

Lab – Randomized Response

Gunnar Yonker

The randomized response technique is a novel way of collecting sensitive survey information, ensuring respondent privacy. By introducing randomness, we ensure that a respondent's true response remains obscured. This fosters honesty while maintaining confidentiality.

Complete Code:

```
import numpy as np
```

```
import pandas as pd
```

```
def flip_coin():
```

```
    s = np.random.rand()
```

```
    if(s <= 0.5):
```

```
        return "Head"
```

```
    else:
```

```
        return "Tail"
```

```
def flip_response(response):
```

```
    return "No" if response == "Yes" else "Yes"
```

```
def mechanism(true_response):
```

```
    first_flip = flip_coin()
```

```
    if first_flip == "Head":
```

```
        second_flip = flip_coin()
```

```
        return true_response if second_flip == "Head" else
```

```
flip_response(true_response)
```

```
    else:
```

```
        second_flip = flip_coin()
```

```
        return "Yes" if second_flip == "Head" else "No"
```

```
df = pd.read_csv('survey_dataset.csv', delimiter = ",")
```

```
for index, row in df.iterrows():  
    row['Response'] = mechanism(row['Response'])  
  
df.to_csv('survey_randomized.csv', sep=',', index=False)
```

Implementation Details:

Coin Flip Simulation: We define a function `flip_coin()` that simulates the flip of a fair coin, returning either head or tail.

```
def flip_coin():  
    s = np.random.rand()  
  
    if(s <= 0.5):  
        return "Head"  
  
    else:  
        return "Tail"
```

Response Flipping Function: The `flip_response()` function is designed to flip the given response. if the input is Yes it will return No and vice versa.

```
def flip_response(response):  
    return "No" if response == "Yes" else "Yes"
```

Mechanism Implementation: The `mechanism()` function, when given a true response, will execute the randomized response mechanism. Depending on the outcomes of two coin flips, the respondent's true answer might be preserved, flipped, or changed to a fixed Yes or No.

```
def mechanism(true_response):  
    first_flip = flip_coin()  
    if first_flip == "Head":  
        second_flip = flip_coin()
```

Lab – Randomized Response
Gunnar Yonker

```
        return true_response if second_flip == "Head" else  
flip_response(true_response)  
  
    else:  
  
        second_flip = flip_coin()  
  
        return "Yes" if second_flip == "Head" else "No"
```

Reading and Randomizing the Dataset: We use the pandas library to read the survey data from the CSV file, apply the randomized response mechanism to each respondent's true response, and save the randomized responses to a new CSV file.

```
df = pd.read_csv('survey_dataset.csv', delimiter = ",")  
  
for index, row in df.iterrows():  
  
    row['Response'] = mechanism(row['Response'])  
  
df.to_csv('survey_randomized.csv', sep=',', index=False)
```

First 20 Rows Before Running the Code:

```
Profession,Marital Status,Response  
Fashion Designer,Single,Yes  
Electrician,Married,No  
Composer,Married,Yes  
Accountant,Single,Yes  
Economist,Single,No  
Teacher,Single,Yes  
Engineer,Single,No  
Interpreter,Single,No  
Graphic Designer,Married,No  
Salesman,Single,Yes  
Florist,Married,No  
Singer,Married,No  
Meteorologist,Married,No  
Social Worker,Married,Yes  
Meteorologist,Single,No  
Pharmacist,Married,No  
Programmer,Married,Yes  
Chemist,Single,No  
Economist,Married,No  
Teacher,Single,Yes
```

First 20 Rows After Running the Code:

```
Profession,Marital Status,Response  
Fashion Designer,Single,Yes  
Electrician,Married,No  
Composer,Married,Yes  
Accountant,Single,Yes  
Economist,Single,No  
Teacher,Single,No  
Engineer,Single,Yes  
Interpreter,Single,Yes  
Graphic Designer,Married,Yes  
Salesman,Single,No  
Florist,Married,Yes  
Singer,Married,No  
Meteorologist,Married,No  
Social Worker,Married,No  
Meteorologist,Single,Yes  
Pharmacist,Married,Yes  
Programmer,Married,No  
Chemist,Single,No  
Economist,Married,No  
Teacher,Single,No
```

Observations and Insights:

Before employing the privacy mechanism, every response in the dataset directly reflected a respondent's true opinion. After employing the mechanism, the true nature of the response becomes obscured due to the random process, making it harder for malicious actors to discern individual opinions. The technique may introduce some level of inaccuracy in the direct interpretation of the results. However, it's possible to estimate the true prevalence of an opinion or behavior using the known probabilities of the randomization process. If the process is not known or the probabilities are not known, this becomes much more difficult. By using this technique, we aim to foster an environment where respondents feel safer to provide honest answers, increasing the credibility of the survey's findings.

Conclusion:

By integrating the randomized response technique, we can achieve a balance between gathering essential data and preserving respondent anonymity. Though this method introduces a layer of uncertainty to direct results, the overarching goal is to encourage truthfulness and enhance the credibility of the data.

Takeaways:

Techniques such as the randomized response show how privacy and data collection can coexist. This method emphasizes the importance of ensuring respondents' trust by protecting their privacy. With enhanced privacy, there's an inherent trade-off with data accuracy at the individual level. However, with sufficient sample sizes, we can still derive meaningful insights at the aggregate level. While this lab focuses on a binary Yes or No response, the principles behind randomized responses can be adapted for more complex scenarios, demonstrating the flexibility of the method. As data privacy becomes a more significant concern, integrating such techniques might become standard practice for sensitive topics, reinforcing the need for data scientists and researchers to understand and apply them correctly. One last takeaway that I think is important is that this method introduces randomness, which emphasizes the importance of having a sufficiently large sample size. With larger samples, the random errors introduced by individual respondents tend to cancel each other out, leading to more accurate aggregate results.