# Problem set 2: More probability

**Due October 14, 2024, at 10am**

(Your name here)

*NOTE: Start with the file* `ps2_2024_more_probability.qmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F24/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the* `Render` *button). Submit both the qmd file and the PDF via Canvas.*

## Problem 1: Bayes' Rule

Bayes Rule expresses the relationship between a conditional probability, e.g. $P(A \mid B)$, and the "reverse" conditional probability $P(B \mid A)$.

One formulation of Bayes' Rule states that, if $\{A_1, A_2, \ldots A_n\}$ is a partition of $\Omega$ with $P(A_i) > 0$ for $i = 1, 2, \ldots, n$ and $B \in S$ with $P(B) > 0$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

$\forall i$.

Here is a proof of Bayes Rule that is missing explanations for the steps:

Step 1:

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

Step 2:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$

Step 3:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

Explain each step in the proof: what definition(s)/rule(s)/law(s)/axiom(s)/condition(s)/mathematical operation(s) is the proof relying on?

**Answer**:

- Step 1: Definition of conditional probability.
- Step 2: Product rule.
- Step 3: Law of total probability.

## Problem 2: Error rates in hypothesis testing

You have a fancy device that tests null hypotheses. Null hypotheses are statements about the world that can be either true or false. The device is designed to turn red when a null hypothesis is false and green when it is true, but it doesn't work perfectly: when a null hypothesis is false it turns red with probability 3/4 (i.e. it mistakenly turns green with probability 1/4), and when a null hypothesis is true it turns green with probability 19/20 (i.e. it mistakenly turns red with probability 1/20). Tests of different null hypotheses are independent, and 4/5 of the null hypotheses you test are true.

(1a) If you test 12 true null hypotheses in a row, what is the probability that the alarm turns red at least once? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

**Answer**:

The probability of the alarm turning red in a single test of a true null hypothesis is given as 1/20; the probability of the alarm turning green is given as 19/20.

We could compute the probability of getting one red light, two red lights, etc all the way to 20 red lights. But by the **complement rule**, the probability of the the alarm turning red at least once in 12 tries is one minus the probability of the alarm turning green 12 times in a row. That's easier to compute.

Given that tests are independent, the probability of the alarm turning green 12 times in a row is $\left(\frac{19}{20}\right)^{12}$.

Therefore the probability of the light turning red at least once in 12 tests of a true null hypothesis is

$$1 - \left(\frac{19}{20}\right)^{12} \approx .46$$

(1b) Write a simulation to check your answer to (1a). That is, use `R` to generate many draws according to the random process described (each time testing twelve true null hypotheses in a row), and confirm that the proportion of draws with at least one red light is approximately the same as in your answer above.

```
# your code here
samp <- rbinom(n = 100000, size = 12, prob = 1/20)
mean(samp > 0)
```

```
[1] 0.46216
```

Indeed, the proportion with at least one red light is very close to the answer I gave above.

(Here I am using the fact that `samp > 0` generates a vector that is `TRUE` for every entry where `samp > 0` and `FALSE` for the others, and that `mean()` applied to a vector of `TRUE` and `FALSE` gives the proportion of `TRUE` entries in the vector.)

(1c) Now suppose you come upon a null hypothesis at random; you don't know if it is true or false in this case. What is the probability of getting a red light when you run your test? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

**Answer**:

By the law of total probability,

$$P(R) = P(R \mid T)P(T) + P(R \mid F)P(F),$$

where $T$ indicates the event that the null hypothesis is true and $F$ indicates the event that the null hypothesis is false, and $R$ indicates the event that the light turns red. Note that $P(R \mid T) = 1/20$ and $P(R \mid F) = 3/4$ are given in the problem, and $P(T) = 4/5$ is also given (so that $P(F) = 1/5$). Applying these facts, we have $P(R) = \frac{1}{20}\frac{4}{5} + \frac{3}{4}\frac{1}{5} = \frac{19}{100}$.

(1d) If the light turns red in a given test, what is the probability that the null hypothesis is true? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

**Answer**:

The objective here is to compute $P(T \mid R)$, a conditional probability.

By Bayes Rule, we have

$$P(T \mid R) = \frac{P(R|T)P(T)}{P(R)}.$$

(It can also be written $P(T \mid R) = \frac{P(R|T)P(T)}{P(R|T)P(T)+P(R|F)P(F)}$.)

We computed the denominator in the previous question (19/100). The numerator is $\frac{1}{20}\frac{4}{5}$, so the answer is $\frac{4}{19}$.

(1e) Write a simulation to check your answer to (1c) and (1d). That is, use `R` to generate many draws according to the random process described (testing null hypotheses), and confirm your answer about the proportion of red-light-producing draws (1c) and the proportion of red-light-producing draws in which the null hypothesis is actually true (1d).

**Answer**:

The objective is to produce a dataset based on the specified random process and then confirm that you can compute proportions that correspond to the desired marginal and conditional probabilities.

```
# one approach
n <- 100000
null_is_true <- sample(x = c(0, 1), size = n, replace = T, prob = c(1/5, 4/5))
red_light <- rep(NA, n) # making an empty "holder"
red_light[null_is_true == 1] <- sample(c(0, 1),
                                       size = sum(null_is_true == 1),
                                       replace = T,
                                       prob = c(.95, .05))
red_light[null_is_true == 0] <- sample(c(0, 1),
                                       size = sum(null_is_true == 0),
                                       replace = T,
                                       prob = c(.25, .75))
# checking 1c
mean(red_light)
```

```
[1] 0.19083
```

```
#checking 1d
mean(null_is_true[red_light == 1])
```

```
[1] 0.2122308
```

The proportion of simulated tests that produce a red light is 0.191 (compared to a theoretical value of 0.19).

The proportion of null hypotheses that are true *given a red light* is 0.212 (compared to a theoretical value of 0.2105263).

Note that in this case I used numerical values (0 and 1) for both vectors. An even more flexible option is to use the "logical" values `F` and `T` (short for `FALSE` and `TRUE`). The key point here is that `R` turns `F` into 0 and `T` into 1 if you try to do a numerical operation, so that `mean(c(F, T, F, T, T))` evaluates to .6.

```
n <- 100000
null_is_true <- sample(x = c(T, F), size = n, replace = T, prob = c(4/5, 1/5))
red_light <- rep(NA, n) # making an empty "holder"
red_light[null_is_true] <- sample(c(F, T),
                                  size = sum(null_is_true),
                                  replace = T,
                                  prob = c(.95, .05))
red_light[!null_is_true] <- sample(c(F, T),
                                  size = sum(!null_is_true),
                                  replace = T,
                                  prob = c(.25, .75))
# checking 1c
mean(red_light)
```

```
[1] 0.18923
```

```
#checking 1d
mean(null_is_true[red_light])
```

```
[1] 0.2101675
```

Just for `R` practice purposes, here is a different approach, using character "F" and "T":

```
# another approach:
# first we draw the null hypotheses -- true or false
nulls <- sample(x = c("F", "T"), size = n, replace = T, prob = c(1/5, 4/5))
# now the test results
results_for_false_nulls <- sample(x = c("R", "G"), size = sum(nulls == "F"), replace = T,
```

```r
results_for_true_nulls <- sample(x = c("R", "G"), size = sum(nulls == "T"), replace = T, p
# concatenate the results
test_results <- c(results_for_false_nulls, results_for_true_nulls)
# concatenate the truth of the null hypothesis
null_status <- c(rep("F", sum(nulls == "F")), rep("T", sum(nulls == "T")))
# proportion of reds (1c)
mean(test_results == "R")
```

```
[1] 0.18952
```

```r
# P(F | R)  (1d)
mean(null_status[test_results == "R"] == "T")
```

```
[1] 0.2096876
```

(1f) Using a similar calculation, John Ioannides reported in a famous 2005 paper that "most published research findings are false". By this he meant that the probability of the null hypothesis being true, given that the researcher rejected the null hypothesis (i.e. the light turns red), is above 1/2. Show one way to change the assumptions in the problem to produce that result.

**Answer**:

## Problem 3: discrete random variables

Suppose I am sampling a voter at random from a population. I care about two characteristics of the voter, which I characterize using numbers: whether she supports the populist candidate ($X = 1$ if so, 0 otherwise) and her level of education ($Y = 0$ if less than college degree, $Y = 1$ if college degree, $Y = 2$ if more than a college degree).

You happen to know the joint distribution of these characteristics in the population from which I am sampling. The resulting joint PMF for my random variables is:

$$f_{X,Y}(x,y) = \begin{cases} 1/12 & x = 0, y = 0 \\ 1/6 & x = 1, y = 0 \\ 1/4 & x = 0, y = 1 \\ 1/4 & x = 1, y = 1 \\ 3/16 & x = 0, y = 2 \\ 1/16 & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3a) What is the marginal PMF of $X$, $f_X(x)$? Replace the question marks with your answers.

$$f_X(x) = \begin{cases} ? & x = 0 \\ ? & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

(3b) What is the marginal PMF of $Y$, $f_Y(y)$?

$$f_Y(y) = \begin{cases} ? & y = 0 \\ ? & y = 1 \\ ? & y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3c) What is the conditional PMF $f_{X|Y}(x|y)$?

$$f_{X|Y}(x|y) = \begin{cases} ? & x = 0, y = 0 \\ ? & x = 1, y = 0 \\ ? & x = 0, y = 1 \\ ? & x = 1, y = 1 \\ ? & x = 0, y = 2 \\ ? & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3d) What is the conditional PMF $f_{Y|X}(y|x)$?

$$f_{X|Y}(x|y) = \begin{cases} ? & x = 0, y = 0 \\ ? & x = 1, y = 0 \\ ? & x = 0, y = 1 \\ ? & x = 1, y = 1 \\ ? & x = 0, y = 2 \\ ? & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

## Problem 4: continuous random variables

(4a) Let $X$ be uniformly distributed between -5 and 3. Compute $\Pr[X < 2]$ and $\Pr[-4 < X < -1/2]$ analytically (e.g. by computing the length and height of the area to be integrated) and confirm your results using a simulation in R.

**Answer**:

The height of the pdf is 1/8 (because the length is 8 and the area under the pdf must be 1). So $\Pr[X < 2]$ is $7 \times 1/8 = 7/8$.

$\Pr[-4 < X < -1/2]$ is $3.5 \times 1/8 = 7/16$.

Confirmation in R:

```
samps <- runif(1000000, min = -5, max = 3)
mean(samps < 2)
```

```
[1] 0.875272
```

```
mean(samps > -4 & samps < -1/2)
```

```
[1] 0.438326
```

(4b) Let $X$ be normally distributed with mean 2 and standard deviation 1.25. Using R, compute (i) $\Pr[X < 0]$, (ii) $\Pr[1 < X < 3]$, and (iii) $\Pr[X > 3.5]$

**Answer**:

```
#(i)
pnorm(0, mean = 2, sd = 1.25)
```

```
[1] 0.05479929
```

```
#(ii)
pnorm(3, mean = 2, sd = 1.25) - pnorm(1, mean = 2, sd = 1.25)
```

```
[1] 0.5762892
```

```r
#(iii)
1 - pnorm(3.5, mean = 2, sd = 1.25)
```

```
[1] 0.1150697
```

(4c) As we discussed in class, if $X$ is a continuous random variable, $f(x)$ can be used to approximate the probability of getting a value near $x$. Suppose $X$ is normally distributed with mean 3 and standard deviation 1.5. Use $f(x)$ (`dnorm()`) to approximate the probability of obtaining a value within .01 of 2. Then use $F(x)$ (`pnorm()`) to obtain the exact value. Finally, draw a large number of random samples using `rnorm()` to obtain a numerical estimate of the same value.

**Answer**:

```r
# approximation via f(x)
dnorm(2, mean = 3, sd = 1.5)*.02
```

```
[1] 0.004259307
```

```r
# exact answer via F(x)
pnorm(2 + .01, mean = 3, sd = 1.5) - pnorm(2 - .01, mean = 3, sd = 1.5)
```

```
[1] 0.004259289
```

```r
# approximation via sampling
samps <- rnorm(n = 100000, mean = 3, sd = 1.5)
mean(samps > 2 - .01 & samps < 2 + .01)
```

```
[1] 0.00379
```