [1:1659] 0 1 0 0 0 0 0 0 0 0 ...
##  $ FamilyHistoryHypertension    : num [1:1659] 0 1 0 0 0 0 1 0 0 0 ...
##  $ FamilyHistoryDiabetes        : num [1:1659] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousAcuteKidneyInjury    : num [1:1659] 0 0 0 0 0 0 0 0 0 0 ...
##  $ UrinaryTractInfections       : num [1:1659] 0 0 0 0 0 0 0 0 1 0 ...
##  $ SystolicBP                   : num [1:1659] 113 120 147 117 98 134 179
138 91 95 ...
##  $ DiastolicBP                  : num [1:1659] 83 67 106 65 66 79 76 95 67
119 ...
##  $ FastingBloodSugar            : num [1:1659] 72.5 100.8 161 188.5 82.2
...
##  $ HbA1c                        : num [1:1659] 9.21 4.6 5.43 4.14 4.26 ...
##  $ SerumCreatinine              : num [1:1659] 4.96 3.16 3.7 2.87 3.96 ...
##  $ BUNLevels                    : num [1:1659] 25.6 31.3 39.7 22 12.2 ...
##  $ GFR                          : num [1:1659] 45.7 55.8 67.6 33.2 56.3
...
##  $ ProteinInUrine               : num [1:1659] 0.745 3.052 1.158 3.746
2.571 ...
##  $ ACR                          : num [1:1659] 123.8 88.5 21.2 123.8 184.9
```

```
...
##  $ SerumElectrolytesSodium      : num [1:1659] 138 138 143 137 141 ...
##  $ SerumElectrolytesPotassium   : num [1:1659] 3.63 5.33 4.33 3.81 4.87
...
##  $ SerumElectrolytesCalcium     : num [1:1659] 10.31 9.6 9.89 10 8.91 ...
##  $ SerumElectrolytesPhosphorus  : num [1:1659] 3.15 2.86 4.35 4.02 3.95
...
##  $ HemoglobinLevels             : num [1:1659] 16.1 15.3 13 15.1 16.7 ...
##  $ CholesterolTotal             : num [1:1659] 208 189 284 235 258 ...
##  $ CholesterolLDL               : num [1:1659] 85.9 86.4 132.3 93.4 171.8
...
##  $ CholesterolHDL               : num [1:1659] 22 87.6 20 58.3 21.6 ...
##  $ CholesterolTriglycerides     : num [1:1659] 212 255 252 392 371 ...
##  $ ACEInhibitors                : num [1:1659] 0 0 0 0 1 1 0 1 0 0 ...
##  $ Diuretics                    : num [1:1659] 0 0 1 0 1 0 0 1 0 0 ...
##  $ NSAIDsUse                    : num [1:1659] 4.56 9.1 3.85 7.88 4.18 ...
##  $ Statins                      : num [1:1659] 1 0 1 0 1 0 1 0 1 0 ...
##  $ AntidiabeticMedications      : num [1:1659] 0 0 0 0 0 1 0 0 0 0 ...
##  $ Edema                        : num [1:1659] 0 0 0 0 0 1 0 0 0 0 ...
##  $ FatigueLevels                : num [1:1659] 3.56 5.33 4.86 8.53 1.42
...
##  $ NauseaVomiting               : num [1:1659] 6.992 0.356 4.674 5.691
2.273 ...
##  $ MuscleCramps                 : num [1:1659] 4.52 2.2 5.97 2.18 6.8 ...
##  $ Itching                      : num [1:1659] 7.56 6.84 2.14 7.08 3.55
...
##  $ QualityOfLifeScore           : num [1:1659] 76.08 40.13 92.87 90.08
5.26 ...
##  $ HeavyMetalsExposure          : num [1:1659] 0 0 0 0 0 0 0 0 0 0 ...
##  $ OccupationalExposureChemicals: num [1:1659] 0 0 1 0 0 0 0 0 0 0 ...
##  $ WaterQuality                 : num [1:1659] 1 0 1 0 1 0 1 0 1 0 ...
##  $ MedicalCheckupsFrequency     : num [1:1659] 1.019 3.924 1.43 3.226
0.285 ...
##  $ MedicationAdherence          : num [1:1659] 4.97 8.19 7.62 3.28 3.85
...
##  $ HealthLiteracy               : num [1:1659] 9.87 7.16 7.35 6.63 1.44
...
##  $ Diagnosis                    : num [1:1659] 1 1 1 1 1 1 1 0 1 1 ...
##  $ DoctorInCharge               : chr [1:1659] "Confidential"
"Confidential" "Confidential" "Confidential" "Confidential" ...
```

**After loading the Chronic Kidney Disease dataset, I checked its structure using `str()`. This revealed that the dataset contains 1,659 observations with 54 variables, including information like Age, Gender, and Smoking status.**

```
Chronic_Kidney_Disease_data$Gender <-
factor(Chronic_Kidney_Disease_data$Gender, levels = c(0, 1), labels =
```

```
c("Male", "Female"))
gender_proportions <- prop.table(table(Chronic_Kidney_Disease_data$Gender))
gender_proportions

##
##      Male    Female
## 0.4846293 0.5153707
```

**When I calculated the gender proportions, I found that approximately 48.5% of the participants are male and 51.5% are female. This gives me a fairly balanced representation of genders in the dataset.**

```
Chronic_Kidney_Disease_data$Smoking <-
factor(Chronic_Kidney_Disease_data$Smoking, levels = c(0, 1, 2),
                                                labels = c("Non-smoker",
"Former smoker", "Current smoker"))
smoking_proportions <- prop.table(table(Chronic_Kidney_Disease_data$Smoking))
smoking_proportions

##
##     Non-smoker  Former smoker Current smoker
##      0.7070524      0.2929476      0.0000000
```

**For smoking status, I found that about 70.7% of the participants are non-smokers, 29.3% are former smokers, and none are current smokers. This indicates that the majority of participants do not smoke.**

```
prob_male <- gender_proportions["Male"]
prob_female <- gender_proportions["Female"]
number_employees <- 150
set.seed(3000)
gender_simulation <- sample(c("Male", "Female"), size = number_employees,
                            prob = c(prob_male, prob_female), replace = TRUE)
table(gender_simulation)

## gender_simulation
## Female   Male
##     81     69
```

**In my simulation of gender distribution among 150 employees, I observed 69 males and 81 females. This simulates the actual proportions I calculated from the dataset.**

```
prob_non_smoker <- smoking_proportions["Non-smoker"]
prob_former_smoker <- smoking_proportions["Former smoker"]
```

```r
prob_current_smoker <- smoking_proportions["Current smoker"]
smoking_simulation <- sample(c("Non-smoker", "Former smoker", "Current
smoker"),
                             size = number_employees,
                             prob = c(prob_non_smoker, prob_former_smoker,
prob_current_smoker),
                             replace = TRUE)
table(smoking_simulation)
```

```
## smoking_simulation
## Former smoker    Non-smoker
##            49           101
```

## During the smoking simulation, I found that among the 150 employees, there were 101 non-smokers and 49 former smokers. Again, this reflects the proportions from the dataset.

```r
prob_at_least_one_male <- 1 - (1 - prob_male) ^ number_employees
prob_at_least_one_male
```

```
## Male
##    1
```

## I calculated the probability of having at least one male in a group of 150 employees, and it turned out to be 1. This means I am certain that at least one male would be present in the group.

```r
number_replicates <- 100000
results_gender <- numeric(number_replicates)
set.seed(3000)
for (k in 1:number_replicates) {
  gender_replicate <- sample(c("Male", "Female"), size = number_employees,
                             prob = c(prob_male, prob_female), replace =
TRUE)
  results_gender[k] <- sum(gender_replicate == "Male")
}
table(results_gender)
```

```
## results_gender
##   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62
63
##    5    3    6   11   16   37   49  101  144  236  360  540  729 1045 1360
1848
##   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78
79
## 2418 2920 3573 4242 4930 5517 5849 6227 6423 6496 6402 6139 5750 5098 4490
3790
##   80   81   82   83   84   85   86   87   88   89   90   91   92   93   94
```

```
95
## 3162 2524 2056 1550 1189  870  660  433  321  191  129   61   51   22   14
4
##   96   97   98
##    7    1    1
```

**In my simulation of 100,000 replicates, I counted the number of males in each group. The results showed that the number of males varied, with some replicates having as many as 85 males. Most replicates had at least one male.**

```
at_least_one_male <- (results_gender >= 1)
table(at_least_one_male)

## at_least_one_male
##   TRUE
## 100000

prob_at_least_one_male_sim <- sum(at_least_one_male) / number_replicates
prob_at_least_one_male_sim

## [1] 1
```

**The results indicate that in all 100,000 simulations, there was at least one male present in each replicate. This reinforces the certainty that at least one male would appear in a group of 150 employees.**

```
prob_at_least_one_current_smoker <- 1 - (1 - prob_current_smoker) ^
number_employees
prob_at_least_one_current_smoker

## Current smoker
##              0
```

**I calculated the probability of having at least one current smoker among the 150 employees and found it to be 0. This indicates that in this dataset, it is impossible to have a current smoker given that there were none in the original data.**

```
results_smoking <- numeric(number_replicates)
set.seed(3000)
for (k in 1:number_replicates) {
  smoking_replicate <- sample(c("Non-smoker", "Former smoker", "Current
smoker"),
                              size = number_employees,
```

```
                                  prob = c(prob_non_smoker, prob_former_smoker,
prob_current_smoker),
                                  replace = TRUE)
  results_smoking[k] <- sum(smoking_replicate == "Current smoker")
}
table(results_smoking)

## results_smoking
##      0
## 100000
```

**In the smoking simulation with 100,000 replicates, the results indicated that there were 0 instances of having a current smoker, reinforcing the earlier finding that there are no current smokers in the dataset.**

```
at_least_one_current_smoker <- (results_smoking >= 1)
table(at_least_one_current_smoker)

## at_least_one_current_smoker
##  FALSE
## 100000

prob_at_least_one_current_smoker_sim <- sum(at_least_one_current_smoker) /
number_replicates
prob_at_least_one_current_smoker_sim

## [1] 0
```

**Overall, my simulations confirmed that there were no current smokers in any of the 100,000 replicates, aligning with the original dataset, which indicates that none of the participants smoke currently.**