LECTURE 04

Distributions and Pandas

September 25, 2023

PBHLTH 198, Fall 2023 @ UC Berkeley

Andrew O'Connor

Class Outline

- Recap of lecture 3
- Famous Distributions
  - Binomial, Poisson, Normal
- Pandas
- HW 4 time

Announcements

- Gradescope
- Two options for running notebooks
  - Locally
  - Deepnote
- Resources
  - Front page of bCourses
  - Module: Statistics and Coding Resources
- Readings and key takeaways updated for weeks 1-3
  - optional readings: good for building intuition!
- Final project?

Announcements

International Student Job Sponsorship #1964

Anonymous Dugong
2 days ago in General

STAR WATCH 219 VIEWS

2

Are there any support systems/resources on campus to help international EECS/CS/DS students get a job/internship through sponsorship?

Graduating senior, so I desperately need one. Do any of y'all have advice on any resources I can utilize (or any other useful info)?

Comment •••

1 Answer

A Amanda Dillon STAFF
12 hours ago #1964a

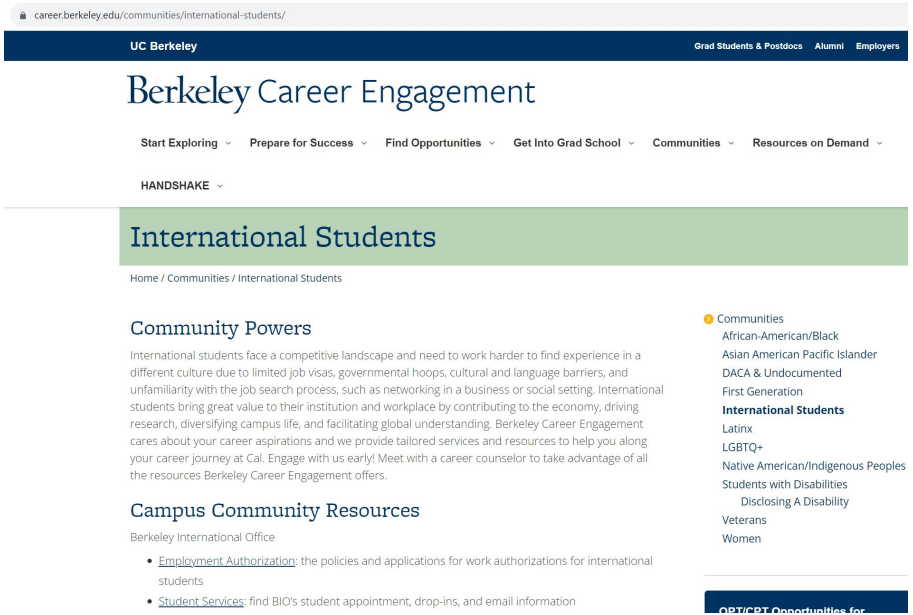
1

Couple of things on this:

- There is an advisor in the career center that specializes in helping international students navigate the job market. Her name is Jing Han. You can schedule a 1:1 session with her through Handshake.
- There is a tool through the career center called "GoinGlobal" that helps international students identify US companies that have provided sponsorship in the past.
- Tomorrow there is an event focused on helping international students understand H-1B regulations. You can learn more and register on Handshake.

Hope this helps!

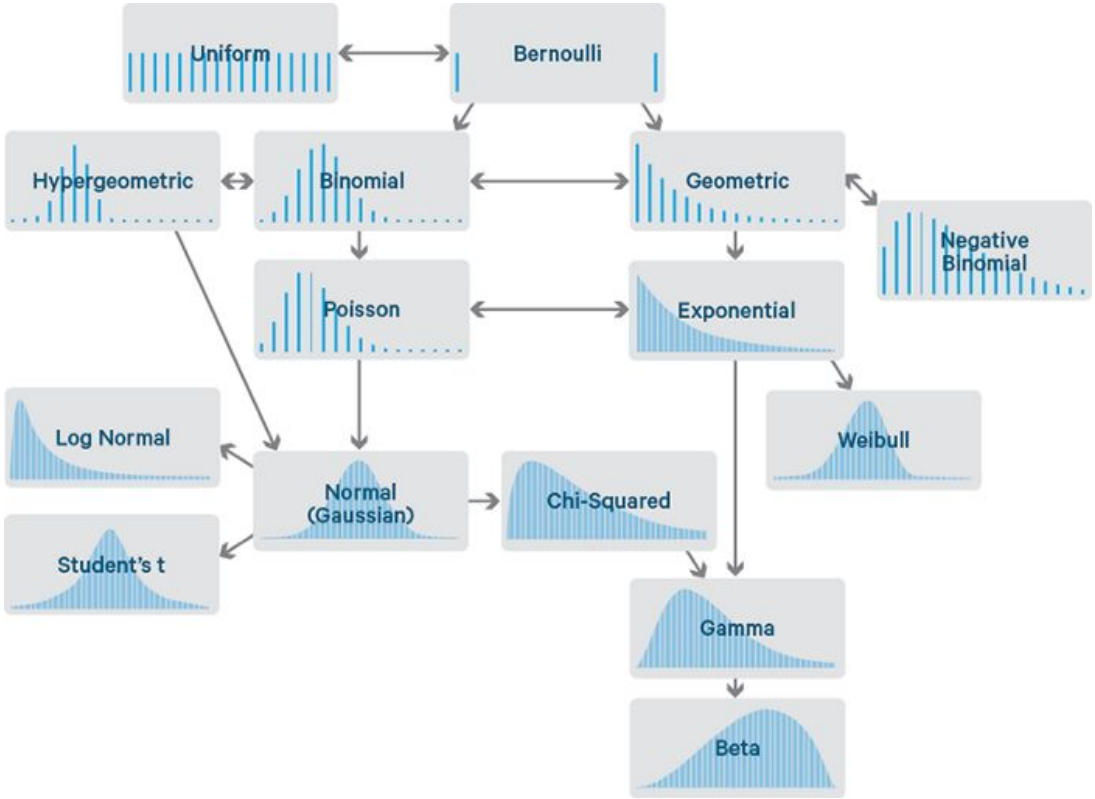
Comment •••



Random Variables have Distributions

- A probability distribution describes how the values of a random variable are spread out or distributed
  - the likelihood of each possible value occurring
- Some properties of distributions
  - PDF (continuous), PMF (discrete)
  - Mean/Expected value
  - Variance
  - Standard deviation
  - Covariance

Famous Distributions



Distributions

Binomial

Poisson

Normal

Binomial

- Discrete distribution
  - Defined by a probability mass function (pmf)
- Interpretation: describes the number of successes in a series of independent Yes/No experiments all with the same probability of success

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Binomial - Examples

Example

An elementary school adminsters eye exams to 800 students. How many students have perfect vision?

Let X represent the number of students with perfect vision. Then

$$X \sim Binom(n = 800, p = ?)$$

where n is the number of students in the school, and p is the probability of having perfect vision. Here we don't know p because it wasn't provided in the question.

- Is X a discrete or a continuous random variable?
- What is the range of values that X can take, hypothetically?

Example 2

Forty-five percent of the population is blood type O. Consider the next five blood donations from unrelated individuals. The number who have type O is the count X of successes in 5 independent observations. Here,

$$X \sim Binom(n = 5, p = 0.45)$$

- Is X a discrete or a continuous random variable?
- What is the range of values that X can take, hypothetically?

Poisson

- Discrete distribution
  - Defined by a probability mass function (pmf)
- Interpretation: describes a very large number of individually **unlikely** events that happen in a certain time interval

$$P(X = x) = \frac{e^{-\mu}\,\mu^{x}}{x!}$$

Poisson - Examples

Example 1 of the Poisson distribution

The Poisson distribution can be used to model rare, but infectious diseases. For example, the number of deaths X attributed to typhoid fever over 100 years follows a Poisson distribution if:

a) The probability of a new death from typhoid fever in any one day is very small.
b) The number of cases reported in any two distinct periods of time are independent random variables.

Each year, the number of deaths from typhoid fever in the US could be recorded. This sequence of deaths over time might look like this: 0, 1, 0, 0, 1, 1, 0, 2, and so on.

Example 2 of the Poisson distribution

The Poisson distribution can also be used to model rare events occurring on a surface area. For example, the distribution of number of bacterial colonies growing on an agar plate. The number of bacterial colonies over the entire agar plate follows a Poisson distribution if:

a) The probability of finding any bacterial colonies in a small area is very small.
b) The events of finding bacterial colonies in any two areas are independent.

The agar plates surface can be divided into several small areas. For each area, you could count the number of bacterial colonies and record this information in a variable in R and it might look like this: 0, 1, 0, 0, 1, 1, 0, 2, and so on.

Normal

- Continuous
  - Defined by probability density function
- Also called the Gaussian distribution, Bell curve

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

Normal

Properties of the Normal distribution

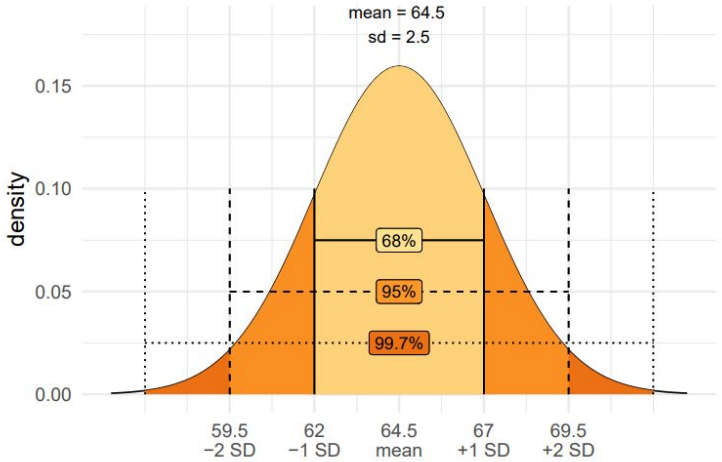
- the density can be drawn by knowing just two parameters , the mean (μ) and SD (σ): $f(x) = \phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- the mean μ can be any value, positive or negative
- the standard deviation σ must be a positive number
- the mean is equal to the median (both $= \mu$)
- the standard deviation captures the spread of the distribution
- the area under the Normal distribution is equal to 1 (i.e., it is a density function)

The 68-95-99.7 rule for all Normal distributions

- Approximately 68% of the data fall within one standard deviation of the mean
- Approximately 95% of the data fall within two standard deviations of the mean
- Approximately 99.7% of the data fall within three standard deviations of the mean

Written probabilistically:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 68\%$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 99.7\%$



Pandas

Overview

Terminology

Pandas

- Pandas is a Python library that specializes in cleaning, managing and manipulating tabular data

In [7]:
```
wine_reviews = pd.read_csv("../input/wine-reviews/winemag-data-130k-v2.csv")
```

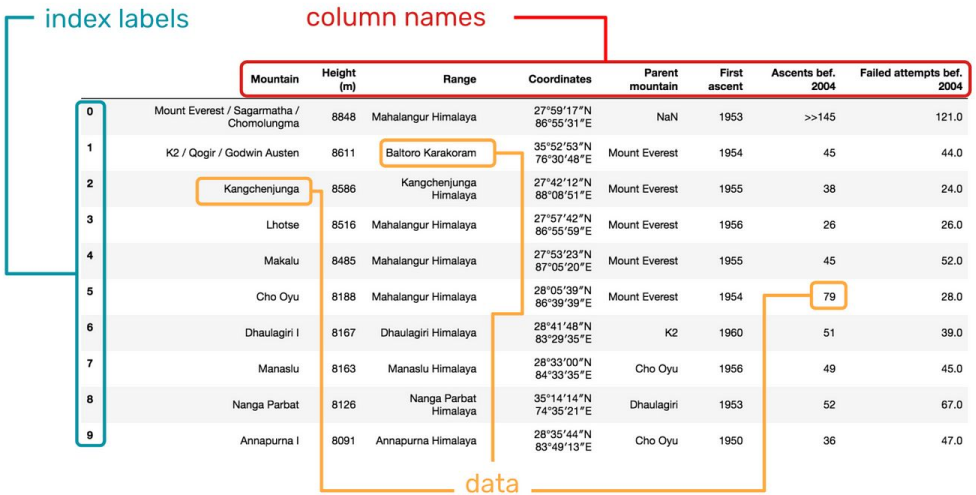
In [9]:
```
wine_reviews.head()
```

Out[9]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_han
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe
1	1	Portugal	This is ripe and fruity, a wine that is	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger

Terminology

- Officially,
  - Tabular data is stored in objects called **Dataframes**
  - The columns of the dataframe are also objects, called **Series**



	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

read: kaggle tutorials

hw time