Topic No. 1

Basic Background Topics of Statistics for Plots and Graphs

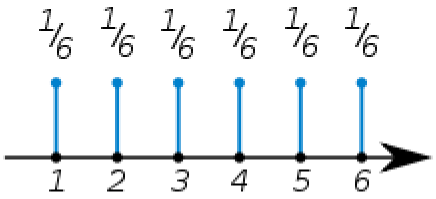
- **Mean, Variation, Standard Deviation, PMF, PDF, CDF.**
- **Histogram vs. Various Probability distributions**
- **Beta Distribution**

PMF, PDF, CDF

- PMF (Probability mass function) : discrete random variable

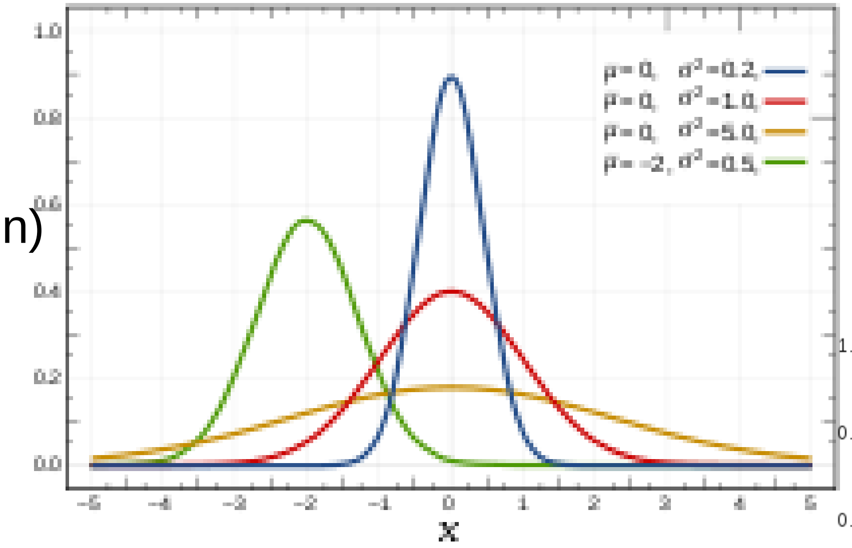
$$\boxed{p_X(x) = P(X = x)}$$

Bernoulli distribution, Binomial distribution



- PDF (Probability density function) : continuous random variable, **relative likelihood**

$$\Pr[a \le X \le b] = \int_a^b f_X(x)\,dx.$$



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

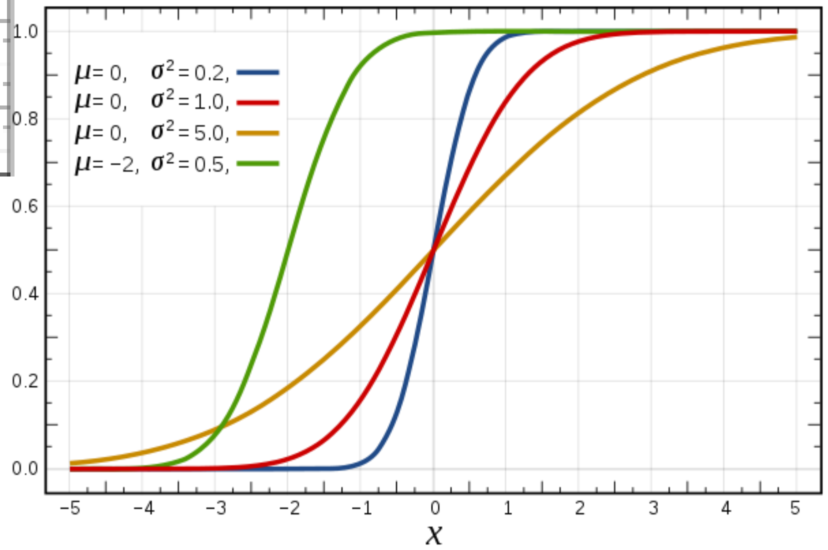
- CDF (Cumulative density function)

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du,$$

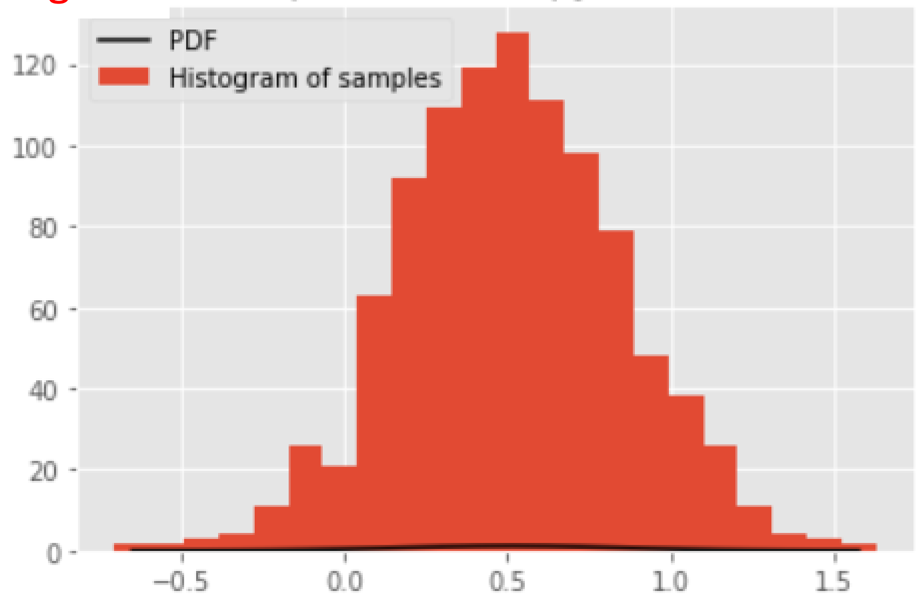
$$F_X(x) = \mathrm{P}(X \le x)$$

$$\mathrm{P}(a < X \le b) = F_X(b) - F_X(a)$$

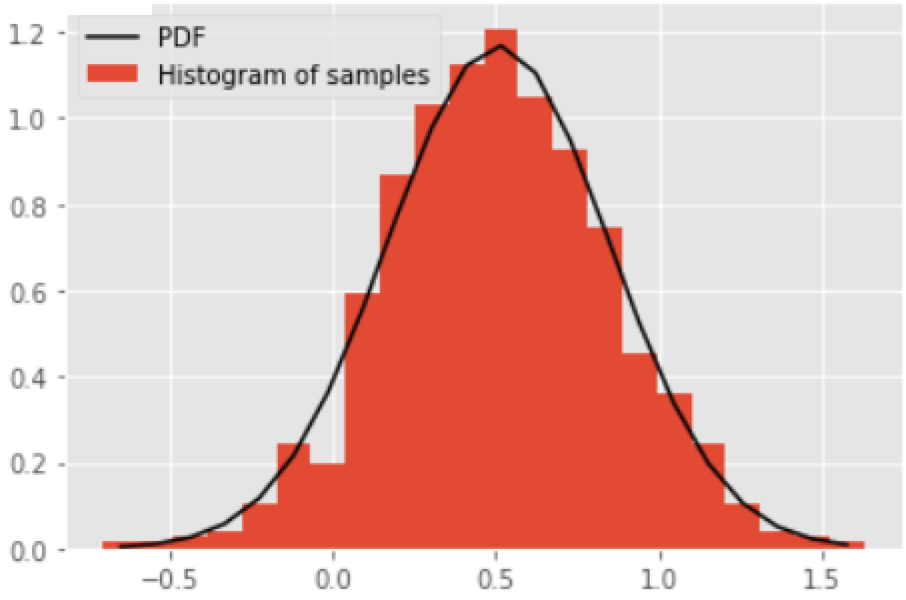
$$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$



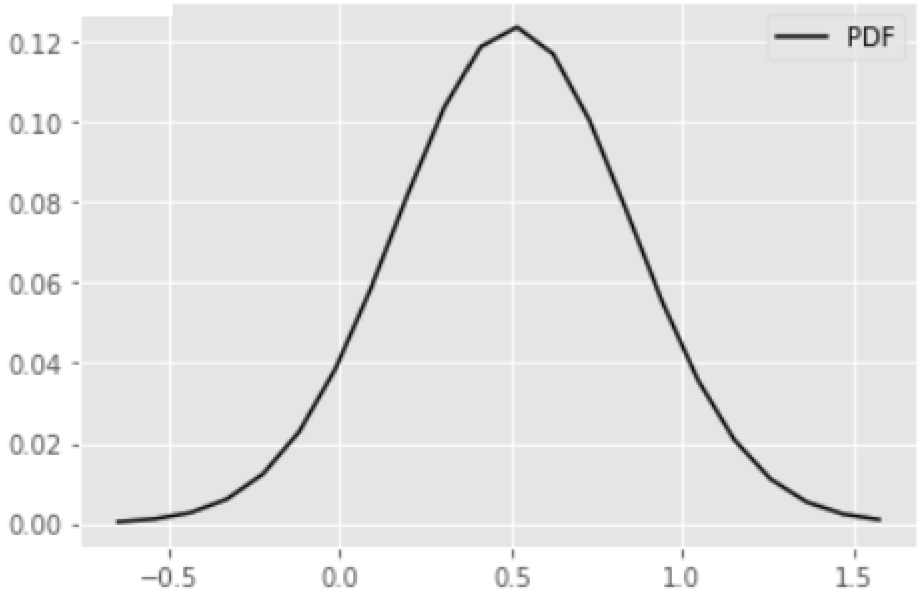
Histogram of samples from numpy.random.normal()



PDF of samples from numpy.random.normal()



Probability Distribution of samples from numpy.random.normal()



Mean, Variation, Standard Deviation

- Mean

$$\mu = \mathrm{E}[X]$$

$$= \int_{\mathbb{R}} x\, \underline{f(x)}\, dx$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance

$$\operatorname{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right]$$

$$= \int_{\mathbb{R}} (x-\mu)^2\, \underline{f(x)}\, dx$$

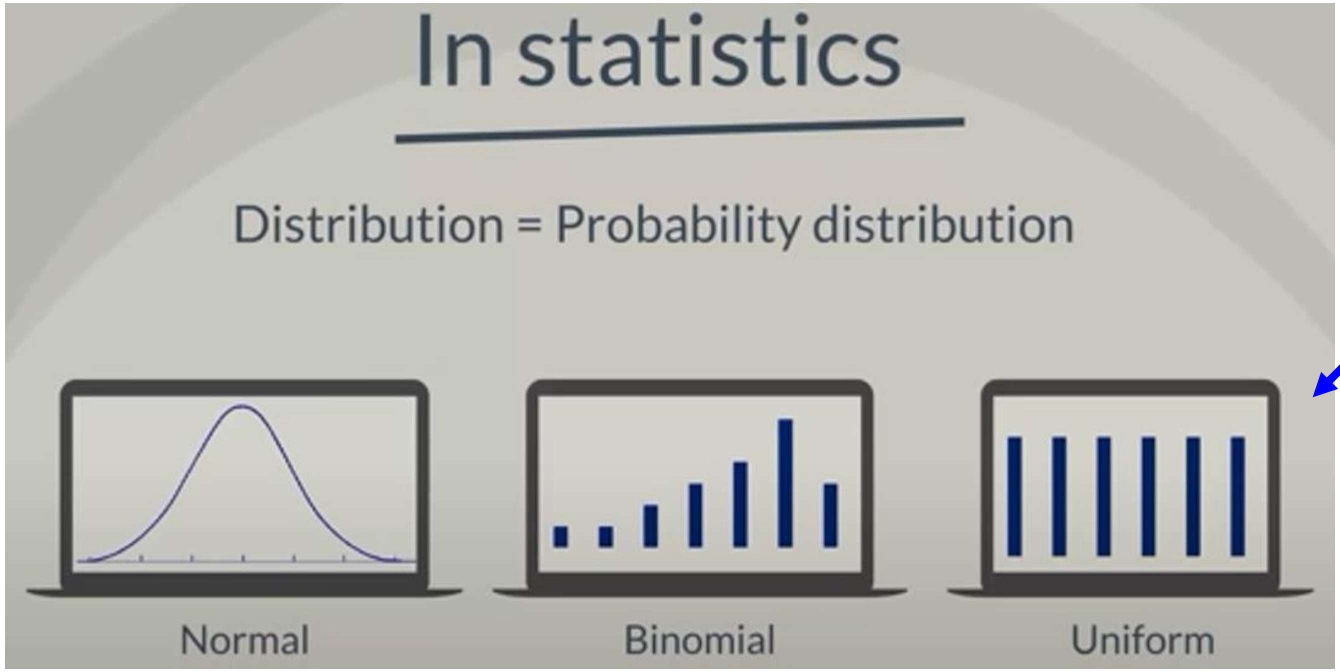
$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

What is a Distribution in Statistics?

When we use the term normal distribution in statistics, we usually mean **a probability distribution.**
Examples are the **Normal(Gaussian) distribution**, the **Binomial distribution**, and the **Uniform distribution.**

A distribution in statistics is a function that shows the **possible values for a variable** and **how often they occur**.

Y axis: the occurrence of every event



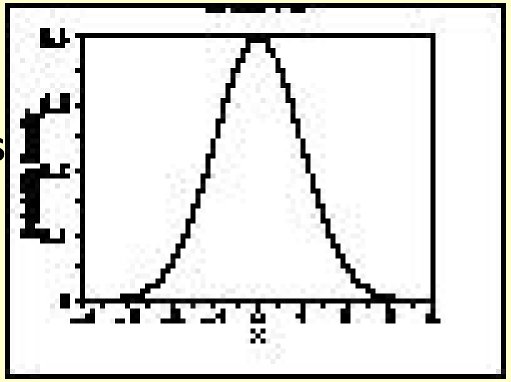
When you roll a die, what is the probability of getting x? What would be the x-axis labels?

The distribution of an event consists not only of the input values that can be observed, but is made up of **all possible values (X axis).**

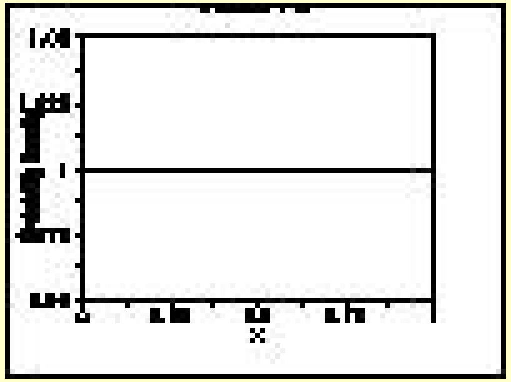
https://365datascience.com/tutorials/statistics-tutorials/distribution-in-statistics/

Different Types of Distributions

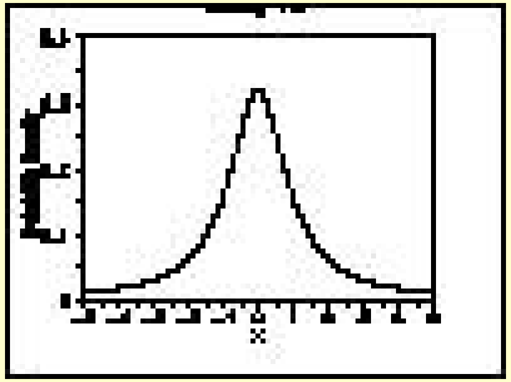
Continuous Distributions



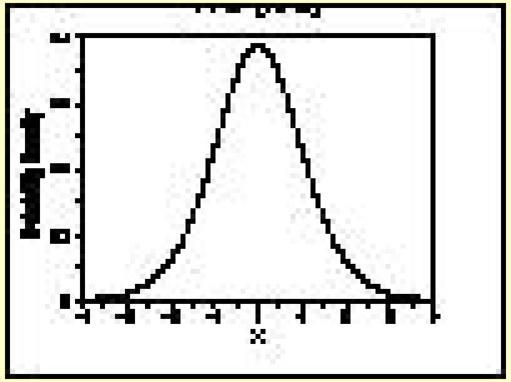
Normal Distribution



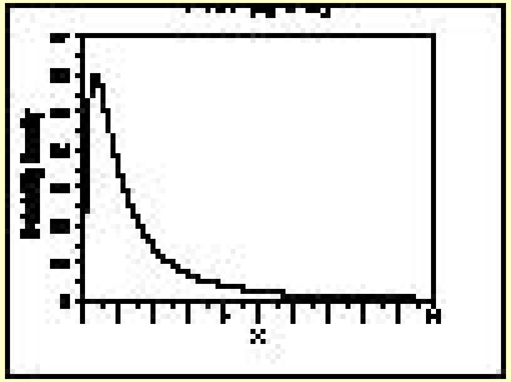
Uniform Distribution



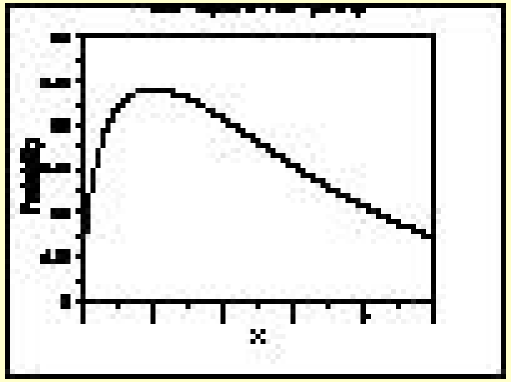
Cauchy Distribution



t Distribution



F Distribution

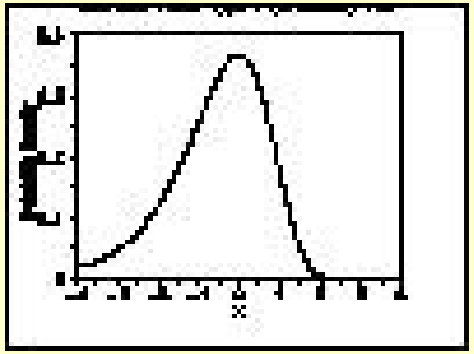


Chi-Square Distribution

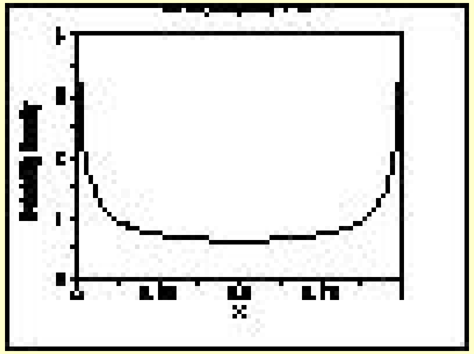
https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

Different Types of Distributions

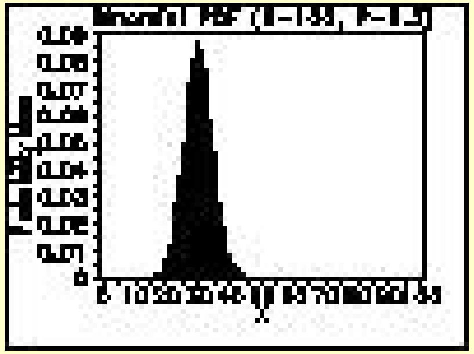
Discrete Distributions



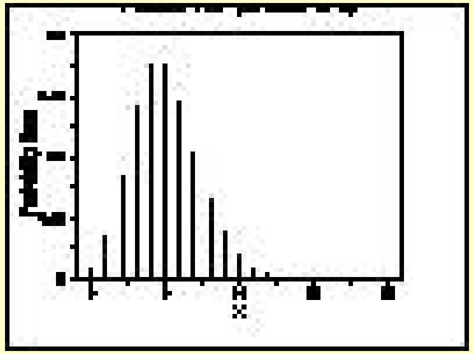
Extreme Value Type I Distribution



Beta Distribution



Binomial Distribution



Poisson Distribution

# of houses a real estate agent may sell a house in a year.

# of deer a hunter can may take in a season.

https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

Histogram

A Motivating Example:

```
import pandas as pd
import os

paths = "E:/Data/"

# read and write to csv
# Add column headers and then print as a .csv file
with_header = pd.read_csv(paths + 'nba.csv', sep=',')
with_header
```

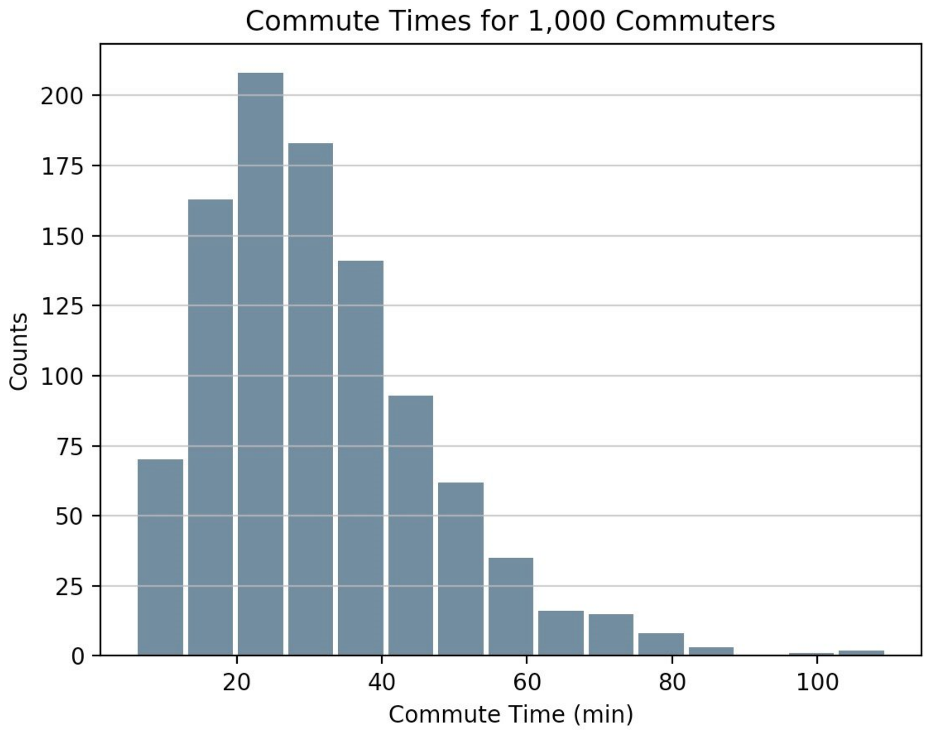
	Name	Team	Number	Position	Height	Age	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	74	25	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	78	25	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	77	27	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	77	22	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	82	29	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
452	Trey Lyles	Utah Jazz	41	PF	82	20	234	Kentucky	2239800.0
453	Shelvin Mack	Utah Jazz	8	PG	75	26	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	73	24	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	87	26	256	NaN	2900000.0
456	Jeff Withey	Utah Jazz	24	C	85	26	231	Kansas	947276.0

457 rows × 9 columns

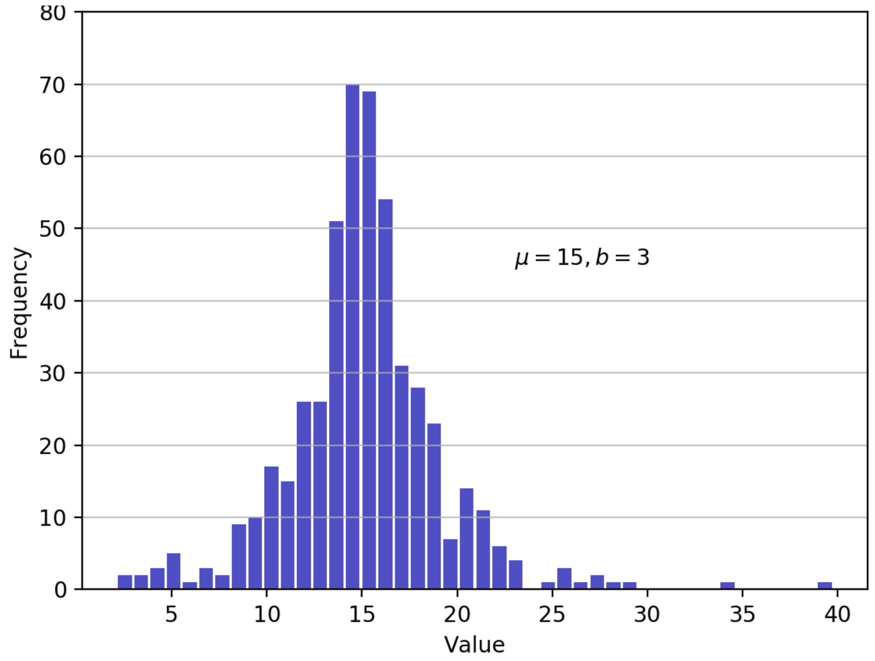
Histogram

Simple python code to plot a histogram

```
import pandas as pd
# Generate data on commute times.
size, scale = 1000, 10
commutes = pd.Series(np.random.gamma(scale, size=size) ** 1.5)
commutes.plot.hist(grid=True, bins=20, rwidth=0.9, color='#607c8e')
plt.title('Commute Times for 1,000 Commuters')
plt.xlabel('Counts')
plt.ylabel('Commute Time')
plt.grid(axis='y', alpha=0.75)
```




```
import matplotlib.pyplot as plt
# An "interface" to matplotlib.axes.Axes.hist() method
n, bins, patches = plt.hist(x=d, bins='auto', color='#0504aa',
                            alpha=0.7, rwidth=0.85)
plt.grid(axis='y', alpha=0.75)
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.title('My Very Own Histogram')
plt.text(23, 45, r'$\mu=15, b=3$')
maxfreq = n.max()
# Set a clean upper y-axis limit.
plt.ylim(ymax=np.ceil(maxfreq / 10) * 10 if maxfreq % 10 else maxfreq + 10)
```



https://realpython.com/python-histograms/

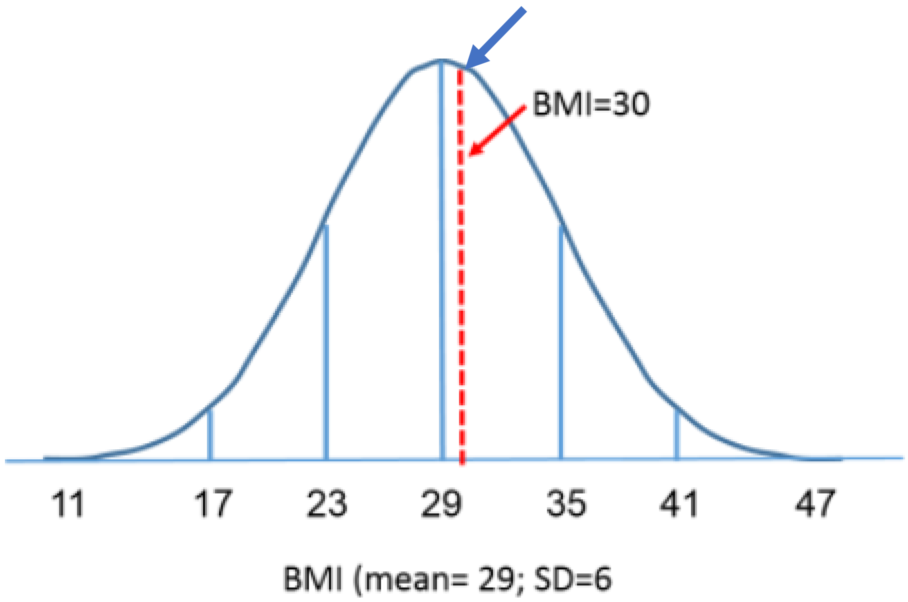
The Standard Normal (Gaussian) Distribution

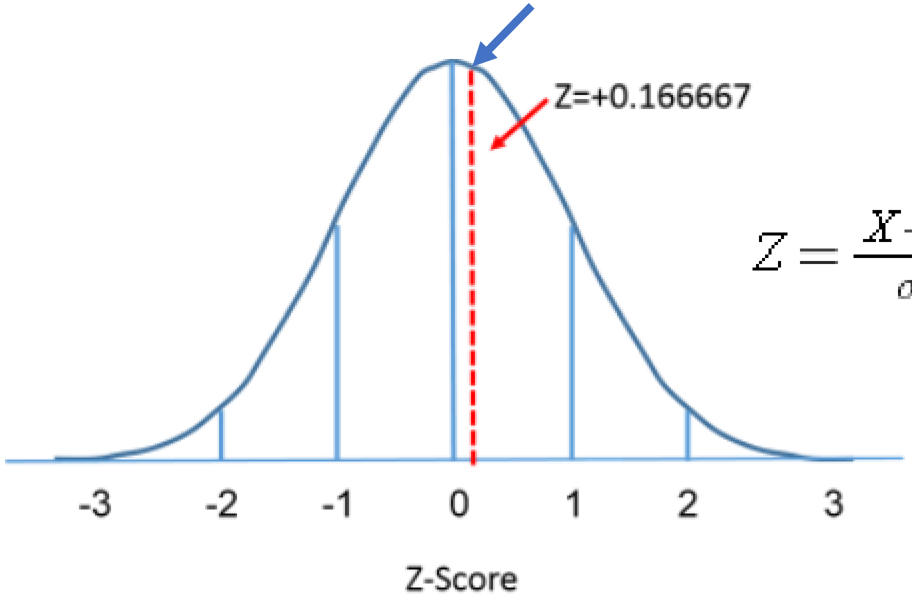
The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1. The standard normal distribution is centered at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation. For the standard normal distribution, 68% of the observations lie within 1 standard deviation of the mean; 95% lie within two standard deviation of the mean; and 99.9% lie within 3 standard deviations of the mean. To this point, we have been using "X" to denote the variable of interest (e.g., X=BMI, X=height, X=weight). However, when using a standard normal distribution, we will use "Z" to refer to a variable in the context of a standard normal distribution. After standardization, the BMI=30 discussed on the previous page is shown below lying 0.16667 units above the mean of 0 on the standard normal distribution on the right.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Example: Show both plots

The distributions of BMI for men aged 60 and the standard normal distribution side-by-side.





$$Z = \frac{X-\mu}{\sigma}$$

https://sphweb.bumc.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability9.html

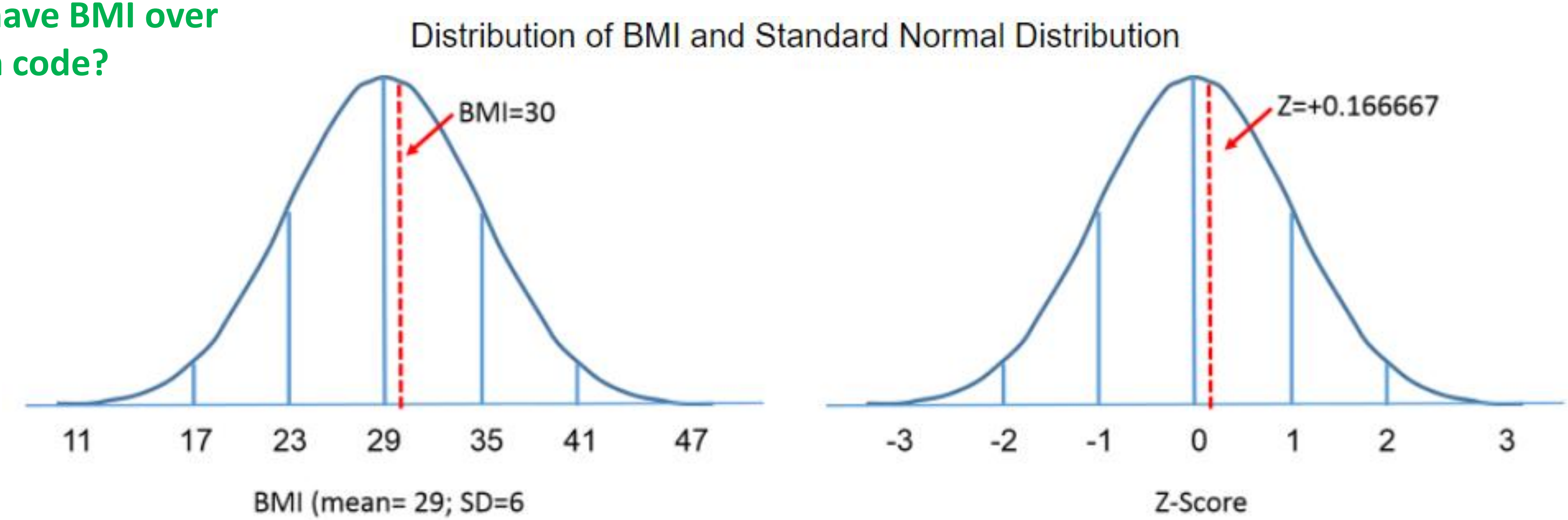
Probabilities of the Standard Normal Distribution Z

The area under the curve to the left of or less of a specified value or "Z value". The area is the probability of observing a value less than that particular Z value.

Example: The probability that the BMI is less than 30, i.e., P(X<30).

Z score, also called a standardized score: $Z = \frac{X - \mu}{\sigma}$

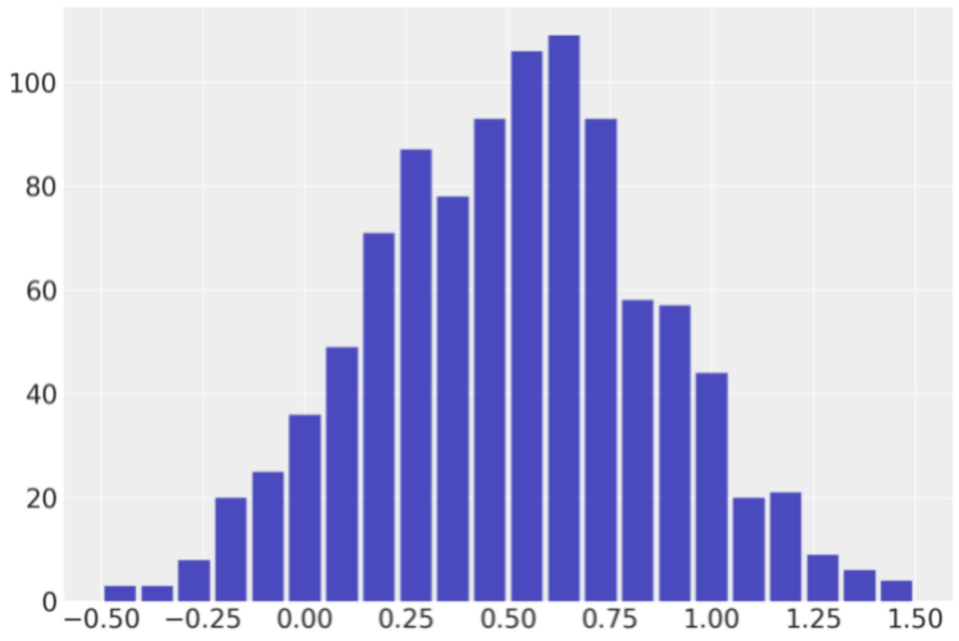
Q: Given 1,000 people, how many would have BMI over 41? → Python code?



Q: Given 1,000 people, how many would have BMI over 35? → Python code?

the areas to the left of the dashed line are the same.

P(X < 30) = P(Z < 0.17)



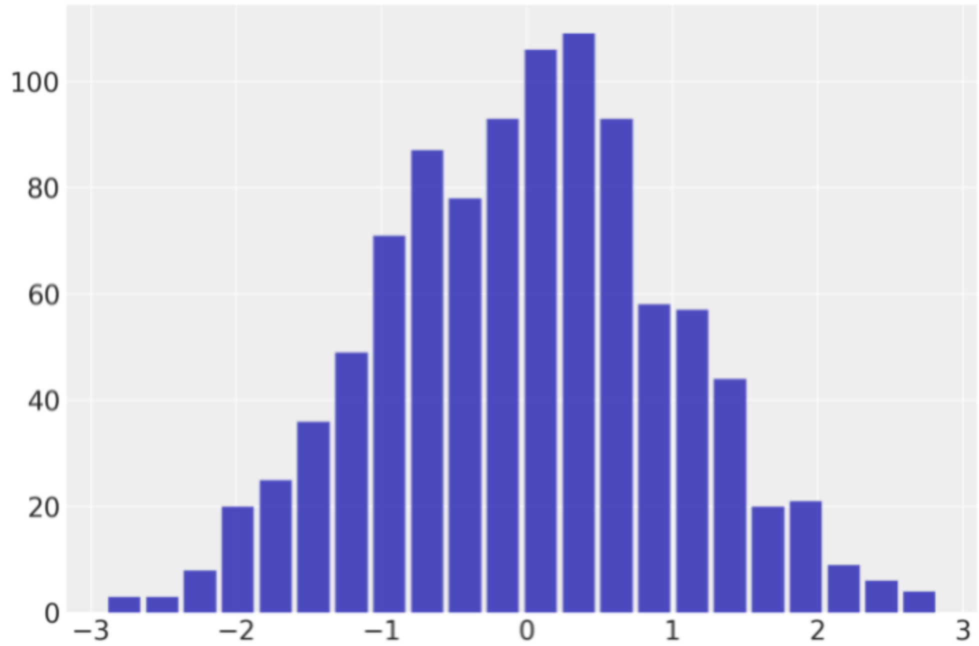

```
import statistics
# 1. generate the artificial dataset. The distribution is mu=0.5, sd=0.35
obs_y = np.random.normal(0.5, 0.35, 1000)

# produce mu and sigma
data =obs_y
mu = statistics.mean(data)
sigma = statistics.stdev(data)
print("Mean is :", mu)
print("STDEV is :", sigma)
```

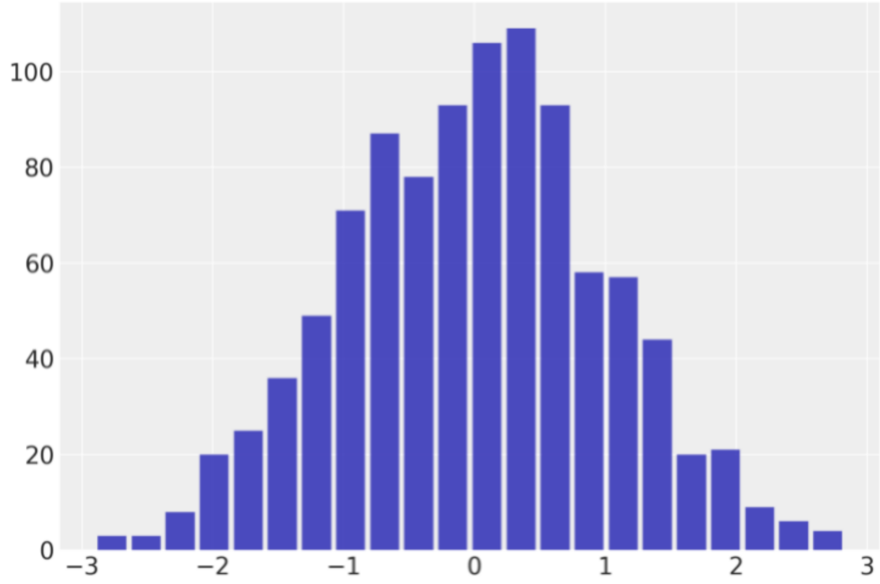
Mean is : 0.5122480231496037
STDEV is : 0.34864815495422685

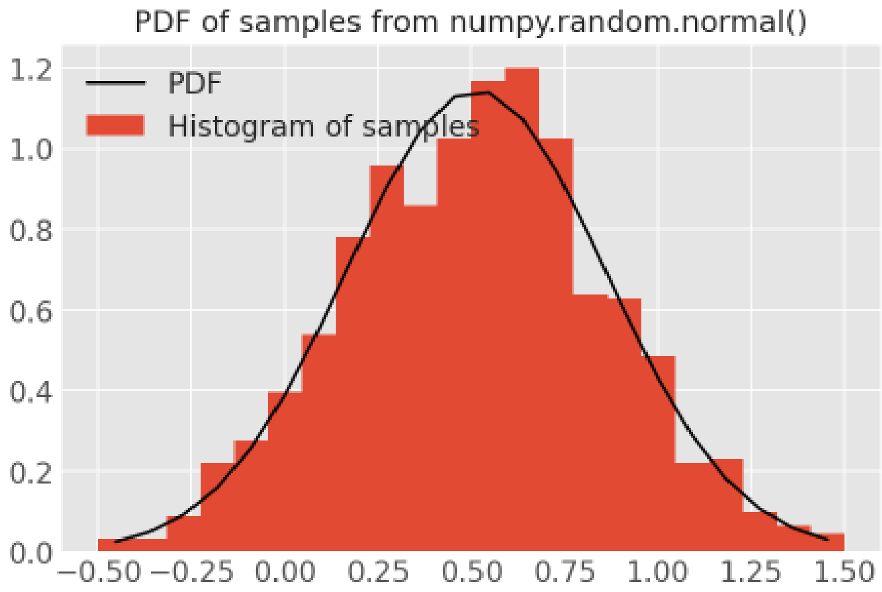
Convert it z-value histogram

```
z = (obs_y - mu)/sigma

n, bins, patches = plt.hist(x=z, bins='auto', color='#0504aa',
                            alpha=0.7, rwidth=0.85)
```

See the change of x-axis







Converting histogram to PDF

Matplotlib histogram and estimated PDF in Python

Typically, if we have a vector of random numbers that is drawn from a distribution, we can estimate the PDF using the histogram tool. Matplotlib's hist function can be used to compute and plot histograms. If the **density argument is set to 'True',** the hist function computes the normalized histogram such that the area under the histogram will sum to 1.

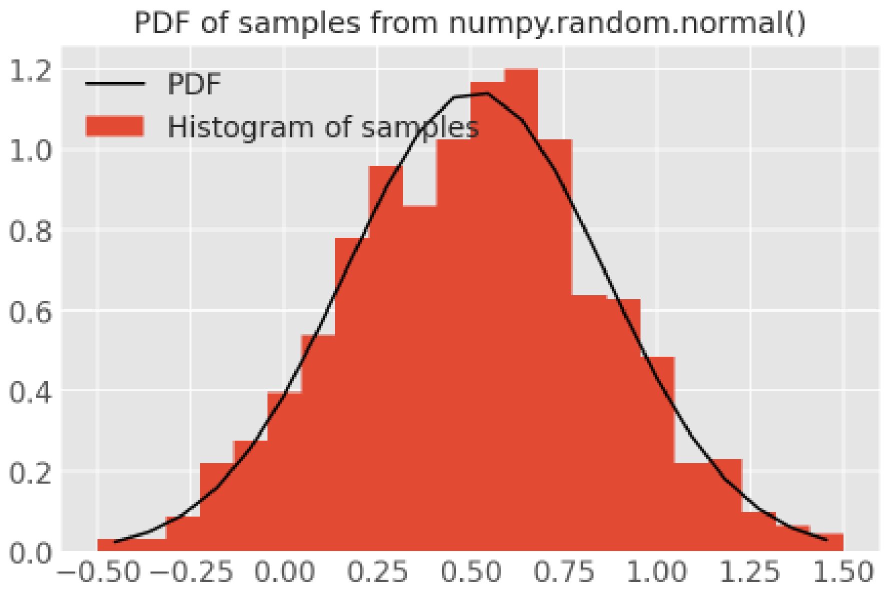
```
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('ggplot')

fig, ax0 = plt.subplots(ncols=1, nrows=1) #creating plot axes
(values, bins, _) = ax0.hist(data, bins=22, density=True, label="Histogram of samples")
#Compute and plot histogram, return the computed values and bins

from scipy import stats
bin_centers = 0.5*(bins[1:] + bins[:-1])
pdf = stats.norm.pdf(x = bin_centers, loc=mu, scale=sigma) #Compute probability density function
ax0.plot(bin_centers, pdf, label="PDF",color='black') #Plot PDF
ax0.legend()#Legend entries
ax0.set_title('PDF of samples from numpy.random.normal()');
```

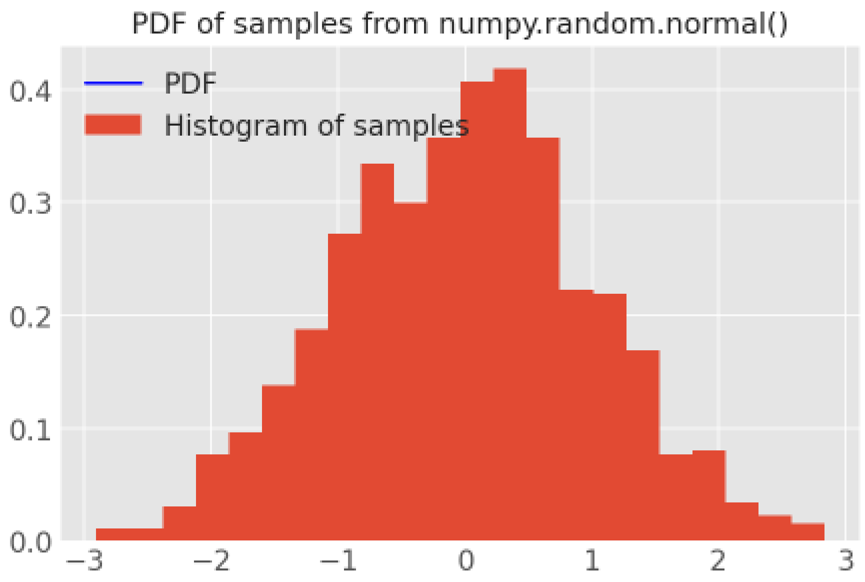
What if you use z instead of data?

Theoretical PDF for normal distribution

https://www.gaussianwaves.com/2020/06/using-matplotlib-histogram-in-python/

What if you use z instead of data?





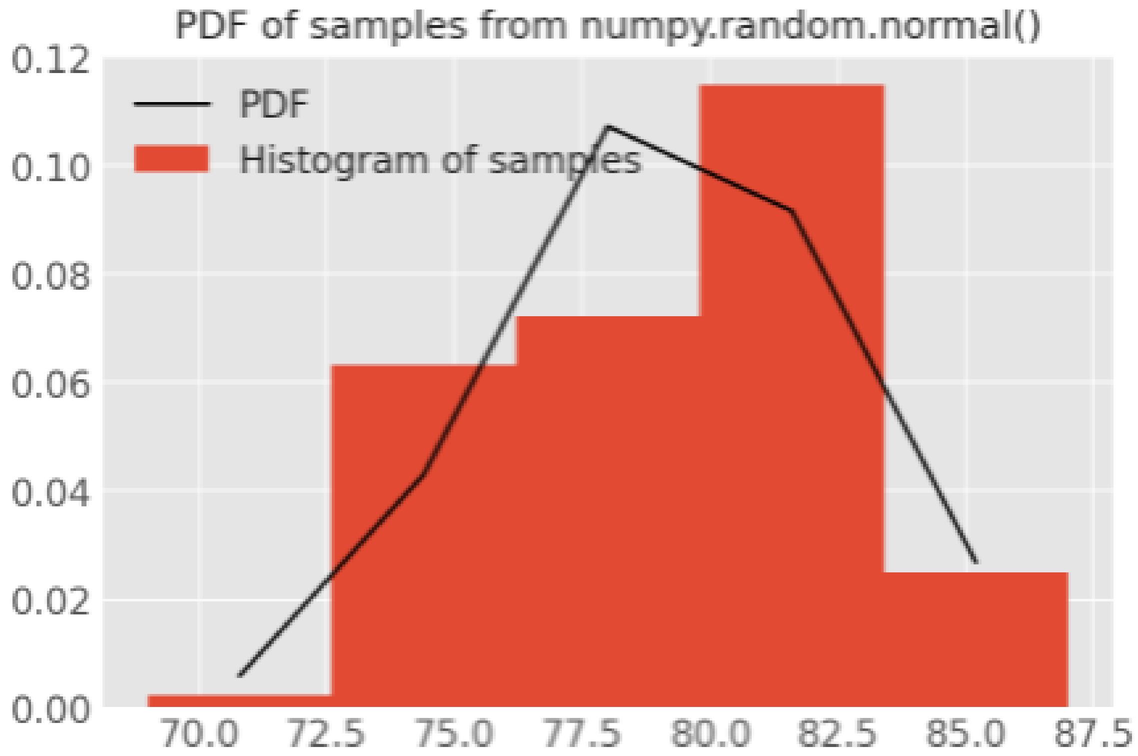
See the change of x-axis

Why theoretical PDF disappear?

Homework 2 – Can you use NBA Height data to do that same?

```
#import statistics
import seaborn as sns
paths = "E:/Data/"

# read csv file and skip the first header row so that Dataframe can be computed.
df = pd.read_csv(paths + 'nba.csv', sep=',', header=None, skiprows=[0],
                 names=["Name", "Team", "Number", "Position", "Height", "Age", "Weight", "College", "Salary"])
data=df["Height"]

sns.histplot(data=data) # works - create histogram

#players_heigth.to_numpuy()
#statistics.mean(players_height)
mu = data.mean()
sigma = data.std()
#Printing the mean
print("Mean is :", mu)
print("STDEV is :", sigma)

#For plotting
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('ggplot')

fig, ax0 = plt.subplots(ncols=1, nrows=1) #creating plot axes
(values, bins, _) = ax0.hist(data, bins=5, density=True, label="Histogram of samples")
#Compute and plot histogram, return the computed values and bins

from scipy import stats
bin_centers = 0.5*(bins[1:] + bins[:-1])
pdf = stats.norm.pdf(x = bin_centers, loc=mu, scale=sigma) #Compute probability density function
ax0.plot(bin_centers, pdf, label="PDF",color='black') #Plot PDF
ax0.legend()#Legend entries
ax0.set_title('PDF of samples from numpy.random.normal()');
```

```
Mean is : 79.27133479212254
STDEV is : 3.480003808029021
```

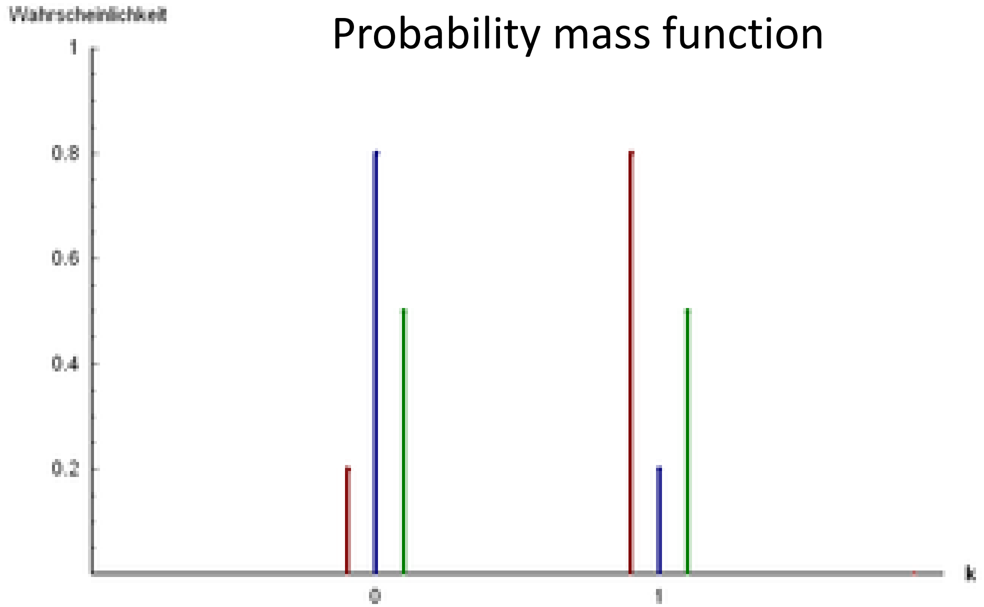


Bernoulli distribution

The Bernoulli distribution, named after Swiss mathematician Jacob Bernoulli,[1] is the **discrete** probability distribution of a random variable which takes the value 1 with probability *p* and the value 0 with probability *q = 1 - p*.

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$

Probability mass function



The expected value of a Bernoulli random variable X is: E(X) = *p*

E[X] = Pr(X=1) * 1 + Pr(X=0) * 0 = p*1 + q*0

The variance of a Bernoulli distributed X is

Var[X] = pq = p(1-p)

Three examples of Bernoulli distribution:

$P(x=0) = 0.2$ and $P(x=1) = 0.8$

$P(x=0) = 0.8$ and $P(x=1) = 0.2$

$P(x=0) = 0.5$ and $P(x=1) = 0.5$

https://en.wikipedia.org/wiki/Bernoulli_distribution

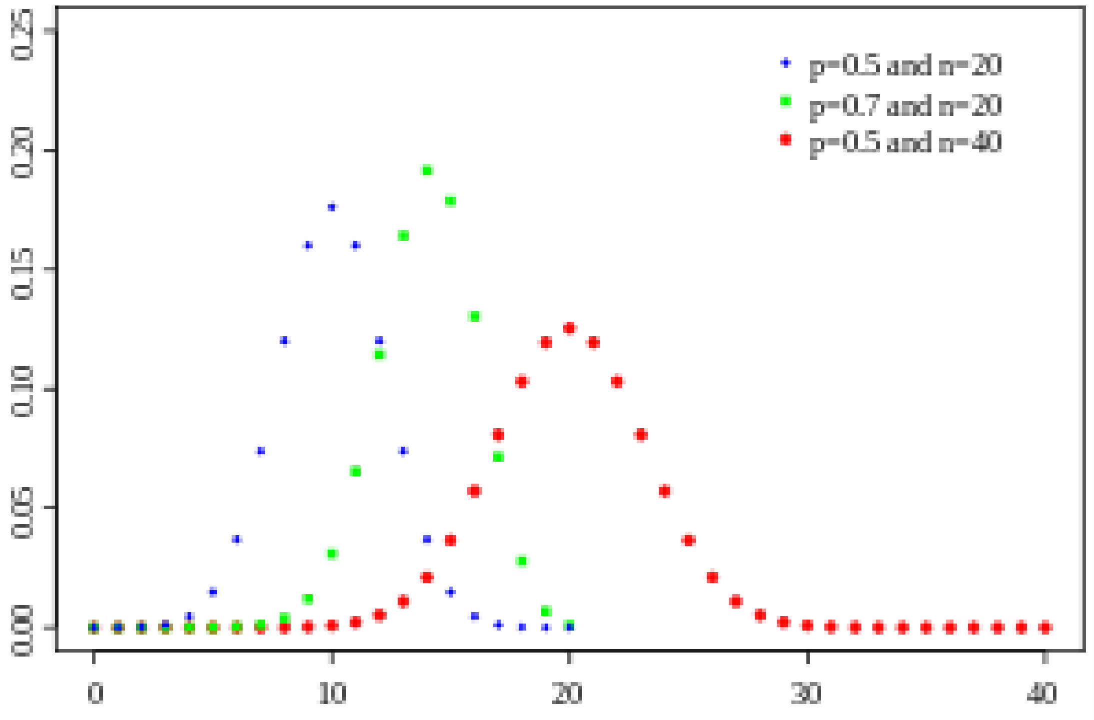
Binomial distribution

The binomial distribution with parameters ***n*** and ***p*** is the **discrete probability distribution of the number of successes** in a sequence of ***n*** independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability *p*) or failure (with probability *q* = 1 − *p*).

A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., ***n*** = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability mass function



# of successes (k)

$E[X] = np$

For n = 20 and p = 0.5

1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 1 0 0 1	8	12
0 0 1 1 0 1 0 0 0 1 1 0 1 0 1 0 1 0 1 0	9	11
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1	19	1

$Var(X) = np(1 - p)$

PDF is not a probability.

The probability density at x can be greater than one but then, how can it integrate to one?

Isn't the PDF f(x) a probability?

No. Because **f(x) can be greater than 1**.
("PD" in PDF stands for "Probability Density," not Probability.)

```
f(x) ≠ P(X = x)

* f(x): PDF for a continuous r.v.
* P(X = x) : PMF for a discrete r.v.
```

f(x) is just a height of the PDF graph at **X = x**. (Are you confused with **X** vs x notation? Check it out here.)

The whole "PDF = probability" misconception comes about because we are used to the notion of "PMF = probability", which is, in fact, correct. However, **a PDF is not the same thing as a PMF**, and it shouldn't be interpreted in the same way as a PMF, because discrete random variables and continuous random variables are not defined the same way.

For discrete random variables, we look up the value of a PMF at a single point to find its probability P(X=x) (e.g. Remember how we plugged x into the Poisson PMF?)

For continuous random variables, we take an integral of a PDF over a certain interval to find its probability that **X** will fall in that interval.

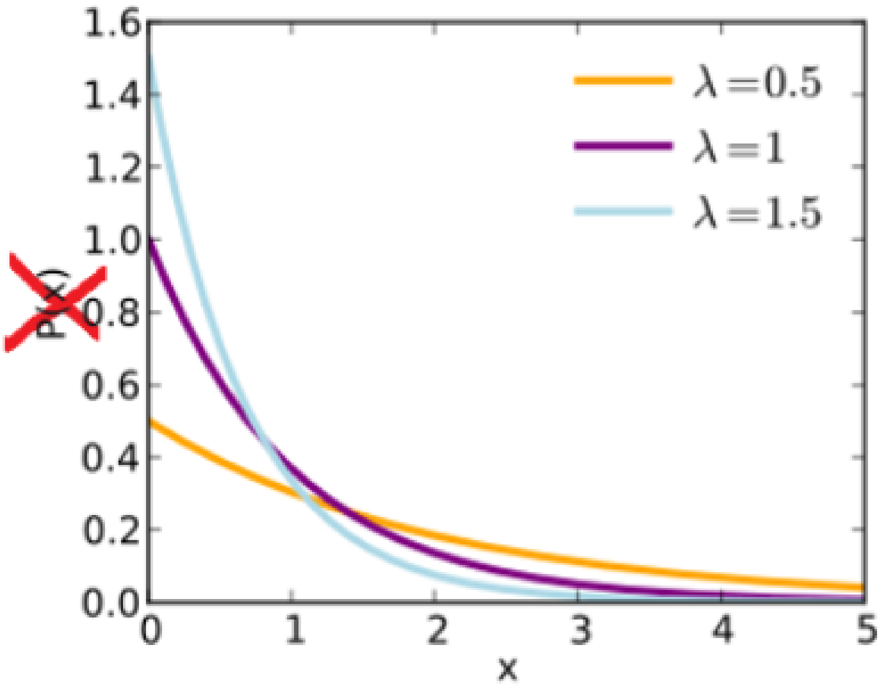
https://towardsdatascience.com/pdf-is-not-a-probability-5a4b8a5d9531

PDF is not a probability.

The probability density at x can be greater than one but then, how can it integrate to one?

4. We need to fix the Wikipedia graph of the exponential distribution. The level of Y-axis **P(X)** sounds like a probability. We need to change it to **f(x)** or **“Probability Density”**.

f(x)
or
Probability Density ←



Wikipedia: The PDF of Exponential Distribution