MET CS 688 Statistics & Probability II

Leila Ghaedi – Spring 2024

Creator: cemagraphics | Credit: Getty Images/iStockphoto

Scraping Instagram

Scraping Instagram

- It’s legal to access public data.
- You must limit the number of requests you are sending in time, otherwise your IP gets restricted.
- If your code worked earlier but not working now you may have reached the API rate limit.
- You need an Instagram credential (after one or two tries without logging in, you will get an error)

```
$ pip3 install instaloader
```

Scraping Instagram

- Steps:
- Import instaloader
- Define the bot
- Choose an account
- Read the content of your desired account and store it

Scraping Instagram

In [1]:
```
#!pip install instaloader
import instaloader
from selenium import webdriver
import time
from datetime import datetime
from itertools import dropwhile, takewhile
from instaloader.exceptions import TwoFactorAuthRequiredException
```

In [2]:
```
# Create an instance of Instaloader class
# Make sure in any loop you are defining there
# is a wait time between consecutive requests
# so you are getting banned
# we are accessing only public data here
bot = instaloader.Instaloader()

#bot.login('USERNAME', 'PASSWORD')

#try:
#    bot.login('USERNAME', 'PASSWORD')
#except TwoFactorAuthRequiredException:
#    bot.two_factor_login(111111)
```

Scraping Instagram

In [3]:
```
# Load a profile from an Instagram handle
# examples bbcnews, tombrady, cnn
profile = instaloader.Profile.from_username(bot.context, "tombrady")
```

In [4]:
```
print(type(profile))
```

```
<class 'instaloader.structures.Profile'>
```

In [5]:
```
print("Username: ", profile.username)
```

```
Username:  tombrady
```

In [6]:
```
print("User ID: ", profile.userid)
```

```
User ID:  1665557140
```

In [7]:
```
print("Number of Posts: ", profile.mediacount)
```

```
Number of Posts:  508
```

Scraping Instagram

```
In [8]:  1 print("Followers: ", profile.followers)
Followers:  15099421

In [9]:  1 print("Following: ", profile.followees)
Following:  427

In [10]:  1 print("Bio: ", profile.biography)
Bio:

In [11]:  1 print("Website: ",profile.external_url)
Website:  http://youtu.be/PZuNTi3nrBs
```

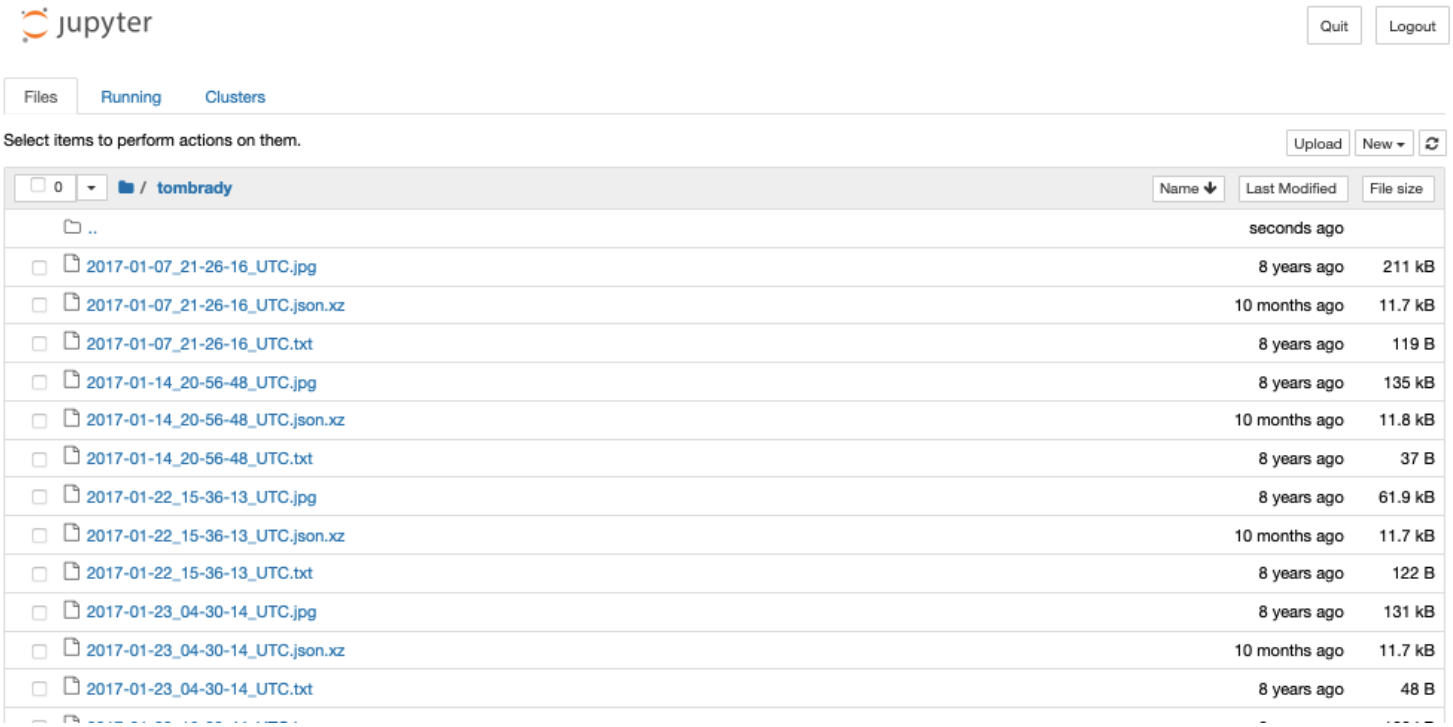
Scraping Instagram

In [12]:

```
posts= profile.get_posts()


SINCE = datetime(2024, 9, 1)
UNTIL = datetime(2024, 9, 23)

for post in posts:
    postdate = post.date
    if (postdate > SINCE and postdate < UNTIL) :
        print(postdate)
        bot.download_post(post, target="tombrady")
        time.sleep(3) # Sleep for 3 seconds
```

```
2024-09-17 18:39:32
tombrady/2024-09-17_18-39-32_UTC_1.jpg tombrady/2024-09-17_18-39-32_UTC_2.jpg tombrady/2024-09-17_18-39-32_UTC_3.j
pg tombrady/2024-09-17_18-39-32_UTC_4.jpg tombrady/2024-09-17_18-39-32_UTC_5.jpg tombrady/2024-09-17_18-39-32_UTC_
6.jpg tombrady/2024-09-17_18-39-32_UTC_7.jpg [What a great night to remembe…] json
2024-09-18 21:59:26
tombrady/2024-09-18_21-59-26_UTC.jpg [Congratulations to @tombrady'…] tombrady/2024-09-18_21-59-26_UTC.mp4 json
2024-09-12 23:00:05
tombrady/2024-09-12_23-00-05_UTC.jpg [I love when my friends come o…] tombrady/2024-09-12_23-00-05_UTC.mp4 json
2024-09-12 20:10:18
tombrady/2024-09-12_20-10-18_UTC.jpg [#Tostitos_Partner WAIT! don't…] tombrady/2024-09-12_20-10-18_UTC.mp4 json
2024-09-10 15:40:53
tombrady/2024-09-10_15-40-53_UTC.jpg [Can't believe @tombrady hit t…] tombrady/2024-09-10_15-40-53_UTC.mp4 json
2024-09-07 13:58:10
tombrady/2024-09-07_13-58-10_UTC.jpg [Time to get back to work 💪 To…] tombrady/2024-09-07_13-58-10_UTC.mp4 json
2024-09-05 20:39:24
tombrady/2024-09-05_20-39-24_UTC.jpg [Get ready for the season with…] tombrady/2024-09-05_20-39-24_UTC.mp4 json
2024-09-05 17:21:23
tombrady/2024-09-05_17-21-23_UTC.jpg [Welcome to the 21st century @…] tombrady/2024-09-05_17-21-23_UTC.mp4 json
```

Scraping Instagram



Selenium WebDriver

- The selenium package is used to automate web browser interaction from Python.
- Several browsers/drivers are supported (Firefox, Chrome, Internet Explorer), as well as the Remote protocol.
- https://pypi.org/project/selenium/

Probability Topics

Probability Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity.

An experiment is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a chance experiment. Flipping one fair coin twice is an example of an experiment.

The result of an experiment is called an outcome.

The sample space of an experiment is the set of all possible outcomes. There are 3 ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram.

The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, S = {H, T} where H = heads and T = tails are the outcomes.

The Law of Large Numbers

- In probability experiments, as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability.

Probability

$$P(\text{of desired event}) = \frac{\text{number of desired event}}{\text{total number of all possible events}}$$

Complement of an event: The complement of event A, denoted by A', consists of all outcomes that are not in A.

Events are mutually exclusive or disjoint if they cannot occur simultaneously.

$$P(A) = 1 - \mathbf{P(A')}$$

Joint and Conditional Probability

$P(A \text{ or } B) = P(A \cup B) =$ P (A **occurs or** B **occurs or both occur**)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \text{ and } B) = P(A \cap B) = P(A$ **occurs and** B **occurs too**)

Koglmorov definition:

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

Bayes definition:

$$P(A \mid B) = \frac{P(B \mid A) \,.\, P(A)}{P(B)}$$

Probability Example 1

- The sample space S is the whole numbers starting at one and less than 20.
- Let event A = the even numbers and event B = numbers greater than 13.
- Calculate P(A), P(B), P(A AND B), P(A OR B), P(A'), P(A) + P(A'), P(A|B), P(B|A)

  - Solution:

S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}

A = {2, 4, 6, 8, 10, 12, 14, 16, 18}

B = {14, 15, 16, 17, 18, 19}

P(A) = 9/19 = 0.47

P(B) = 6/19 = 0.31

Probability Example 1

A AND B = {14,16,18}

A OR B = {2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19}

P(A AND B) = 3/19

P(A OR B) = 12/19

A′ = {1, 3, 5, 7, 9, 11, 13, 15, 17, 19}

P(A′) = 10/19

P(A) + P(A′) = 1 (9 /19 + 10 /19 = 1)

P(A|B) = P(A AND B) P(B) = 3/6

P(B|A) = P(A AND B) P(A) = 3/9

Independent Events

- The occurrence of one event has no effect on the probability of the occurrence of another event.
- Events A and B are independent if one of the following is true:
  - 1. $P(A|B) = P(A)$
  - 2. $P(B|A) = P(B)$
  - 3. $P(A \text{ AND } B) = P(A)P(B)$

Mutually Exclusive Events

- Two events are mutually exclusive if the probability that they both happen at the same time is zero.
- If events A and B are mutually exclusive, then P(A AND B)=0

Sampling With Replacement
VS
Sampling Without Replacement

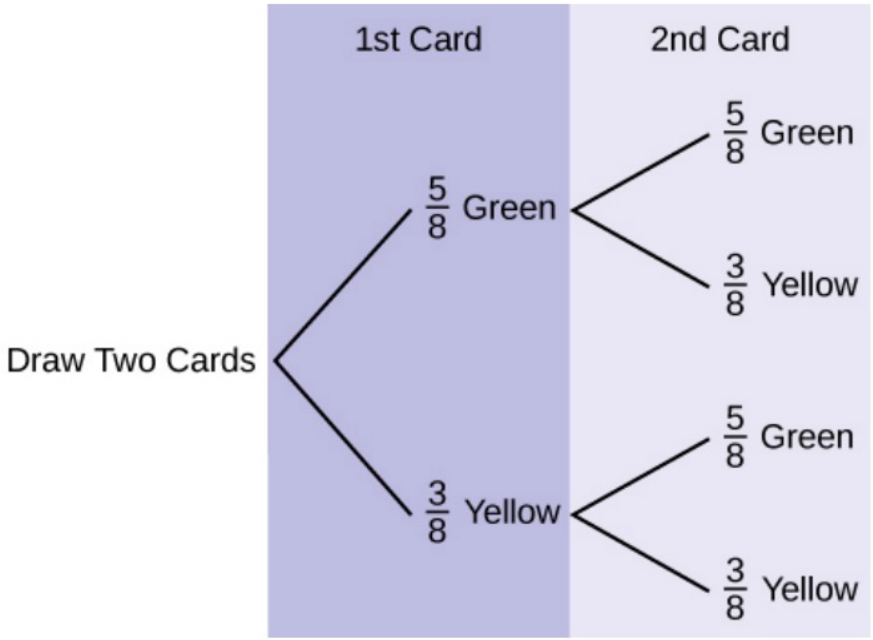
- If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.
- When sampling is done without replacement, each member of a population may be chosen only once.

Probability Example 2

- Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled. Suppose that you randomly draw two cards, one at a time, with replacement.
- Let G1 = first card is green Let G2 = second card is green
  a) Draw a tree diagram of the situation.
  b) Find P(G1 AND G2).
  c) Find P(at least one green).
  d) Find P(G2|G1).
  e) Are G2 and G1 independent events? Explain why or why not.

Probability Example 2

a) Here is the tree diagram:



Probability Example 2

b) P(GG) = (5/8) * (5/8)= 25/64

c) P(at least one green) = P(GG) + P(GY) + P(YG)

= 5/8*5/8 + 5/8*3/8 + 3/8 * 5/8

= (25+15+15)/64= 55/64

d) P(G|G)= 5/8

e) Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Probability Distribution Function (PDF)

- A discrete probability distribution function has two characteristics:
  - Each probability is between zero and one.
  - The sum of the probabilities is one.

Probability Distribution Function (PDF)?

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, x = 0, 1, 2, 3, 4, 5. P(x) = the probability that X takes on value x.

Why is this a discrete probability distribution function?

X	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$

Expected Value

- The expected value is often referred to as the "long-term" mean.
- This means that over the long term of doing an experiment over and over, you would expect this average.
- Expected Value= $\sum \quad x \ * P(x)$

Probability Example 3

- Suppose you play a game with a biased coin. You play each game by tossing the coin once. P(heads) = 2/3 and P(tails) = 1/3. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, do you expect to profit or lose money?

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

Probability Example 3

- Expected Value = 10/3 + - 12/3 = -2/3
- If you play this game, on average, each time you lose 67 cents. According to negative EV, this is not profitable game for you.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

PDF - PMF - CDF

- Probability Density Function (PDF) shows how **dense** is the probability at **each data point** in a **continuous vector.**

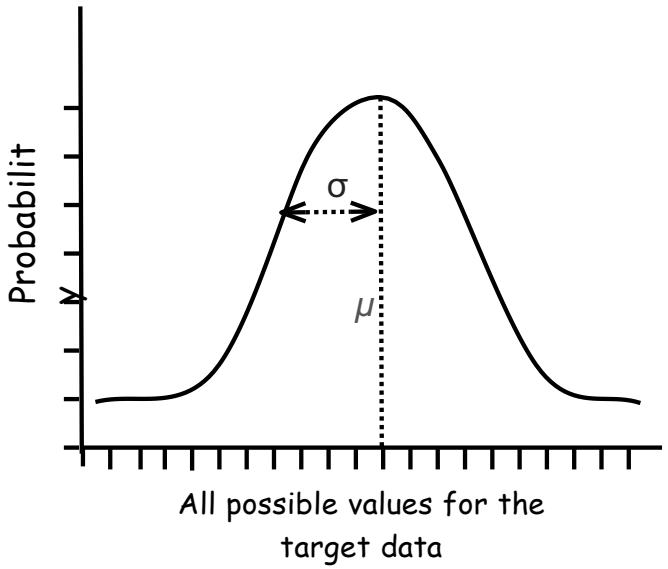
- Probability Mass Function (PMF) shows how **dense** is the probability at **each data point** in a **discrete vector**.

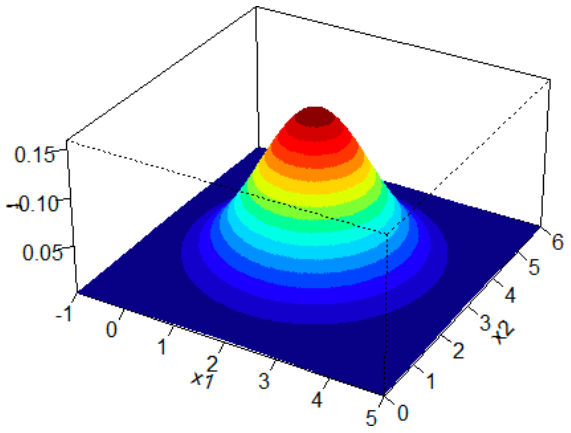
- Cumulative Distribution Function CDF is a function that describes a **distribution** of a variable (either discrete or continuous variable).

Common Probability Density Functions

- Some of the more common discrete probability functions are:
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution

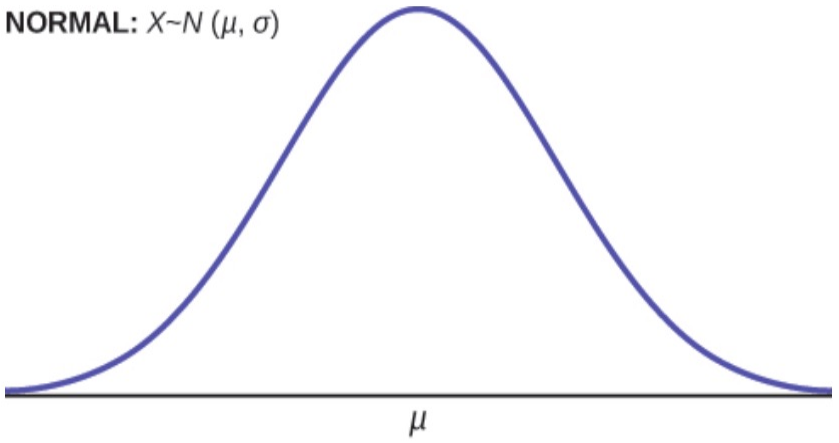
Normal (Gaussian) Distribution





Normal Distribution

- Bell shape (symmetrical)
- Mean = Mode = Median
- Has two parameters Mean and Std
- 68% of the data falls within 1 standard deviation
- Following equation presents the PDF of normal distribution for any given variables of *x*. However, you do not need to memorize it.



$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

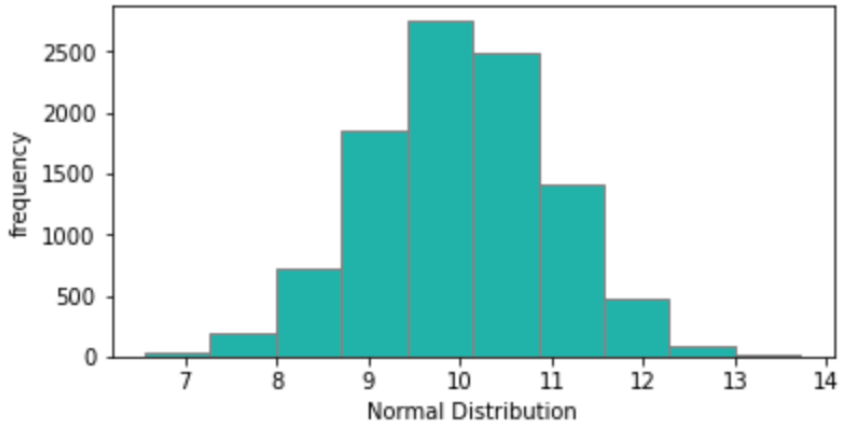
Normal Distribution

In [1]:
```
import numpy as np
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
# Generates a normal continuous random variable.
# The location (loc) keyword specifies the mean.
# The scale (scale) keyword specifies the standard deviation.
normal_data =stats.norm(scale=1, loc=10).rvs(10000)
```

Normal Distribution

In [3]:
```
fig, axs = plt.subplots(figsize =(6, 3))
axs .hist(normal_data,  color = "lightseagreen", ec="grey" )

# add labels and titles
plt.xlabel("Normal Distribution")
plt.ylabel("frequency")

# show plot
plt.show()
```



Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

$$z = (x - \mu) / \sigma$$

The z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

Binomial Distribution

- There are three characteristics of a binomial experiment:
  1. There are a fixed number of trials. A trials is repetitions of an experiment (n is a finite number).
  2. There are only two possible outcomes, called "success" and "failure," for each trial (binary outcome). If p is the probability of success and qnis the probability of failure p+q=1.
  3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial.

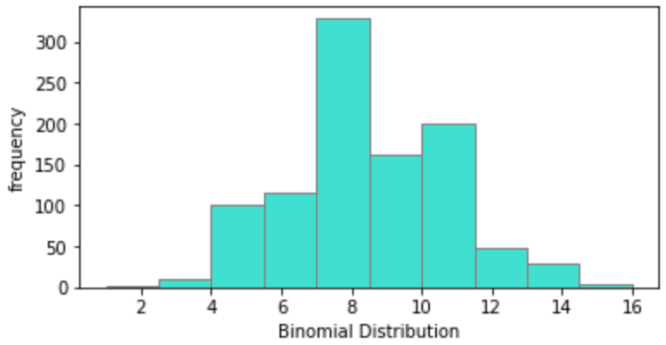
Binomial Distribution Example

- It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?
- Let X = the number of workers who have a high school diploma but do not pursue any further education. X takes on the values 0, 1, 2,..., 20 where n = 20, p = 0.41, and q = 1 – 0.41 = 0.59. $X \sim B(20, 0.41)$ Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$.

Binomial Distribution

In [4]:
```
# generate data for binomial distribution
# n== number of trials,p== probability of success in each trial
n, p = 20, .41
binomial_data = np.random.binomial(n, p, 1000)
```

In [5]:
```
fig, axs = plt.subplots(figsize =(6, 3))
axs .hist(binomial_data,  color = "turquoise", ec="grey" )

# add labels and titles
plt.xlabel("Binomial Distribution")
plt.ylabel("frequency")

# show plot
plt.show()
```



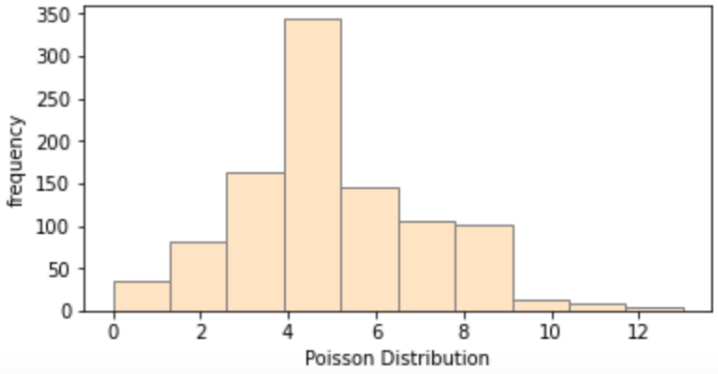
Poisson Distribution

- There are two main characteristics of a Poisson experiment.
  1. The Poisson probability distribution gives the probability of a number of events occurring in a fixed interval of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
  2. The Poisson distribution may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1,000). You will verify the relationship in the homework exercises. n is the number of trials, and p is the probability of a "success." The random variable X = the number of occurrences in the interval of interest.

Poisson Distribution

In [6]:
```
# generate poisson data
# lam: Expected number of events occurring in
# a fixed-time interval, must be >= 0
poisson_data = np.random.poisson(lam=5, size=1000)
```

In [7]:
```
fig, axs = plt.subplots(figsize =(6, 3))
axs .hist(poisson_data,  color = "bisque", ec="grey" )

# add labels and titles
plt.xlabel("Poisson Distribution")
plt.ylabel("frequency")

# show plot
plt.show()
```



Central Limit Theorem & The Law of Large Numbers

- Central Limit Theorem states if we collect enough amount of data the distribution tends toward a normal distribution even if the original variables themselves are not normally distributed.

- The Law of Large Numbers (LLN) is a theorem that explains performing an experiment several times on the sample dataset brings the sample parameters (mean, median, …) closer to the population parameters, and as the number of experiments increases, the sample characteristics should get closer to the population.

Confidence Interval

- Confidence Interval is used to identify the interval or a range of values from sample to estimate the chance whether our sample reflects the data in the population. CI is operated based on a confidence level.
- The confidence level usually is 90%, 95% or 99%. There are other levels as well, but 95% is the most common used value for confidence level.
- Confidence interval of 95% means, there is 5% chance that your data is outside these values.

Confidence Interval

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

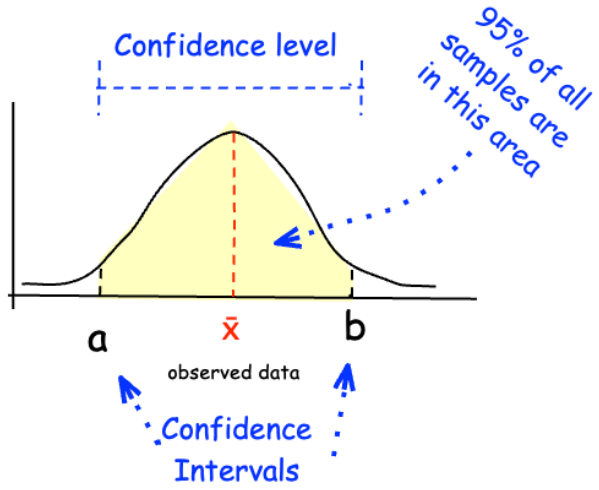
CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size



Confidence Interval Example

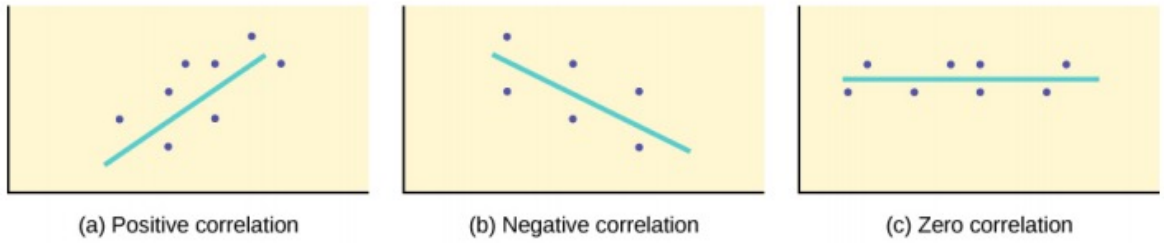
In [1]:
```
import numpy as np
import scipy.stats as st

# define sample data
data = [1, 1, 1, 2,
        3, 3, 3, 3,
        4, 5, 5, 5,
        8, 8, 9, 10]


#create 95% confidence interval for population mean weight
st.t.interval(alpha=0.95, df=len(data)-1,
              loc=np.mean(data), scale=st.sem(data))
```

Out[1]:
```
(2.8812894117023977, 5.993710588297603)
```

Correlation

Correlation is a statistical measure that describes the relationship between two or more variables. It indicates the strength and direction of the association between the variables.

Example: Ice cream consumption is positively correlated to crime rate
(correlation does not imply causation)



Pearson Correlation

- Formula for Pearson's correlation coefficient:

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

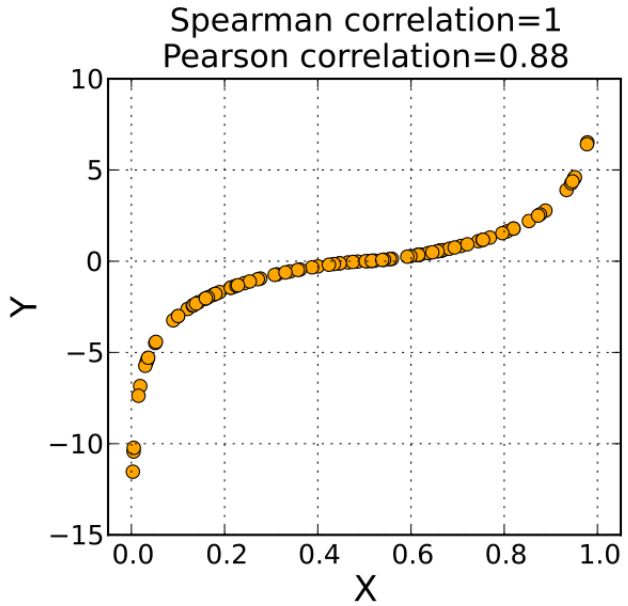
Pearson Correlation Assumptions

- Data from both variables follow normal distributions.
- Your data have no outliers.
- Your data is from a random or representative sample.
- You expect a linear relationship between the two variables.

Spearman Correlation

The Spearman's test is a non-parametric version of the parametric Pearson bivariate correlation coefficient.

Parametric tests and non-parametric tests are distinguished on the basis of assumptions that they make about the nature of the data to be analyzed.



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Pearson vs Spearman Correlation

- The Pearson correlation coefficient assesses the linear relationship between variables, while the Spearman correlation coefficient evaluates the monotonic relationship.

Statistical Significance and P Values

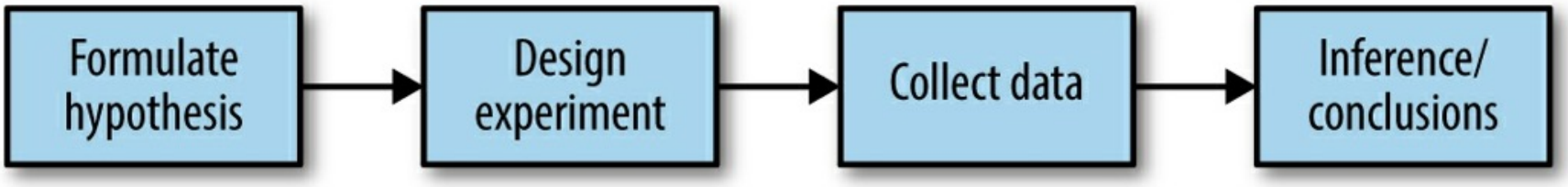
- How do we know our findings are generalizable?

Significance Test

Statistical Experiments & Significance Testing

- Design of experiments is a cornerstone of the practice of statistics, with applications in virtually all areas of research.
- The goal is to design an experiment in order to confirm or reject a hypothesis.

The classical statistical inference pipeline



The classical statistical inference pipeline

- Whenever you see references to statistical significance, t-tests, or p-values, it is typically in the context of the statistical inference pipeline This process starts with a hypothesis:
  - drug A is better than the existing standard drug
  - price A is more profitable than the existing price
- An experiment (example: A/B test) is designed to test the hypothesis — designed in such a way that it will deliver conclusive results.
- The data is collected and analyzed, and then a conclusion is drawn.

A/B Testing

- An A/B test is an experiment with two groups to establish which of two treatments, products, procedures is superior.
- Often one of the two treatments is the standard existing treatment, or no treatment.
- If a standard (or no) treatment is used, it is called the control. A typical hypothesis is that a new treatment is better than the control

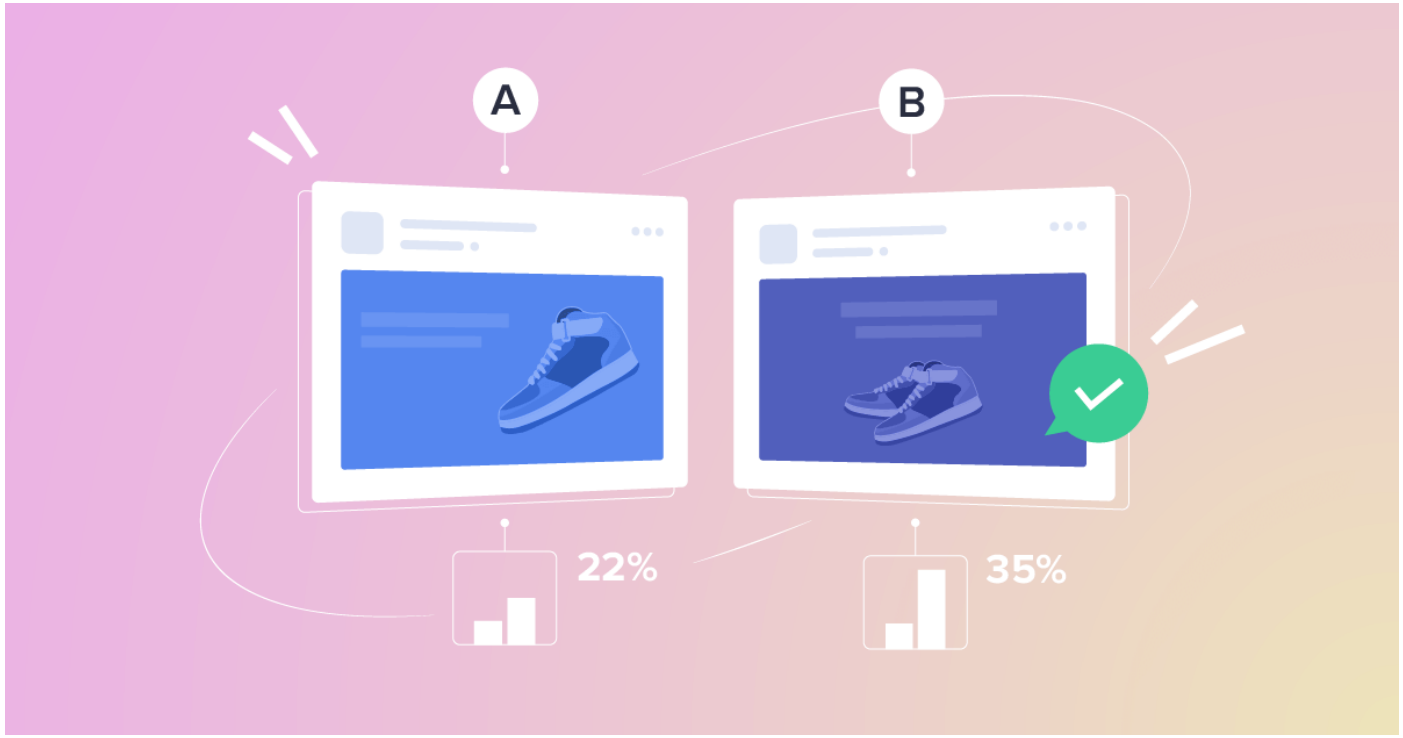
Key Terms of A/B Testing

- Treatment: Something (drug, price, web headline) to which a subject is exposed.
- Treatment group: A group of subjects exposed to a specific treatment.
- Control group: A group of subjects exposed to no (or standard) treatment.
- Randomization: The process of randomly assigning subjects to treatments.
- Subjects: The items (web visitors, patients, etc.) that are exposed to treatments.
- Test statistic: The metric used to measure the effect of the treatment.

A/B Testing

- A/B tests are common in web design and marketing, since results are so readily measured.
- Examples
  - Testing two therapies to determine which suppresses cancer more effectively
  - Testing two prices to determine which yields more net profit
  - Testing two web headlines to determine which produces more clicks
  - Testing two web ads to determine which generates more conversions

A/B Testing on a Facebook Ad



https://madgicx.com/blog/a-b-testing-facebook

Statistical Significance and p-Values

- Statistical significance is how we can measure whether an experiment yields a result more extreme than what chance might produce.
- If the result is beyond the realm of chance variation, it is said to be statistically significant.
- p-value: Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.
- Null Hypothesis that there is no significant difference between specified populations

Null Hypothesis, P_Value

- Null Hypothesis states that there is no significant difference between specified populations and any effect you observe is due to random chance.
- p-value: Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

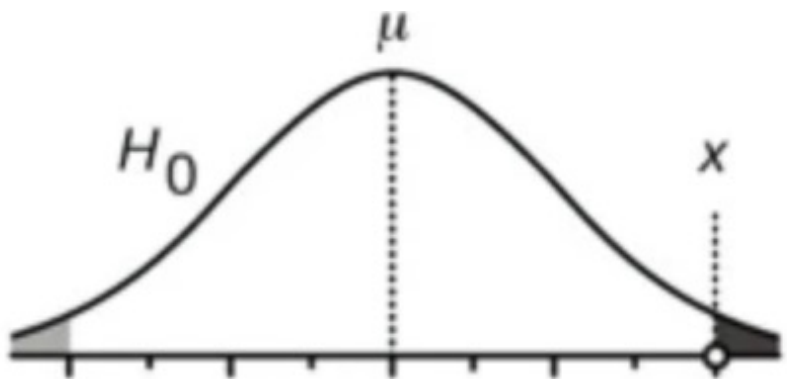
Hypothesis Test

- The hypothesis test assumes that the null hypothesis is true, creates a "null model" (a probability model), and tests whether the effect you observe is a reasonable outcome of that model.
- If P_Value is smaller than significance level (0.05), then you could reject the null hypothesis and conclude your result is statistically significant.

Significance Level

- Alpha
- Researchers decide at what significance level a result is “too unusual” to happen by chance.
- Rather, a threshold is specified in advance, as in “more extreme than 5% of the chance (null hypothesis) results”; this threshold is known as alpha.
- Typical alpha levels are 5% and 1%.

Statistical Significance and p-Values



Parametric vs Non-Parametric Tests

- Parametric significance tests assume that all samples have a normal distribution.

- Non-parametric significance tests do not rely on the normal distribution of samples. Whenever, you encounter the term non-parametric, it is distribution free.

t_Test

- t_Tests is a very common significance test, named after Student's t-distribution.
- t-Test checks whether the means of two groups are reliably different.

t_Test

- Two Sample t-test (Independent): We use this t-test to compare the mean of two groups which are independent and report whether there is a statistical significance among them. (Example: A/B testing)
- Paired t-test: It is used when we have one group of data, that is measured at two different times. It is another form of one-sample t-test. Usually, this test is being used to check if the new treatment, method, etc. is effective and works better than previous method or not.

t_Test Example

In [1]:
```
import numpy as np
from scipy import stats
```

In [2]:
```
rng = np.random.default_rng()
# create two samples with the same mean (loc) and same standard deviation (scale)
random_sample_1 = stats.norm.rvs(loc=5, scale=10, size=1000, random_state=rng)
random_sample_2 = stats.norm.rvs(loc=5, scale=10, size=500, random_state=rng)
```

In [3]:
```
stats.ttest_ind(random_sample_1, random_sample_2)
# pvalue suggest the two samples are from the same distribution
```

Out[3]:
```
Ttest_indResult(statistic=0.40227484354553344, pvalue=0.6875391582946528)
```

t_Test Example

In [4]:
```
# create another samples with the differnt parameters
random_sample_3 = stats.norm.rvs(loc=15, scale=10, size=800, random_state=rng)
```

In [5]:
```
stats.ttest_ind(random_sample_1, random_sample_3)
# pvalue< 0.05 suggest the two samples are from the different distribution
# rejects the null hypothesis
```

Out[5]:
```
Ttest_indResult(statistic=-21.68310517078561, pvalue=8.338233821214685e-93)
```

t_Test Example

In [6]:
```
# Lets explore a paired t test
# Imagine we have the mesurements of A1C% for a group of patients who
# went through a specific diet and excersie for two weeks and repeated A1C

A1C_1 = [5.5, 6.7, 8.4, 9.9, 8.4, 5.2, 9.3, 6.0, 9.1, 9.1, 8.3, 8.9, 7.7, 6.8, 9.1]
A1C_2 = [5.2, 6.5, 8.4, 8.0, 8.1, 5.2, 8.5, 6.1, 7.8, 8.6, 7.2, 8.9, 8.1, 6.4, 7.9]
```

In [7]:
```
#perform the paired samples t-test
stats.ttest_rel(A1C_1, A1C_2)

# pvalue< 0.05 suggest the two samples are from the different distribution
# rejects the null hypothesis
```

Out[7]:
```
Ttest_relResult(statistic=3.05098519142553, pvalue=0.008632904623380018)
```

t_Test Assumptions

- The data are continuous.
- The data is sampled at random from a population.
- Homogeneity of variance is present, indicating that the variability within each group of data is comparable.
- The distribution is approximately normal.

ANOVA
Analysis Of Variance

Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A/B/C/D, each with numeric data. The statistical procedure that tests for a statistically significant difference among the groups is called analysis of variance, or ANOVA.

ANOVA

The H0 in ANOVA assumes that all groups' mean are equal.

H1 assumes at least two of group means are different.

$$H0: \mu 1 = \mu 2 = \mu 3$$

H1: Means are not all equal.

F-statistic: A standardized statistic that measures the extent to which differences among group means exceed what might be expected in a chance model.

F-statistic

$$F = MS_B / MS_W$$

Where:

MS_B = Sum of squares between samples (SS_B) / (k-1)

MS_W= Sum of squares within samples (SS_w) / (n-k)

k is the number of groups
n is the total number of observations

“Sum of squares,” referring to deviations from some average value.

ANOVA

One-Way ANOVA: is a hypothesis test, which tests the equality of three of more population means simultaneously using variance.

- Number of observations could be different in each group.
- Number of independent variables is one.

Two-Way ANOVA: is a statistical technique which studies the interaction between factors, influencing variable.

- Number of observations need to be equal in each group.
- Number of independent variables is two.

One Way ANOVA Assumptions

- The responses for each factor level have a normal population distribution.
- These distributions have the same variance.
- The data-points are independent.

ANOVA Example
Web Stickiness Data for 4 web pages (in seconds)

Page 1	Page 2	Page 3	Page 4
164	178	175	155
172	191	180	159
177	182	178	154
156	165	170	151
195	187	172	150

ANOVA Example
Web Stickiness Data for 4 web pages (in seconds)

In [1]:
```
from matplotlib import pyplot as plt
import numpy as np
from scipy import stats as st
import pandas as pd
from scipy.stats import f_oneway
```

In [2]:
```
# Web stickiness in seconds of four web pages

data=[[164, 178, 175, 155],
      [172, 191, 180, 159],
      [177, 182, 178, 154],
      [156, 165, 170, 151],
      [195, 187, 172, 150]]

dataframe = pd.DataFrame(data, columns=["Page1","Page2","Page3","Page4" ] )
```

ANOVA Example
Web Stickiness Data for 4 web pages (in seconds)

```
In [3]: dataframe.head()
```

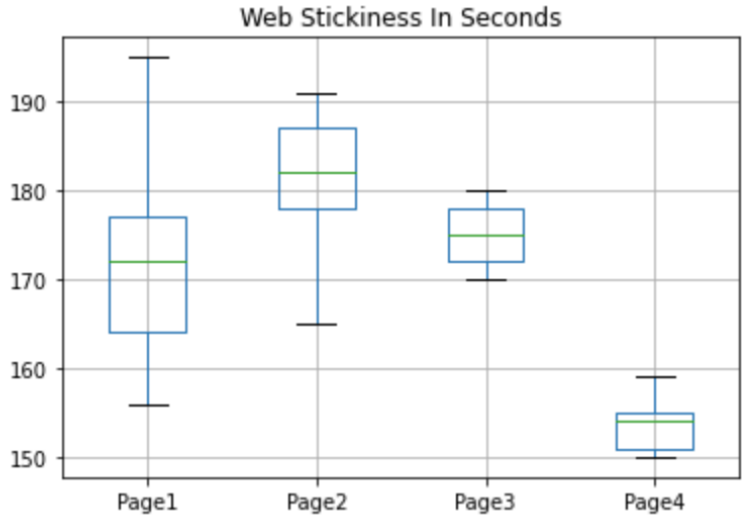
Out[3]:

	Page1	Page2	Page3	Page4
0	164	178	175	155
1	172	191	180	159
2	177	182	178	154
3	156	165	170	151
4	195	187	172	150

ANOVA Example
Web Stickiness Data for 4 web pages (in seconds)

```
In [4]: dataframe.boxplot()
        plt.title('Web Stickiness In Seconds')
```

```
Out[4]: Text(0.5, 1.0, 'Web Stickiness In Seconds')
```



ANOVA Example
Web Stickiness Data for 4 web pages (in seconds)

```
In [5]: ANOVA = f_oneway(dataframe.Page1, dataframe.Page2, dataframe.Page3, dataframe.Page4)
        ANOVA

Out[5]: F_onewayResult(statistic=7.792733199118014, pvalue=0.0019812750431078695)
```

References

Illowsky, Barbara; Dean, Susan. Introductory Statistics. XanEdu Publishing Inc.