Probability & Statistics

Helpful link → Refer here - probability distributions

Probability

- Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true.
- The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility of the event and 1 indicates certainty.
- The higher the probability of an event, the more likely it is that the event will occur.
- Random variable is denoted as X.

Example - Dice roll - $\{1, 2, 3, 4, 5, 6\}$

- $P(X=1) = \frac{1}{6}$
- $P(X=2) = \frac{1}{6}$
- ...
- $P(X=6) = \frac{1}{6}$
- $P(X = even) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Variable

- The variable is the event that changes constantly.

Discrete Random Variable

- When there are a finite (or countable) number of such values, the random variable is discrete.
- Random variables contrast with "regular" variables, which have a fixed (though often unknown) value.
- It is a variable whose value is obtained by counting.

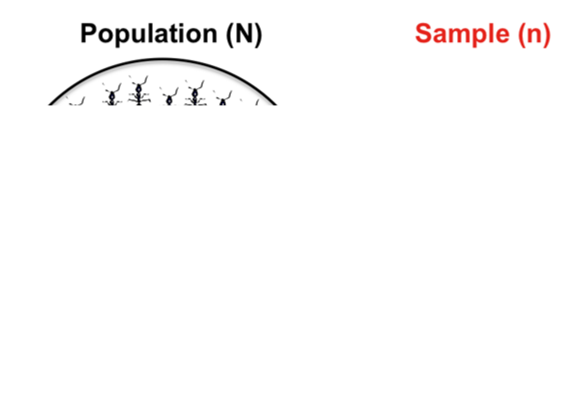
Continuous Random Variable

- A continuous variable is defined as a variable which can take an uncountable set of values or infinite set of values.
- For instance, if a variable over a non-empty range of the real numbers is continuous, then it can take on any value in that range.
- It is a variable whose value is obtained by measuring.

In [ ]:

Population and Sample

- A population is the entire group that you want to draw conclusions about.
- A sample is the specific group that you will collect (randomly) data from. The size of the sample is always less than the total size of the population.



- Mean of -
  - Population is denoted as μ
  - Sample is denoted as $\bar{x}$

Credit - Image from Internet

In [ ]:

Normal or Gaussian Distribution

- Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.
- The general form of PDF is given as -

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

- μ and σ^2 are the two important parameters used here to determine the shape of PDF.
- Let X be a random variable that follows Normal distribution whose mean is μ and variance is σ^2. We can write this in notation form as - $X \sim N(\mu, \sigma^2)$

Conclusion

- As x shifts from μ, $f(x)$ decreases i.e. $\exp(-x^2)$.
- Normal distribution plot is symmetric.

In [1]:
```
[illegible]
```

In [2]:
```
[illegible]
```

Symmetricity

- A symmetric distribution is a type of distribution where the left side of the distribution mirrors the right side.
- By definition, a symmetric distribution is never a skewed distribution.

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

$$skew = \frac{3(\mu - M)}{\sigma}$$

where -

- μ is mean
- M is median
- σ is standard deviation

In [3]:
```
[illegible]
```

In [4]:
```
[illegible]
```

Out[4]: [illegible]

Kurtosis

- It is a measure of the "tailedness" of the probability distribution of a real-valued random variable.
- Like skewness, kurtosis describes the shape of a probability distribution and there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.
- It tells whether there is a problem of outliers in the data.

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4}$$

In [5]:
```
[illegible]
        return kurt
```

In [6]:
```
# data = [1, 4, 3, 2, 5, -10, 12, 14, 10, 6]
data = [88, 85, 82, 97, 67, 77, 74, 86, 81, 95, 77, 88, 85, 76, 81]
get_kurtosis(data=data)
```

Out[6]: -0.2927119837423464

In [7]:
```
get_kurtosis(data=data, excess=False)
```

Out[7]: 2.7072880162576536

In [ ]:

Standard Normal Variate (z)

A random variable z is said to be standard normal varaite iff it has mean 0 and variance 1.

- $z \sim N(0, 1)$
- $z = \frac{(x-\mu)}{\sigma} \implies$ standardization (transformation)

In [ ]:

Central Limit Theorem

Let X be a random variable of population distribution which is not following Normal Distribution.

Let's take (m) samples from the population where each sample is of size (n) is 30.

- $S_1 \rightarrow$ random sample of n to be 30 and mean be $\bar{x_1}$
- $S_2 \rightarrow$ random sample of n to be 30 and mean be $\bar{x_2}$
- ...
- $S_m \rightarrow$ random sample of n to be 30 and mean be $\bar{x_m}$

Now $\bar{x_i} \implies \bar{x_1}, \bar{x_2}, \bar{x_3}, \ldots, \bar{x_m}$ are (m) sample means which also have a distribution.

The distribution of $\bar{x_i}$ is called the **Sampling distribution of Sample means**.

CLT says that if x has finite mean (μ) and variance (σ^2) then

the distribution of

$$\bar{x_i} \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \, as \, (n \rightarrow \infty)$$

Note

- $\mu \approx$ mean of $\bar{x_i}$
- $\frac{\sigma^2}{n} \approx$ variance of $\bar{x_i}$

In [ ]:

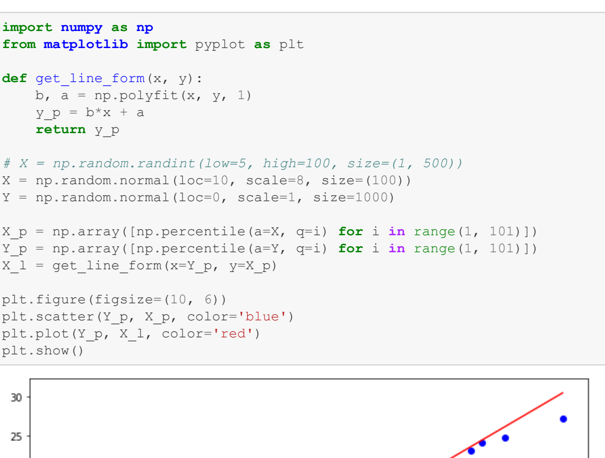
Quantile - Quantile Plot (QQ plot)

- This plot is used to determine if a random variable X is Gaussian Distributed or not by just plotting it.

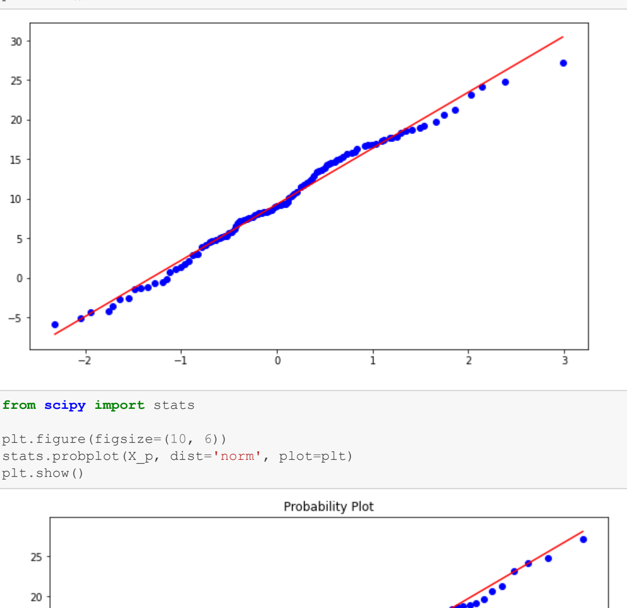
How to plot?

- Sort the data (X) & Compute the percentiles.
- Create a random variable $Y \sim N(0, 1)$, sort the values and compute percentiles.
- Plot the QQ plot using (X) percentiles and (Y) percentiles.
  - Keeping (Y) on x axis and (X) on y axis.
- If all the points roughly lie on the straight line, then the data is Normally or Gaussian distributed.

In [8]:
```
import numpy as np
from matplotlib import pyplot as plt

def get_line_form(x, y):
    b, a = np.polyfit(x, y, 1)
    y_p = b*x + a
    return y_p

# X = np.random.randint(low=5, high=100, size=(1, 500))
X = np.random.normal(loc=10, scale=8, size=(100))
Y = np.random.normal(loc=0, scale=1, size=1000)

X_p = np.array([np.percentile(a=X, q=i) for i in range(1, 101)])
Y_p = np.array([np.percentile(a=Y, q=i) for i in range(1, 101)])
X_l = get_line_form(x=Y_p, y=X_p)

plt.figure(figsize=(10, 6))
plt.scatter(Y_p, X_p, color='blue')
plt.plot(Y_p, X_l, color='red')
plt.show()
```



In [9]:
```
from scipy import stats

plt.figure(figsize=(10, 6))
stats.probplot(X_p, dist='norm', plot=plt)
plt.show()
```



In [ ]:

Chebyshev's Inequality

If X (`we don't know the distribution`) is a random variable with mean (finite) μ and standard deviation (non-zero & finite) σ then -

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$\implies P\left[(X \geq \mu + k\sigma) \, or \, (X \leq \mu - k\sigma)\right] \leq \frac{1}{k^2}$$

$$\implies P\left[(\mu - k\sigma) < X < (\mu + k\sigma)\right] \geq 1 - \frac{1}{k^2}$$

In [ ]:

Uniform Distribution

A probability distribution in which all outcomes are equally likely. In other words, all the values are equi-probable.

- Discrete uniform distribution
  - Eg - Throwing a fair dice. All the values have chance of getting selected.
- Continuous uniform distribution

Discrete Noatations

There are two parameter such as a and b.

- Notation - $U\{a, b\}$ or unif$\{a, b\}$
- $a \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- $b \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}, b \geq a$
- n (total outcomes) $= b - a + 1$
- support $k \in \{a, a+1, \ldots, b-1, b\}$
- PMF = $\frac{1}{n}$
- CDF = $\frac{\lfloor k \rfloor - a + 1}{n}$
- Mean = $\frac{a+b}{2}$
- Median = $\frac{a+b}{2}$
- Mode = NA
- Variance = $\frac{(b-a+1)^2 - 1}{12}$
- Skewness is 0

In [ ]:

Bernoulli Distribution

In probability theory and statistics, the Bernoulli distribution is a discrete probability distribution of a random variable which takes -

- the value `1` with probability p.
- the value `0` with probability $q = 1 - p$.
- parameters $\rightarrow 0 \leq p \leq 1$.

In [ ]:

Binomial Distribution

In probability theory and statistics, 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)
- failure (with probability $q = 1 - p$)
- Notation - $B(n, p)$

In [ ]:

Log Normal Distribution

The random variable X is said to be log normal iff $\log(X)$ is normally distributed.

- $X \sim \text{Log-normal}(\mu, \sigma)$

In [ ]:

How do we find the relationship between two variables?

With the help of

- Co-Variance $\rightarrow \text{cov(X, Y)}$
  - $\text{cov(X, X)} = \text{var(X)}$
  - $\text{cov(X, Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) * (y_i - \mu_y)$
- Pearson's Correlation Coefficient
  - $\frac{\text{cov(X, Y)}}{\sigma_x \sigma_y}$
- Spearman Rank Correlation
  - compute ranks (other than) values
  - apply pearson correlation for ranks values
  - much more robust than pearson's correlation

we can quantify the relationship between two variables.

> Correlation doesn't imply causation.

In [ ]:

Confidence Interval

- In statistics, a confidence interval (CI) is a type of estimate computed from the statistics of the observed data. This gives a range of values for an unknown parameter (for example, a population mean).
- The interval has an associated confidence level that gives the probability with which the estimated interval will contain the true value of the parameter.

In a Normal distribution (data) -

- if we want to compute that population mean would lie in 95% of confidence interval,
- then what is the interval? $\rightarrow [\mu - 2\sigma, \mu + 2\sigma]$

How do we estimate the C.I of μ (population mean) of a random variable?

- case 1:
  - if σ is known, then by applying CLT and Normal distribution $\rightarrow N \sim (\mu, \frac{\sigma}{\sqrt{n}})$
- case 2:
  - if σ is not known, then by applying t-distribution $\rightarrow$ using $(n-1)$ degrees of freedom

In [ ]:

Bootstrap based Confidence Interval (very important)

Let's say we have a random variable $X \sim F(\text{distribution})$. If we take a sample S of size n, how can we compute 95% confidence interval for median or standard deviation or varince? (We shall use bootstaping techniques)

- $S \rightarrow \{x_1, x_2, x_3, \ldots, x_n\} (n = 10)$
- From S, generate k sub-samples with replacement of size m where $(m \leq n)$
  - Let $k = 1000$
  - $s_1 \rightarrow \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)}\}$
  - $s_2 \rightarrow \{x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots, x_m^{(2)}\}$
  - $s_3 \rightarrow \{x_1^{(3)}, x_2^{(3)}, x_3^{(3)}, \ldots, x_m^{(3)}\}$
  - ...
  - $s_k \rightarrow \{x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_m^{(k)}\}$
- Now, compute median or standard deviation or variance for each sub-sample and store it in vector
- Sort the vector and consider the interval $[\text{vector}_{25}, \text{vector}_{975}]$ which is a 95% percent confidence interval

Non-Parametric techniques - doesn't make any assumptions about the distribution.

In [ ]:

Hypothesis Testing (Proof by Contradiction)

Let `c1` and `c2` be two classes of the data showing the heights of the students. Let μ_1 and μ_2 be the means of them respecitively.

Q1 - Is there a difference in the heights of students in `c1` and `c2` ?

Choosing a test static

- Come up with one data value that shows that the students of `c2` are taller than `c1`.
- For obvious reasons we can compare μ_1 and μ_2.

Null Hypothesis

- $H_0 \rightarrow$ Both μ_1 and μ_2 are same - no difference
  - $\mu_1 = \mu_2$

Alternate Hypothesis

- $H_1 \rightarrow$ There is strict difference between two classes
  - $\mu_1 \neq \mu_2$

Assume that H_0 is true and prove that is false and thus, reject H_0 and accept H_1.

p-value

- Tells the probability of observing $x = (\mu_2 - \mu_1)$ if the null hypothesis H_0 is true
  - if H_0 is true and p-value is (say) 0.9 $\rightarrow$ Accept H_0
  - if H_0 is true and p-value is (say) 0.05 $\rightarrow$ Reject H_0 and accept H_1

In [ ]:

Important things while designing a Hypothesis Testing

- Design of the experiment.
- Design of the H_0 - Null Hypothesis.
  - Consider the Null Hypothesis which makes the probability computation easy.
- Design of the test statistic.

In [ ]:

Hypothesis - Task

Q) Given two cities `c1` and `c2` determine if the `population means` of heights of people in these cities are the same or not.

Let

- `c1` $\rightarrow \{p_1, p_2, p_3, \dots, p_\infty\}$
- `c2` $\rightarrow \{p_1^1, p_2^1, p_3^1, \dots, p_\infty^1\}$

and

- $\mu_1 \rightarrow$ Population mean of `c1`
- $\mu_2 \rightarrow$ Population mean of `c2`

(Instead of the population mean, we will use a sample mean)

Take a random sample of size 50 from `c1` and `c2`. And compute the sample mean of each category.

Let

- `s1` $\rightarrow \{h_1, h_2, h_3, \dots, h_{50}\}$
- `s1` $\rightarrow \{h_1^1, h_2^1, h_3^1, \dots, h_{50}^1\}$

and

- $\mu_s^{(1)} \rightarrow$ Sample mean of `c1` (`s1`) which is $162cm$
- $\mu_s^{(2)} \rightarrow$ Sample mean of `c2` (`s2`) which is $167cm$

Test Statistic - $\mu_s^{(2)} - \mu_s^{(1)} = 5cm\ (x) \mid \text{observation}$

Null Hypothesis (H_0) $\rightarrow \mu_1 = \mu_2$ (no difference)

Alternative Hypothesis (H_1) $\rightarrow \mu_1 \neq \mu_2$ (there is a difference)

We shall compute $P\big((x = 5) \,|\, H_0\big) \rightarrow$ p-value

- P(obs | assumption) - obs → cannot be incorrect - assumption → can be incorrect

- Case 1 - Let p-value is 0.2 (20%) which is significant (> 0.05). Therefore, accept H_0 - There is a 20% chance of observations to have a difference of $5cm$ in sample heights of `c1` and `c2` with a sample of size 50 if there is no difference in population means.
- Case 2 - Let p-value is 0.03 (3%) which is lesser significant (< 0.05). Therefore, reject H_1

Method to compute p-value (Resampling and Permutation)

```
delta_list = []
```

1. Make a big set `S` where `s1` ∪ `s2`. - $S = \{h_1, h_2, h_3, \dots, h_{50}, h_1^1, h_2^1, h_3^1, \dots, h_{50}^1\}$
2. Create two sets `e1` and `e2` of size 50 where observations are picked at random from `S`. This is called resampling. - `e1` $\rightarrow \{h_1, h_2^1, \dots, h_{50}\}$ - `e2` $\rightarrow \{h_1^1, h_2, \dots, h_{50}^1\}$
3. Compute means for `e1` and `e2` and find the mean difference (simulated difference) between two. Save the difference value in `delta_list`.
4. Repeat **step 2** and **step 3** `k` times ($k = 1000$). - `delta_list` = $[\delta_1, \delta_2, \dots, \delta_k]$
5. Sort `delta_list` in increasing order (ascending order). - `delta_list` = $[\delta_1^1, \delta_2^1, \dots, \delta_k^1]$

Case - 1

1. Let's say we have values like -

$$\delta_1^1 \leq \delta_2^1 \leq \delta_3^1 \leq \dots \leq \delta_{800}^1 \leq 5cm \leq \dots \leq \delta_k^1$$

1. 80% of the values are $\leq 5cm$ and 20% are $\geq 5cm$.
2. p-value $\rightarrow P\big((x \geq 5) \,|\, H_0\big)$ is 20% i.e., 0.2 which is greater than 0.05. Therefore, the assumption is true (accept H_0).

Case - 2

1. Let's say we have values like -

$$\delta_1^1 \leq \delta_2^1 \leq \delta_3^1 \leq \dots \leq \delta_{970}^1 \leq 5cm \leq \dots \leq \delta_k^1$$

1. 97% of the values are $\leq 5cm$ and 3% are $\geq 5cm$.
2. p-value $\rightarrow P\big((x \geq 5) \,|\, H_0\big)$ is 3% i.e., 0.03 which is lesser than 0.05. Therefore, the assumption is false (accept H_1).

In [ ]:

Refer Python Mandatory Assignment for more details