Recognizing the most common distributions

Prereq (To understand this lecture, you must have already mastered):

1. Probability basics
2. Expectation and Variance

Topics:

1. Continuous Uniform Distribution
2. Discrete Uniform Distribution
3. Exponential Distribution
4. Gaussian Distribution
5. Bernoulli Distribution
6. Categorical Distribution
7. Poisson Distribution
8. How to identify p(x) by recognizing the distribution
9. More histogram practice

Lecture Written by : Prof. Wu
@ chiehwu.com

We Previously learned that Probability Distributions are Really Useful

Once we have the probability distribution, we can

- Calculate the probability of events $p(x)$
- Calculate conditional probabilities $p(x|y) = \frac{p(x,y)}{p(y)}$
- Caclulate dependency with $p(x, y) = p(x)p(y)$.
- Caclulate Bayesian probabilities $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$.
- Make better life choices, etc, etc.

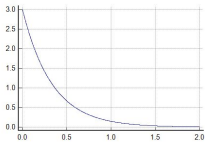
However in life, people don't just hand you a readily available distribution $p(x)$!!

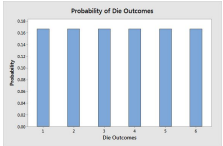
- Instead, you have to collect data and use it to create a model $p(x)$ that represents the data
  - The process of finding $p(x)$ is called "Paramteter Estimation"
  - The resulting model $p(x)$ is called a Generative Model. (The Gen AI today is based on this concept)
- A large portion of data science and machine learning is about modeling the true distribution of the data.

There are many ways to find the probability distribution,

- Histograms
- Recognizing the distribution
- Maximum Likelihood
- Kernel Density Estimation
- Method of Moment
- MAP
- Bayesian Parameter Estimation
- Gaussian Mixture Models
- Flow models
- Variational Auto-encoders
- Graphical Networks
- GANs

We also learned in the previous classes



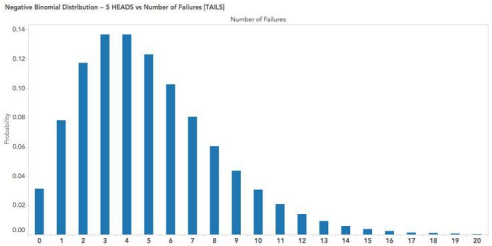


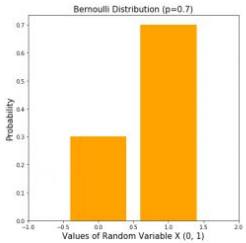
- that the easiest way to visualize data is with histograms.
- However, histograms is not a probability distribution $p(x)$, it is only a visualization tool.
- Today, we are going to learn the easiest way to convert data into and equation, i.e., data $\longrightarrow p(x)$.

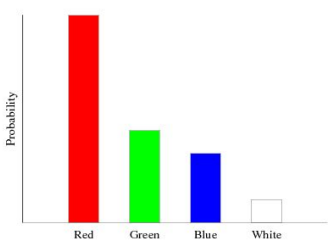
Statisticians have long realized that there exists a list of distributions that often occur in nature.

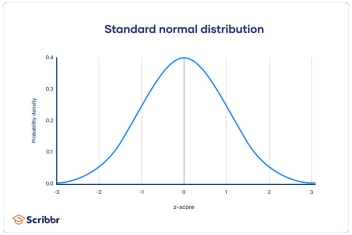
- They have since identified equations to model these natural phenomenons.
- If you recognize the appropriate situations to use these distributions, you can obtain $p(x)$ from data.
- Let's go over today the 6 most commonly seen distributions.

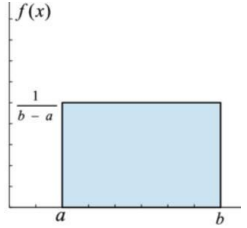
1. Bernoulli Distribution
2. Categorical Distribution
3. Uniform Distribution
4. Poisson Distribution
5. Exponential Distribution
6. Gaussian Distribution











When can we use Bernoulli Distributions?

Bernoulli Distributions describes events $\mathcal{X}$ with exactly 2 possible outcomes (1 or 0).

- Which basketball team will win the game today? (Team A = 1 or B = 0)
- Will it rain today (Yes = 1, or No = 0)
- Will the student pass or fail? (pass = 1 or fail = 0)

For example, let say you have gone to 40 good interviews (as 1s) and 60 bad interviews (as 0s).

- In this case, the probability of having a good interview is $p(x = 1) = 0.4$ and bad interview as $p(x = 0) = 0.6$.
- What equation $p(x)$ would yield this result?

Let's call the probability of 1 as θ, therefore, the probability of 0 must be $(1 - \theta)$.

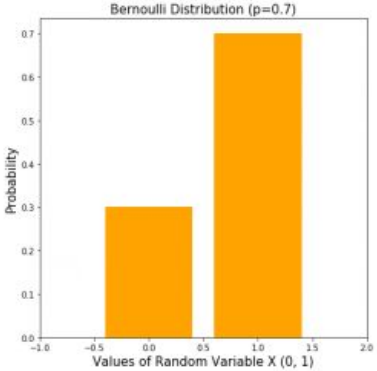
- If you recognize this situation as Bernoulli, then the equation is

$$\underbrace{p(x) = \theta^x(1-\theta)^{1-x}}_{\textbf{Equation of Bernoulli you need to know.}} \quad \Longrightarrow \quad p(x) = 0.4^x(0.6)^{1-x}.$$

- Notice that if you plug in the event 1 or 0, it gives you the correct probability.

$$p(1) = 0.4^1(0.6)^0 = 0.4 \quad \text{and} \quad p(0) = 0.4^0(0.6)^1 = 0.6.$$

Let's go from Data to Bernoulli Distributions?



Given a data of 1s and 0s

- You count the number of 1s and divide that by the total number.
- This would give you the probability of success θ.
- Once you know theta, you know the distribution

$$p(x) = \theta^x (1-\theta)^{1-x}.$$

- A Bernoulli distribution looks like the figure shown.

Try to find $p(x)$ given the following data

We ask 4000 couples at the end of 1 month of dating if they are still together or have broken up. The file can be loaded from

`Prob_of_breakup_within_1_month.csv`.

Identify $p(x)$ given this data.

Try to find $p(x)$ given the following data

We ask 4000 couples at the end of 1 month of dating if they are still together or have broken up. The file can be loaded from

`Prob_of_breakup_within_1_month.csv`.

Identify $p(x)$ given this data.

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
```

You know p(x) once you know θ

$$p(x) = \theta^x (1-\theta)^{1-x}$$

```
X = pd.read_csv("Prob_of_breakup_within_1_month.csv", header=None)
X = LabelEncoder().fit_transform(X[0])
θ = np.sum(X)/len(X)
print('θ = %.3f'%θ)
```

```
θ = 0.417
```

When can we use Categorical Distributions?

Categorical Distributions describes events X with exactly more than 2 possible outcomes.

- Which number will the die roll? (1,2,3,4,5,6)
- What's the weather today? (Sunny, raining, cold)
- Which candidate will win the election. (Candidates A, B, or C)

For example, let say you asked 100 people their favorite meal of the day: 40 people said breakfast, 10 people said lunch and 50 people said dinner. What is the probability distribution $p(x)$ that describes this dataset?

- In this case, we have 3 possible outcomes. Therefore, the input to $p(x)$ is a vector of 3 variables $x = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top$.
- In fact, you only have 3 possible input you can put into the probability

$$\underbrace{p\left(x = \begin{bmatrix}1\\0\\0\end{bmatrix}\right) = \theta_1 = 0.4}_{\textbf{Prob of breakfast}} \quad \text{and} \quad \underbrace{p\left(x = \begin{bmatrix}0\\1\\0\end{bmatrix}\right) = \theta_2 = 0.1}_{\textbf{Prob of lunch}} \quad \text{and} \quad \underbrace{p\left(x = \begin{bmatrix}0\\0\\1\end{bmatrix}\right) = \theta_3 = 0.5}_{\textbf{Prob of dinner}}$$

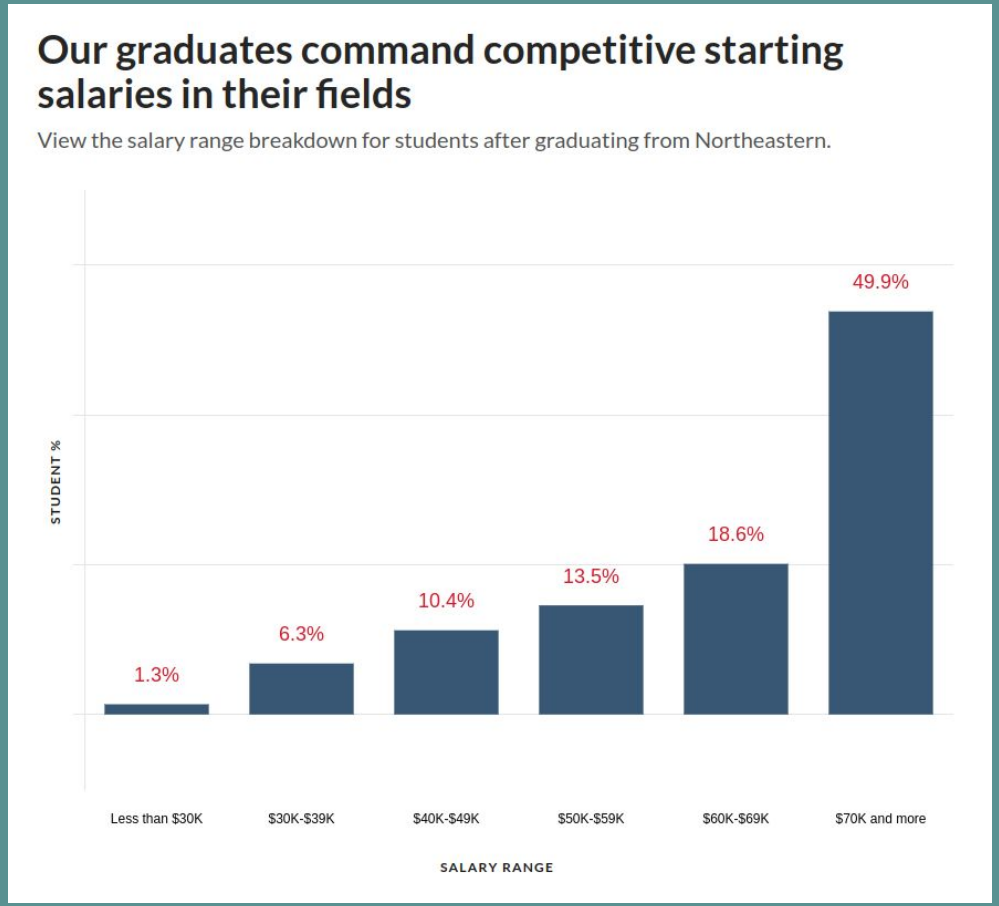
- In this case, the $p(x)$ for categorical distribution is

$$p(x) = \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3} = \prod_{i=1}^{3} \theta_i^{x_i} \quad \Longrightarrow \quad p(x) = 0.4^{x_1}0.1^{x_2}0.5^{x_3}.$$

- Notice how if we plug the vector for breakfast and lunch, we would get the associated probability

$$\underbrace{p\left(x = \begin{bmatrix}1\\0\\0\end{bmatrix}\right) = \theta_1^1\theta_2^0\theta_3^0 = \theta_1 = 0.4}_{\textbf{Prob of breakfast}} \quad \text{and} \quad \underbrace{p\left(x = \begin{bmatrix}0\\1\\0\end{bmatrix}\right) = \theta_1^0\theta_2^1\theta_3^0 = \theta_2 = 0.1}_{\textbf{Prob of lunch}}$$

Graduating Salary of Northeastern Students

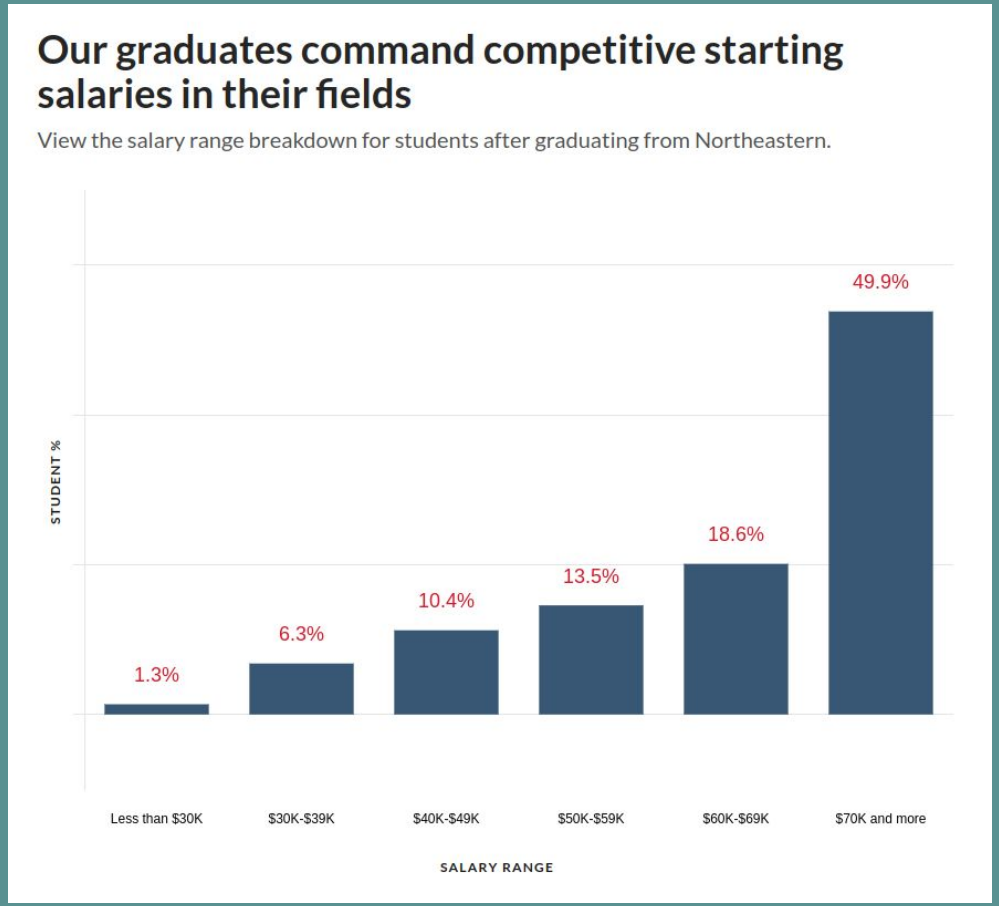


Source

The Plot consists of the starting salaries of Northeastern Graduate from 2016 to 2022.

What is an appropriate p(x) that describes this chart?

Graduating Salary of Northeastern Students



Source

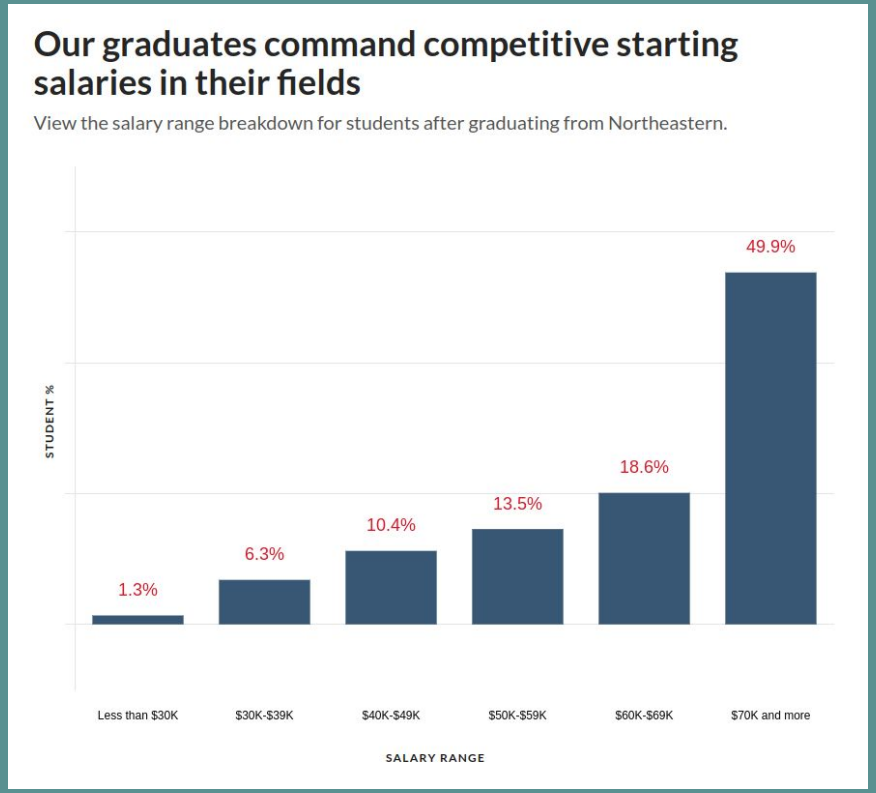
The Plot consists of the starting salaries of Northeastern Graduate from 2016 to 2022.

What is an appropriate p(x) that describes this chart?

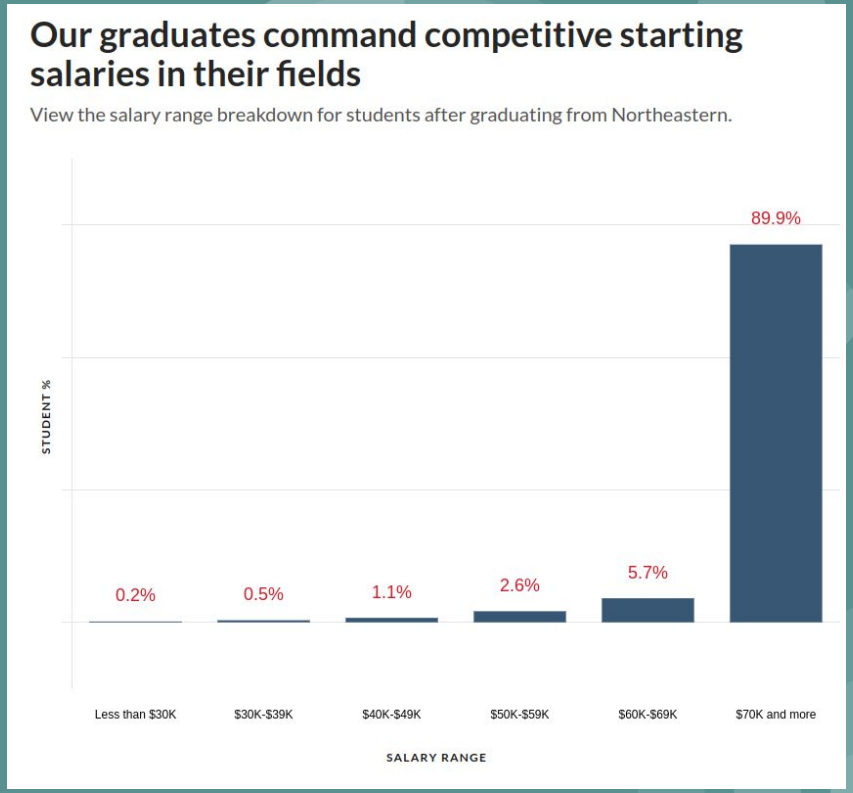
Categorical Distribution

$$p(x) = 0.013^{x_1} 0.063^{x_2} 0.104^{x_3} 0.135^{x_4} 0.186^{x_5} 0.499^{x_6}$$

Graduating Salary of Northeastern Students



Graduating Salary of Khoury Students



When to use Uniform Distribution (A discrete distribution)

The Uniform Distribution describes events that have multiple outcomes that are all equally likely.

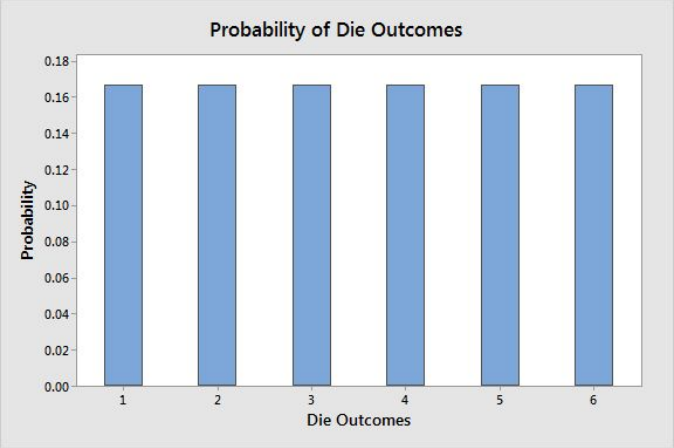
- **Which number will a fair die land on?** (1, 2, 3, 4, 5 or 6, p(x) = ⅙)
- **Which card suit will you get?** (spade, diamond, heart, clubs, p(x) = ¼)

Example: What is the probability distribution that describes the probability you would get a certain number from a deck of cards?

Equation:

$$p(x) = \frac{1}{d} \quad \text{where } d = \text{ number of categories.}$$

Notice there are no θ terms. You automatically know the distribution if you know the number of possible outcomes.



When to use Uniform Distribution (A Continuous distribution)

The Uniform Distribution can be both discrete and continuous (continuous case)

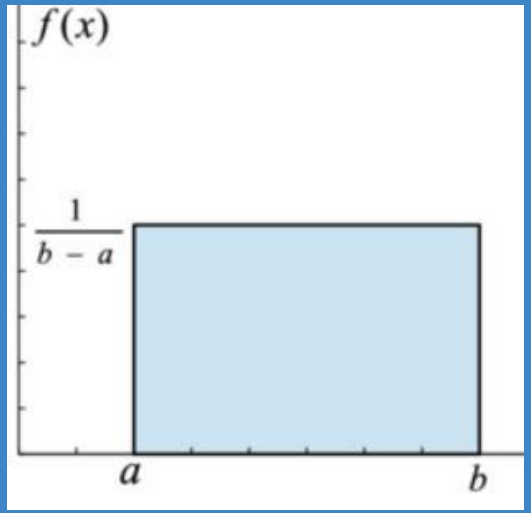
- **Where would the MBTA Train break down?** (equally likely anywhere)
- **What angle would a falling pin point towards?** (equally likely 360 degrees)

Example: If you look at your watch at a random time, what is the probability that the second hand would be between 0 and 10?

Equation:

$$p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{everywhere else} \end{cases}$$

Notice there are no θ terms. You automatically know the distribution if you know b and a.



When to use Poisson Distribution (A discrete distribution)

The Poisson Distribution describes that becomes more and more rare as x increase

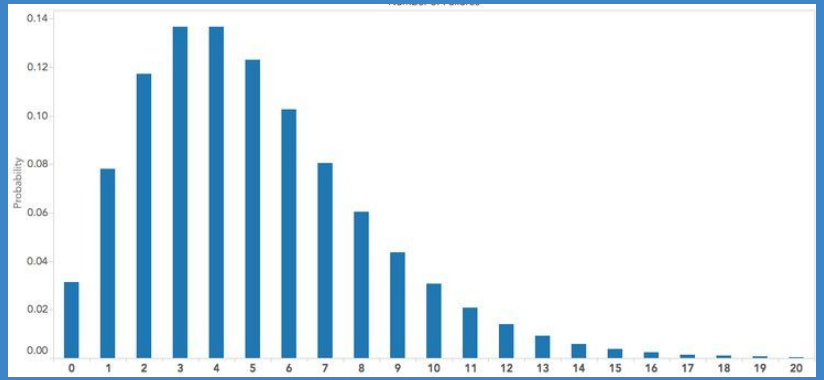
- **What's the probability that n number of chips would be defective?** (1, 2, 3, 4, 5, ...)
- **Distribution that describes the number of people waiting in line at the bank.**
- **The number of deaths a year from shark.**

Example: If you on average you meet 4 new people in a month. What's the probability you will meet 20 new people this month?

Equation:

$$p(x) = \frac{\theta^x e^{-\theta}}{x!} \quad \text{where} \quad x = \{1, 2, 3, ...\}$$

The pattern is obvious from previous examples, if you know θ, you know the entire distribution p(x).



Notice that the probability is highest at 4, but decreases quickly where 20 is unlikely

The expectation for poisson distribution.

$$\mathbb{E}[X] = \sum_{0}^{\infty} x\frac{\theta^x e^{-\theta}}{x!} = e^{-\theta}\sum_{0}^{\infty} x\frac{\theta^x}{x!} = e^{-\theta}\sum_{0}^{\infty}\frac{\theta^x}{(x-1)!} = e^{-\theta}\theta\underbrace{\sum_{0}^{\infty}\frac{\theta^{x-1}}{(x-1)!}}_{\text{famous sum } = e^{\theta}} = e^{-\theta}\theta e^{\theta} = \theta$$

The expected number of occurrence is θ. So if we know the expectation, we know the distribution.

Solve this problem :

If you on average you meet 4 new people in a month. What's the probability you will meet 20 new people this month?

Example: If you on average you meet 4 new people in a month. What's the probability you will meet 20 new people this month?

$$p(x|\theta = 4) = \frac{4^x e^{-4}}{x!}$$

$$p(x = 20|\theta = 4) = \frac{4^{20} e^{-4}}{20!} = 0.0000000082775$$

When to use Exponential Distribution (A continuous distribution)

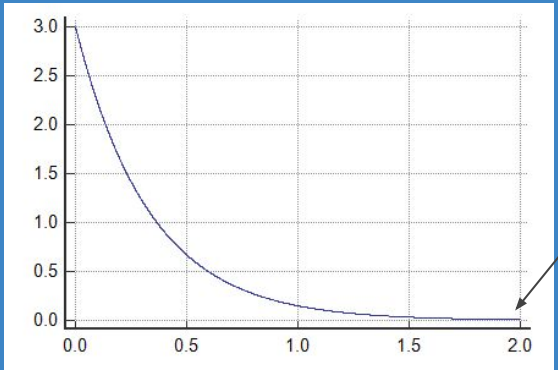
Similar to Poisson Distribution (but continuous) describes that becomes more and more rare as x increase

- **The amount of time a postal worker spend on a customer.**
- **The amount of time someone buys a concert ticket before the concert.**
- **The amount of time a phone is not dropped.**

Example: If on average people first drop their phone in 0.5 year, what's the probability that they never drop the phone within the 2 year plan?

Equation: $p(x) = \theta e^{-\theta x} \quad \text{where} \quad x \geq 0$

The pattern is obvious from previous examples, if you know θ, you know the entire distribution p(x).



Notice that the exponential distribution (continuous) looks very similar to the poisson distribution (discrete). Make sure you don't confuse the 2.

Based on this chart, we see that it's almost impossible for someone to never drop their phone within the 2 year plan.

How do we know the θ value for Exponential Distribution?

The θ value for exponential distribution is

$$\theta = \frac{1}{\mathbb{E}[X]} \quad \text{where} \quad p(X) = \theta e^{-\theta x}.$$

How do we know this? We just calculate the the expected value for p(x) and we will get

$$\mathbb{E}[X] = \int_0^\infty x\, p(x)\, dx = \underbrace{\int_0^\infty x\, \left(\theta e^{-\theta x}\right)\, dx}_{\textbf{We didn't learn to solve this integral}} = \frac{1}{\theta}.$$

But the point from the integral is that $\mathbb{E}[X] = \frac{1}{\theta}$, giving us the conclusion that

$$\theta = \frac{1}{\mathbb{E}[X]}.$$

Once we identified θ, we have $p(x)$ and we can use it to find all sorts of probabilities. For example is the average of an exponential event is $1/2$, then $\theta = 2$ and

$$p(1 \leq x \leq \infty) = \int_1^\infty 2e^{-2x}\, dx.$$

Practice question

You work at Google, the average load time after search is 0.5 seconds. Your boss comes to you and say

“Hey machine learning specialist, under normal operations, what’s the probability that a user would experience longer than 2 seconds? (unacceptable)”

1. Identify the appropriate p(x) that models this problem.
2. Use python to perform the integration to find the probability of experiencing longer than 2 second delay.

3. If google has 1 billion users at any given time, how many users will experience a lag time of higher than 2 seconds?

Answer

Exponential Distribution

- The equation for exponential distribution is

$$p(x) = \theta\, e^{-\theta x}$$

- to know the exact equation, we just need to find θ which is $\frac{1}{Avg(x)}$.
- Since the average wait time is 0.5 sec, then $\theta = 2$, giving us the equation $p(x) = 2e^{-2x}$. So we just need to solve the integral

$$\int_2^{\infty} 2e^{-2x}\, dx$$

```
from scipy.integrate import quad
import numpy as np
from numpy import exp
from numpy import inf as ∞
```

```
def integrand(x):
    return 2*exp(-2*x)
```

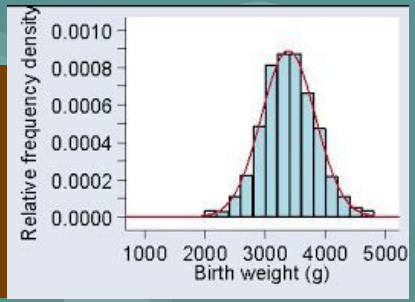
```
(result, error) = quad(integrand, 2, ∞  )
print('Probability: ', result)
print('Num of people experience longer than 2 seconds: %d'%(1000000000*result))
```

```
Probability:  0.018315638888710205
Num of people experience longer than 2 seconds: 18315638
```

When to use Gaussian Distribution (A continuous distribution)

The Normal/Gaussian Distribution describes any phenomenon with a bell curve

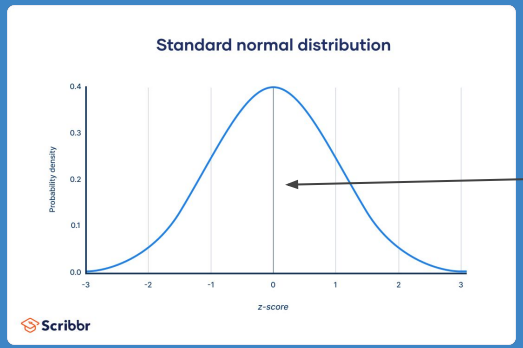
- **The IQ score of a population.**
- **The height of a population.**
- **Birth weight.**



Example: Given the retirement ages of NBA basketball players, what is the appropriate probability distribution?

Equation: $$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Here, we have 2 parameters to find out μ and σ



The normal/Gaussian distribution is one of the most commonly seen and used distribution.

So many aspects of our lives reflects this chart. Most of us are just average, and few are truly exceptional.

The expectation for Gaussian distribution.

The expectation has a long and complicated integral, let's just skip it and get to the point.

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

$$Var[x] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

Now that you know that the expectation is μ, we can approximate it with the average value

$$\mathbb{E}[x] \approx \frac{1}{n}\sum_{i}^{n} x_i = \bar{x}$$

Similarly, you can obtain the variance via approximation with the empirical variance formulate we learned

$$Var[x] \approx \frac{1}{n}\sum_{i}^{n} (x_i - \bar{x})^2$$

Once you know $\bar{x}$ and σ^2, you also know the equation for the Gaussian distribution.

Practice the Concept

If an average Northeastern Student graduated high school with a GPA of 4.1 and a standard deviation of 0.3, then

1. What is the $p(x)$ that describes this data using a Gaussian Distribution.
2. Take your own High School GPA and calculate what percentage of the NEU population had a lower GPA (use cdf).

Practice question

You are trying to model how much time you spend texting. You found that you receive on average 10 texts an hour. What's the probability that you will miss 8 texts the next hour where you have your phone turned off.

Practice question

You are trying to model how much time you spend texting. You found that you receive on average 10 texts an hour. What's the probability that you will miss 8 texts the next hour where you have your phone turned off.

$$p(x) = \frac{\theta^x e^{-\theta}}{x!} \quad \text{where} \quad x = \{1, 2, 3, ...\}$$

$$p(x = 8|\theta = 10) = \frac{10^8 e^{-10}}{8!}$$

Python tools once you have $p(x)$

- Once you identify the common distributions $p(x)$ from data, Python provides a bunch of automatic tools.
- We have previously learned the Gaussian distribution tools

```
from scipy.stats import norm

norm.pdf(x_value, μ, σ)
norm.cdf(x_value, μ, σ)
norm.ppf(area, μ, σ)
```

- We have other distributions as well (exponential)

```
from scipy.stats import expon

expon.pdf(x_value, start_x, μ)
expon.cdf(x_value, start_x, μ)
expon.ppf(area, μ, σ)
```

- For a list of all distributions see

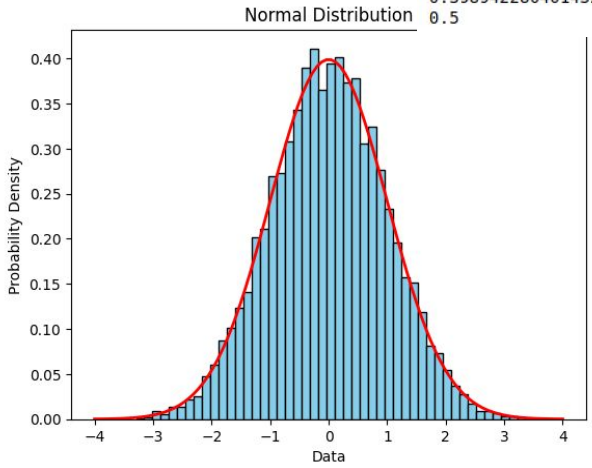
https://docs.scipy.org/doc/scipy/reference/stats.html

```python
#!/usr/bin/env python
#
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
```

```python
mean = 0  # Mean of the distribution
std_dev = 1  # Standard deviation of the distribution
size = 10000  # Number of data points to generate
#
data = np.random.normal(mean, std_dev, size)
#
# Plot the histogram of the data
plt.hist(data, bins=50, density=True, color='skyblue', edgecolor='black')
#
# Plot the probability density function (PDF) of the normal distribution
x = np.linspace(mean - 4 * std_dev, mean + 4 * std_dev, 100)
y = 1 / (std_dev * np.sqrt(2 * np.pi)) * np.exp(-(x - mean)**2 / (2 * std_dev**2))
plt.plot(x, y, color='red', linewidth=2)
#
# Set plot title and labels
plt.title('Normal Distribution')
plt.xlabel('Data')
plt.ylabel('Probability Density')
#
# Display the plot
plt.show()
#
```

```python
v1 = norm.pdf(0)
print(v1)
#
#——»you can also obtain the inte
v2 = norm.cdf(0)
print(v2)
#
```

```
0.3989422804014327
0.5
```



Python tools once you have $p(x)$

- Once you identify the common distributions $p(x)$ from data, Python provides a bunch of automatic tools.
- We have previously learned the Gaussian distribution tools

```
from scipy.stats import norm

norm.pdf(x_value, μ, σ)
norm.cdf(x_value, μ, σ)
norm.ppf(area, μ, σ)
```

- We have other distributions as well (exponential)

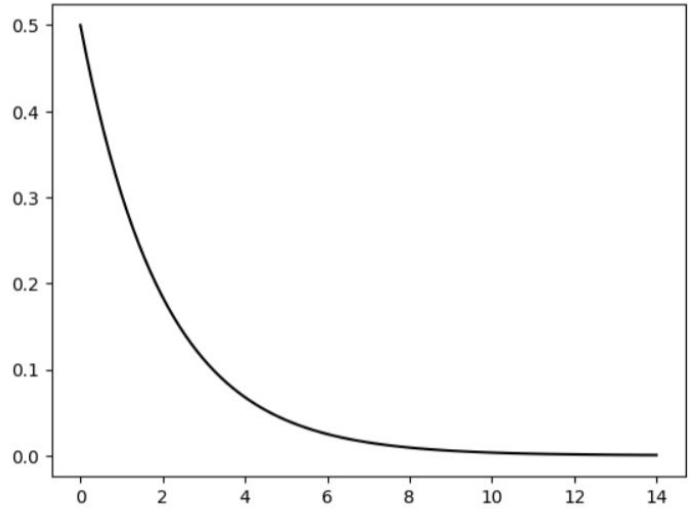
```
from scipy.stats import expon

expon.pdf(x_value, start_x, μ)
expon.cdf(x_value, start_x, μ)
expon.ppf(area, μ, σ)
```

- For a list of all distributions see

```
https://docs.scipy.org/doc/scipy/reference/stats.html
```

```
import numpy as np
from numpy import linspace
from scipy.stats import expon
from matplotlib import pyplot as plt
```

```
x = linspace(0,14,100)
y = expon.pdf(x, 0, 2) # p(x) starting at 0, μ=2
plt.plot(x, y, 'k-') # draw it
plt.show()
```



You can find the area with cdf

```
print(expon.cdf(3, 0, 2))
```

```
0.7768698398515702
```

You can also find the inverse cdf

```
print(expon.ppf(0.2, 0, 2))
```

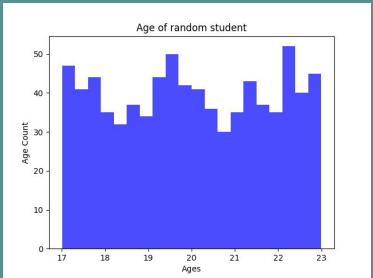
```
0.44628710262841953
```

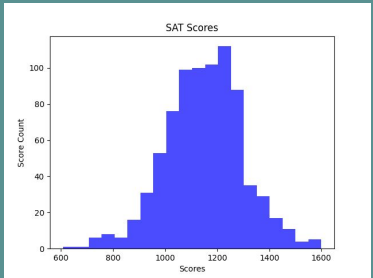
Practice Practice Practice

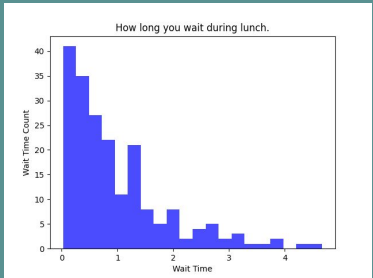
Go to the course website and find the 3 csv files : SAT.csv, Lunch_wait_time.csv, Student_age.csv

1. For each file, use the data to approximate the pdf p(x).
2. Once you have the pdf, plot the probability distribution right over the histogram and show that your function is a good approximation.
3. For the SAT scores
   a. Find the p(x < 1000) by counting the number of students score less than 1000 and divide by total.
   b. This time, use the python cdf function to find the probability p(x < 1000), and compare the results. Should they be the same?
   c. Use ppf to find the score that's higher than 80% of the SAT population.

4. Once you know the wait time p(x)
   a. use the the python cdf function to find the probability that someone waits longer than 3s.
   b. Is 90% of the population currently waiting less than 2s? (use ppf)







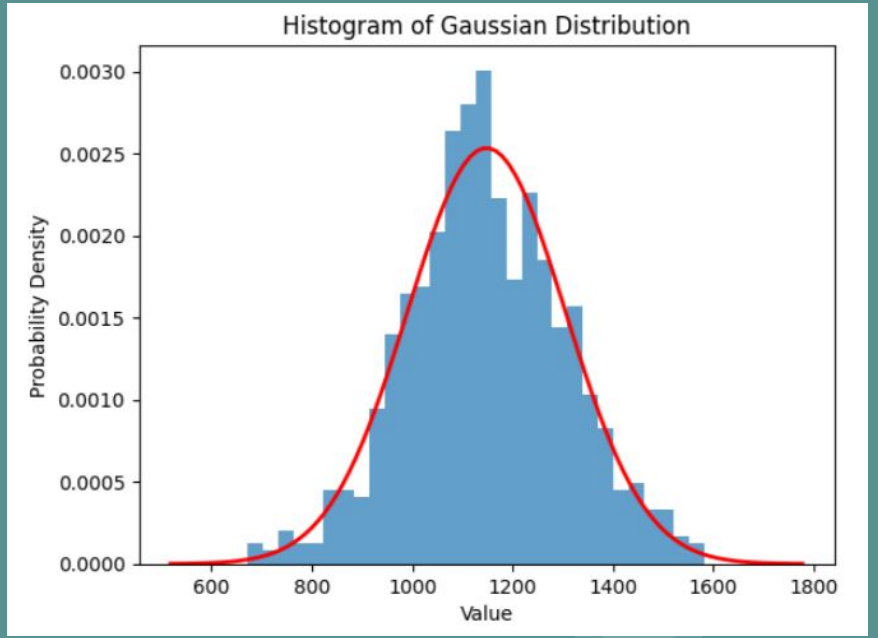
```
import numpy as np
import matplotlib.pyplot as plt
from numpy import genfromtxt
#
```

```
samples = genfromtxt('SAT.csv', delimiter=',')
n = len(samples)
#
μ = np.mean(samples)
σ = np.std(samples)
#
print('mean: ', μ)
print('std: ', σ)
#
```

```
mean:  1147.82875
std:  157.41561842281564
```

Plot histogram

```
plt.hist(samples, bins=30, density=True, alpha=0.7)
#
# Plot the Gaussian distribution curve for comparison
x = np.linspace(μ-4*σ, μ+4*σ, 100)
y = (1 / (σ * np.sqrt(2 * np.pi))) * np.exp(-0.5 * ((x - μ) / σ) ** 2)
plt.plot(x, y, color='red', linewidth=2)
#
# Set plot labels and title
plt.xlabel('Value')
plt.ylabel('Probability Density')
plt.title('Histogram of Gaussian Distribution')
#
# Show the plot
plt.show()
```



End of the day...

Remember to submit your In-class work

Drawn by AI @ https://lexica.art/