<https://chatgpt.com/share/105d1e08-f782-4f21-b832-295ec4d4e7e9>

Explore how to use probability distributions in descriptive statistics with a real-world dataset. We'll use the \*\*Telco Customer Churn\*\* dataset from IBM Watson, a common dataset used to predict customer churn in a telecom company.

### Dataset Overview

The Telco Customer Churn dataset includes information about:

- Customer account information (e.g., tenure, contract type, payment method)

- Services subscribed (e.g., internet service, online security)

- Customer demographics (e.g., gender, senior citizen status)

- Churn label (whether the customer churned or not)

### 1. Descriptive Statistics

#### Understanding Data with Probability Distributions

Probability distributions help us understand the properties of the data, such as central tendency, dispersion, and shape.

- \*\*Central Tendency\*\*: Mean, median, and mode.

- \*\*Dispersion\*\*: Variance, standard deviation, and range.

- \*\*Shape\*\*: Skewness and kurtosis.

#### Visualization

Visualizations like histograms, PDFs, and CDFs provide insights into the distribution of the data.

- \*\*Histogram\*\*: Shows the frequency of data points within specified ranges.

- \*\*PDF (Probability Density Function)\*\*: Describes the likelihood of a continuous random variable to take on a specific value.

- \*\*CDF (Cumulative Distribution Function)\*\*: Represents the probability that a continuous random variable will take a value less than or equal to a specific value.

### Why Create PDF, CDF, and PMF?

- \*\*PDF (Probability Density Function)\*\*: Useful for understanding the distribution and density of continuous data points.

- \*\*CDF (Cumulative Distribution Function)\*\*: Helps in understanding the probability of a variable falling within a range of values.

- \*\*PMF (Probability Mass Function)\*\*: Used for discrete data to represent the probability of each possible value.

### Real-World Example in ML

Let's use Python to demonstrate how to create these distributions using the Telco Customer Churn dataset.

#### Step-by-Step Implementation

1. \*\*Load the Dataset\*\*:

```python

import pandas as pd

# Load the dataset

url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/telco-customer-churn.csv'

df = pd.read\_csv(url)

```

2. \*\*Descriptive Statistics\*\*:

```python

# Summary statistics for numerical columns

summary\_stats = df.describe()

print(summary\_stats)

```

3. \*\*Histograms and PDFs\*\*:

```python

import matplotlib.pyplot as plt

import seaborn as sns

# Histogram of 'MonthlyCharges'

plt.figure(figsize=(10, 6))

sns.histplot(df['MonthlyCharges'], kde=True)

plt.title('Histogram and PDF of Monthly Charges')

plt.xlabel('Monthly Charges')

plt.ylabel('Frequency')

plt.show()

```

4. \*\*CDF\*\*:

```python

import numpy as np

# CDF of 'MonthlyCharges'

monthly\_charges = np.sort(df['MonthlyCharges'])

cdf = np.arange(len(monthly\_charges)) / float(len(monthly\_charges))

plt.figure(figsize=(10, 6))

plt.plot(monthly\_charges, cdf)

plt.title('CDF of Monthly Charges')

plt.xlabel('Monthly Charges')

plt.ylabel('CDF')

plt.show()

```

5. \*\*PMF\*\*:

```python

# PMF of 'Contract' type (categorical variable)

contract\_counts = df['Contract'].value\_counts(normalize=True)

plt.figure(figsize=(10, 6))

contract\_counts.plot(kind='bar')

plt.title('PMF of Contract Type')

plt.xlabel('Contract Type')

plt.ylabel('Probability')

plt.show()

```

### Interpretation and Use Cases

- \*\*PDF and Histogram\*\*:

- \*\*Understanding Distribution\*\*: The histogram and PDF of `MonthlyCharges` show the distribution of charges. A peak in the PDF indicates the most common charge range.

- \*\*Model Assumptions\*\*: Many machine learning models assume normality in the data. PDFs help in verifying this assumption.

- \*\*CDF\*\*:

- \*\*Percentiles and Probabilities\*\*: The CDF can help in finding percentiles and understanding probabilities. For instance, we can determine what percentage of customers have monthly charges below a certain value.

- \*\*Risk Management\*\*: In a churn prediction model, CDFs can help in setting thresholds for taking preventive actions.

- \*\*PMF\*\*:

- \*\*Category Analysis\*\*: The PMF of `Contract` type shows the probability distribution of different contract types. This helps in understanding the proportion of customers on different contract plans.

- \*\*Decision Making\*\*: Businesses can use this information to design retention strategies for specific contract types.

By using probability distributions, we gain deeper insights into the data, which helps in better modeling and decision-making in machine learning projects.