# Project - Probability Course

Understanding Bayes Theorem

Haris Rasyidi | 2023

*Bayes theorem is a method to update our previous belief, based on the new evidence given to us*

Let's use the insurance dataset that is available to use, given the following distribution

Suppose our main objective today is to find how **" how many smokers are in the dataset?"**

*Ps: we will assume we already aware with conditional probability*

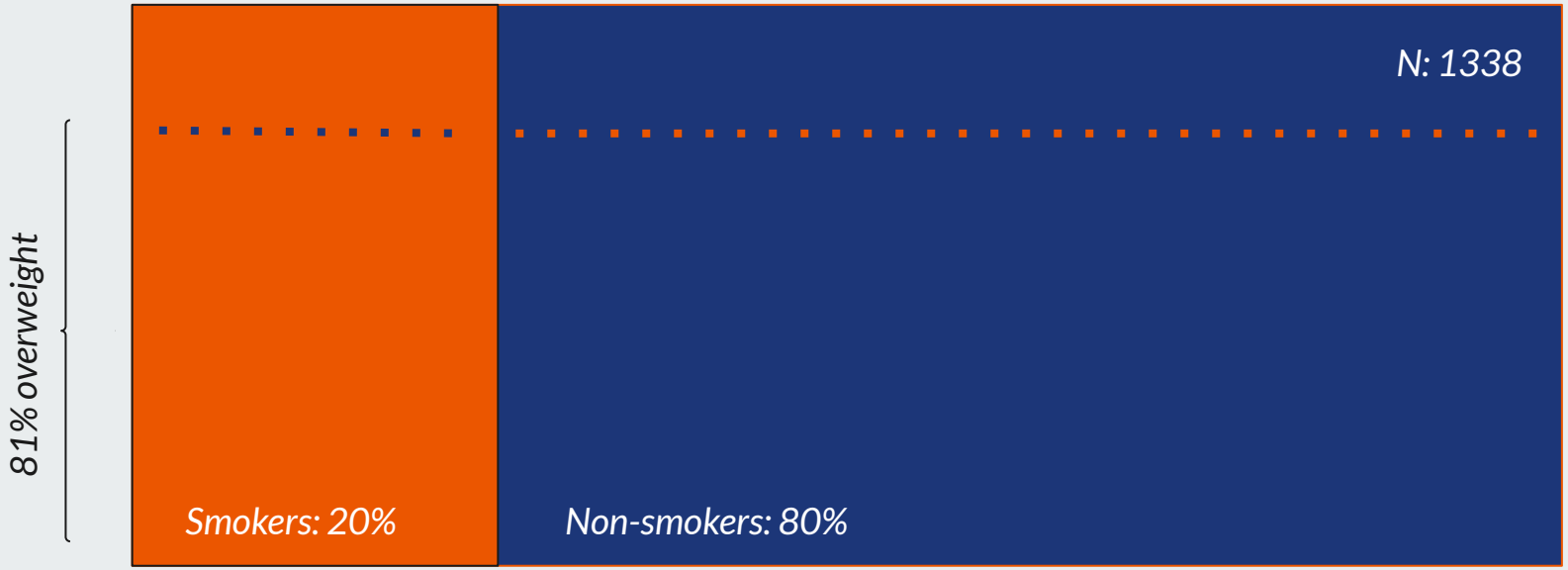
| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 0 | 19 | female | 27.900 | 0 | yes | southwest | 16884.92400 |
| 1 | 18 | male | 33.770 | 1 | no | southeast | 1725.55230 |
| 2 | 28 | male | 33.000 | 3 | no | southeast | 4449.46200 |
| 3 | 33 | male | 22.705 | 0 | no | northwest | 21984.47061 |
| 4 | 32 | male | 28.880 | 0 | no | northwest | 3866.85520 |

| | age | bmi | children | charges |
|---|---|---|---|---|
| count | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| mean | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| std | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| min | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| 25% | 27.000000 | 26.296250 | 0.000000 | 4740.287150 |
| 50% | 39.000000 | 30.400000 | 1.000000 | 9382.033000 |
| 75% | 51.000000 | 34.693750 | 2.000000 | 16639.912515 |
| max | 64.000000 | 53.130000 | 5.000000 | 63770.428010 |

*We calculate that the probability of* ***getting a smokers in the dataset is 20%***


*Meaning every 20 people we pull from the dataset, there is a chance 1 of them is a smoker*

*Now we are given a new information,* ***that at least 81% of the entire dataset contain individuals who are overweight***




*Knowing this new information, we want to update* ***"what the likelihood that the information is true (81% are overweight) given that my hypothesis is true (20% are smokers)***

*Which will look something like this*

*The result are* ***" the evidence will hold true  79.9% of the time, given the hypothesis is true"***

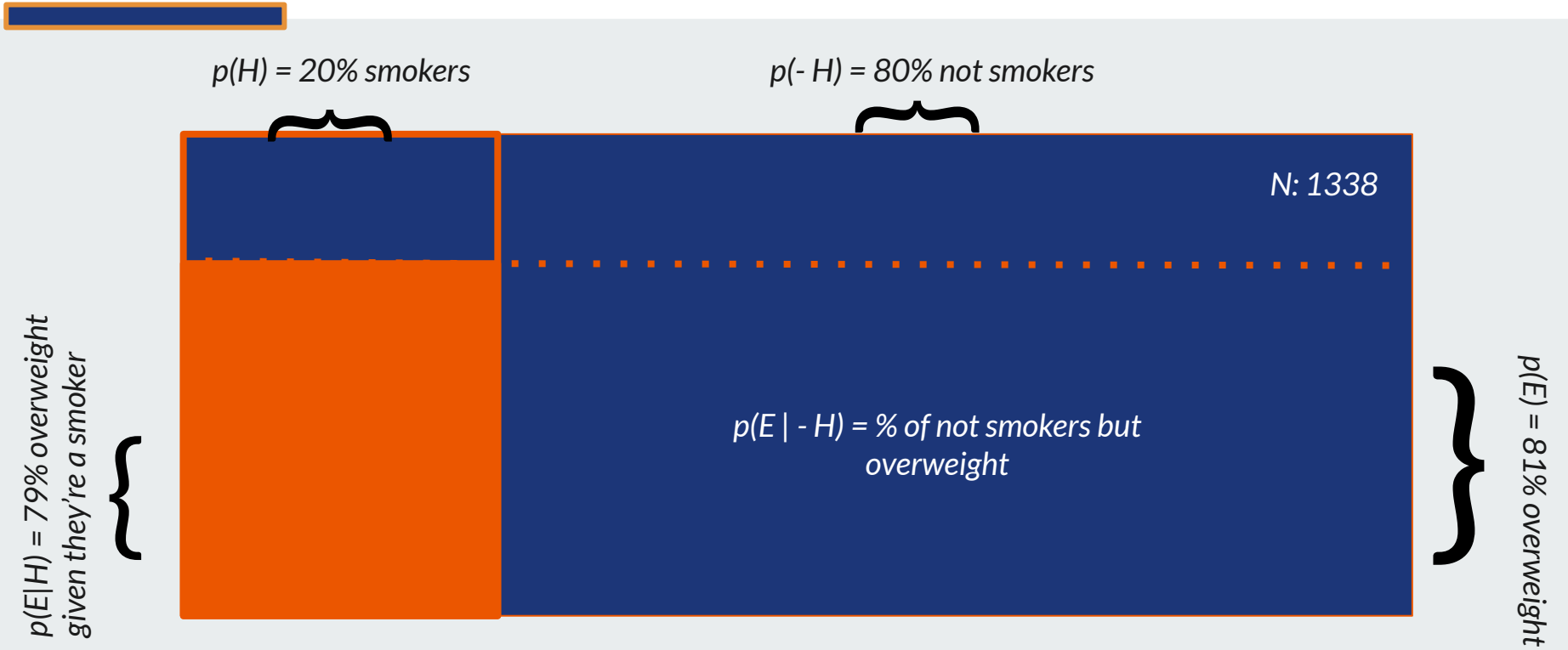
$$p\left(\ overweight \mid smoker\ \right) = \frac{n\left(\ overweight\ \cap\ smoker\ \right)}{n\left(\ smoker\ \right)}$$

```
# data filter to set up the overweight & smoker union
filter_overweight_smoker = (df["bmi"] > 24.9) & (df["smoker"] == "yes")
overweight_smoker = df[filter_overweight_smoker]["smoker"].count()

# counting data on smokers
smoker = df[df['smoker']=='yes']['smoker'].count()

# result
p_overweight_smoker = (overweight_smoker/smoker * 100).round(1)
print(p_overweight_smoker)
```

```
79.9
```

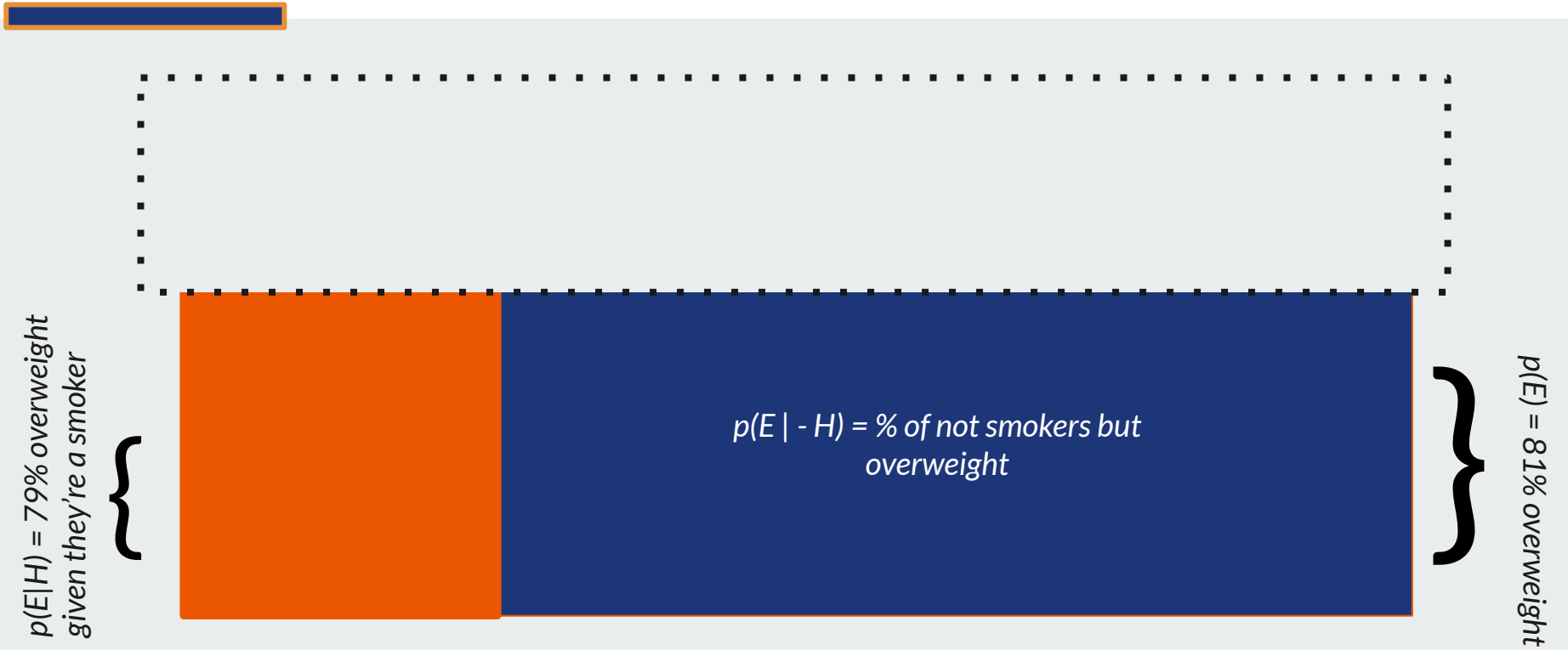
*Knowing this, we know gonna call the p ( evidence |  hypothesis) as a* ***likelihood***

*Now this is what our final geometry will look like*



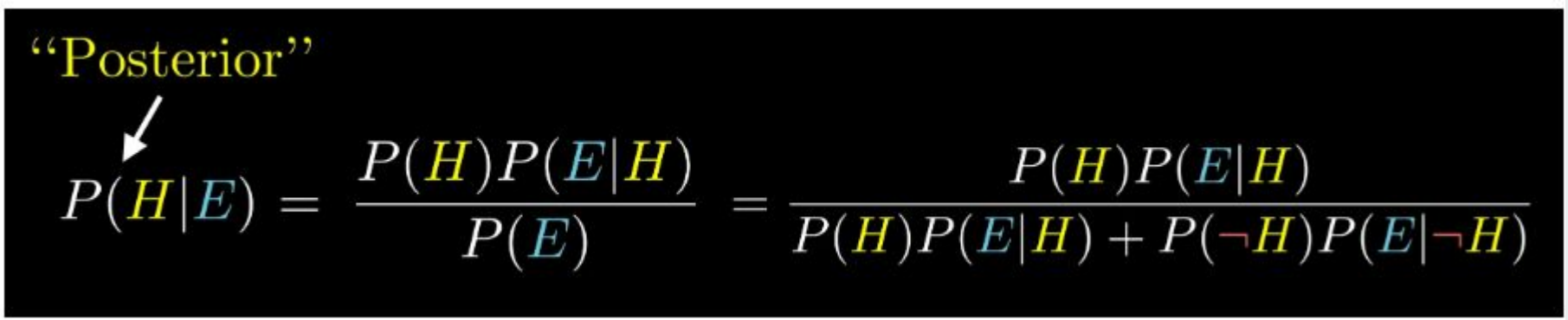

*Now we want to know,* ***the probability of our hypothesis is true given the new information, which means finding out p(H|E)***

*Visualized, it'll look something like this*




*Now we want to know,* ***the probability of our hypothesis is true given the new information, which means finding out p(H|E)***

*To update our hypothesis, we can use* ***bayes theorem***




*In this case, its easier to use the left formula since we have all the relevant data*

***p(H) = 20%***
***p(E|H) = 79.9 %***
***p(E) = 81%***

*Which we **get 19.7%***

$$p(\text{ smokers} \mid \text{overweight }) = \frac{p(\text{ smokers})\ p(\text{overweight} \mid \text{smokers})}{p(\text{ overweight })}$$

$$p\left(\text{ smokers} \mid \text{overweight}\right) = \frac{20\% \ . 79.9\%}{81\%}$$

$$p\left(\text{ smokers} \mid \text{overweight}\right) = 19.72\%$$

*Translated to common english, p(H|E) = 19.7% means that our our **original hypothesis of 20% chances of smokers** from the entire dataset can be **updated to 19.7%** chances of smoker **given the new evidence of overweight***