**SELF STUDY QUESTIONS**

1. **Expand the following parentheses as far as possible for the following expressions.**
2. **E(*a*x+*b*y) for x, y variables and *a, b* as scalars**.

=E(*a*x)+E(*b*y)

=*a*E(x)+*b*E(y)

1. **E(*a*xy) for x, y independent variables and *a* a scalar.**

Cov(xy) = 0 when x and y are independent variables,

Since Cov(xy)=E(xy)-E(x)E(y)

Therefore E(xy)-E(x)E(y)=0 ;

E(xy)=E(x)E(y)

Solving for E*(a*xy):

= *a*E(xy)

= *a*E(x)E(y)

1. **E(*a*xy) for x, y correlated variables and *a* a scalar.**

Cov(xy)=1 when x and y are correlated variables,

Since Cov(xy)=E(xy)-E(x)E(y)

Therefore E(xy)-E(x)E(y)=1

E(xy)=1+E(x)E(y)

Solving for E(axy):

=aE(xy)

=1+aE(x)E(y)

1. **a) Explain the difference between PDF and a CDF.**

A PDF is the derivative of a CDF with respect to some dummy variable, while a CDF is the probability that a random variable less than or equal to x. The difference between the two is that a PDF is always non-zero for all values of x while a CDF its value is always between zero and one.

CDF:

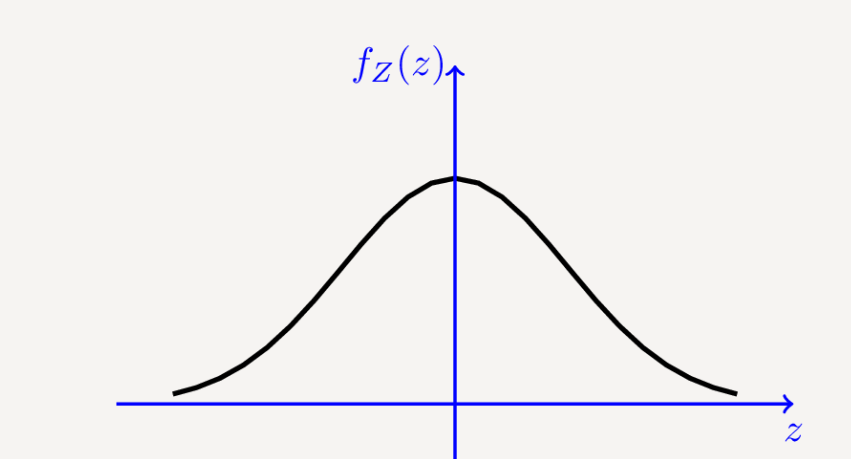
PDF: (x) = for)≥0

**b) What shapes are the PDF and CDF for a normally distributed random variable?**

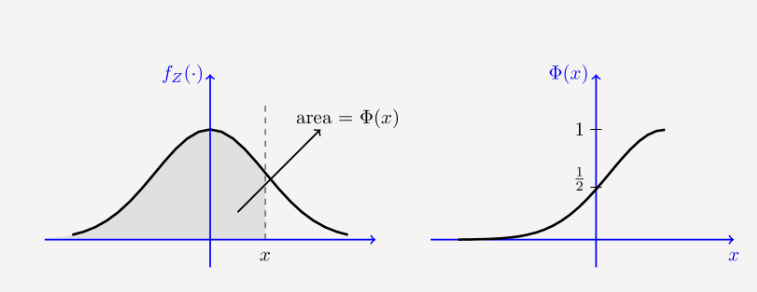
**Shape of a PDF**

A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, shown as Z∼N(0,1)Z∼N(0,1), if its PDF is given by

∈R



**Shape of a CDF**



1. **What is the central limit theorem and why is it important in statistics?**

Central limit theorem states that when repeated sampling is done on a population the shape of the distribution of the means will be known.  It allows us to understand the behavior of estimates across repeated sampling and thereby conclude if a result from a given sample can be declared to be “statistically significant,” that is, different from some null hypothesized value.

1. **Explain the difference between mean mode and median, which is the most useful measure of an average and why?**

The mean is the sum of all the values in the data set divided by the number of values in the data set. Mean as an average is easily influenced by outliers (very small or large values in the data set that are not typical).

The median is the point at which there are an equal number of data points whose values lie above and below the median value. . Thus, the median is truly the middle of the data set and the best measure of central tendency and most useful average.

Model is the most common score or the score that appears the highest number of times in a distribution.

1. **Which is a more useful measure of central tendency for stock returns - the arithmetic mean or the geometric mean? Explain your answer**.

Geometric mean is the most useful measure because it takes into account the compounding that occurs from period to period. Because of this, investors usually consider the geometric mean a more accurate measure of returns than the arithmetic mean.

* The geometric mean is most appropriate for series that exhibit [serial correlation](https://www.investopedia.com/terms/s/serial-correlation.asp). This is especially true for investment portfolios.
* Most returns in finance are correlated, including yields on bonds, stock returns, and [market risk premiums](https://www.investopedia.com/terms/m/marketriskpremium.asp). The longer the [time horizon](https://www.investopedia.com/terms/t/timehorizon.asp), the more critical [compounding](https://www.investopedia.com/articles/06/compoundingdarkside.asp) becomes, and the more appropriate the use of geometric mean.
* For volatile numbers, the geometric average provides a far more accurate measurement of the true return by taking into account year-over-year compounding.

1. **The covariance between variables is 0.99, are they strongly related? Explain your answer.**

**Yes the two variables are strongly related.**

Covariance shows how two variables vary together, while correlation quantifies how the two variables relate, thus the higher the covariance (towards 1) the higher the correlation and thus strong relationship.

1. **Explain the difference between the following pairs of terms.**
2. **Continuous and discrete data**

Discrete Data can only take certain values while Continuous Data can take any value (within a range)

1. **Ordinal and nominal data**

Nominal data simply names something without assigning it to an order in relation to other numbered objects or pieces of data. An example of nominal data might be a "pass" or "fail" classification for each student's test result.

Ordinal data involves some order; ordinal numbers stand in relation to each other in a ranked fashion. For example, suppose you receive a survey from your favorite restaurant that asks you to provide feedback on the service you received. You can rank the quality of service as "1" for poor, "2" for below average, "3" for average, "4" for very good and "5" for excellent.

1. **Time series and panel data**

Time series data - It is a collection of observations (behavior) for a single subject (entity) at different time intervals (generally equally spaced).

Panel Data (Longitudinal Data) - It is usually called as Cross-sectional Time-series data as it a combination of above mentioned types, i.e., collection of observations for multiple subjects at multiple time periods.

1. **Noisy and clean data**

Noisy data is data that cannot be understood and interpreted correctly by machines, such as [unstructured](https://searchbusinessanalytics.techtarget.com/definition/unstructured-data) text,  [data](https://searchdatamanagement.techtarget.com/definition/data) that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.

Clean data is data without corrupt or inaccurate records, incomplete, incorrect, irrelevant parts.

1. **Simple and continuously compounded returns**

Simple returns are estimate of the returns on investment, calculated simply by finding the investment profit before taxes and interest expenses. The simple rate of return is easy to calculate but is not always accurate because it considers the investment's profit rather than cash flow, it also does not account the effects of compounding.

Continuously compounded returns result when the interest earned on an investment is calculated and reinvested back into the account for an infinite number of periods. The interest is calculated on the principal amount and the interest accumulated over the given periods and reinvested back into the cash balance.

1. **Nominal and real series**

Nominal series is a series that has not been deflated for inflation while a real series is a series that has been deflated for inflation.

1. **Bayesian and classical statistics**

Bayesian statistics is used to describe uncertainty in using probability, whereby prior beliefs are updated using data to arrive to a posteriors belief and inference is made based on that posterior probability.

Classical statistics deals with parameter estimation, whereby parameters are regarded as fixed, but unknown. A parameter is estimated using data. The resulting parameter estimate is subject to uncertainty resulting from random variation in the data, known as sampling variability

1. **Present and explain a problem that can be approached using time series regression, another one using cross-sectional regression and another panel data.**

**Time series regression example: US consumption expenditure**

A time series of quarterly percentage changes (growth rates) of real personal consumption expenditure, y, and real personal disposable income, x, for the US from 1970 Q1 to 2016 Q3.

Model assumptions include:

1. The relationship between the forecast variable and the predictor variable satisfies this linear equation.
2. Second, we make the following assumptions about the errors (ε1… εT) (ε1,…, εT):

* They have mean zero; otherwise the forecasts will be systematically biased.
* They are not auto correlated; otherwise the forecasts will be inefficient, as there is more information in the data that can be exploited.
* They are unrelated to the predictor variable; otherwise there would be more information that should be included in the systematic part of the model.

The scatter plot between consumption expenditure and disposable income is used to show the relationship between the two.

**Cross-sectional regression example**

For example, we might believe that highly leveraged companies perform differently from companies with low debt, or that performance varies according to industry affiliation. In either case, we are defining an *attribute*—not a factor. The factor that causes low-debt companies to perform differently from high-debt companies most likely has something to do with interest rates. Industry affiliation, of course, measures sensitivity to factors that affect industry performance (such as military spending or competition). Once we specify a set of attributes that we feel measure sensitivity to the common sources of risk, we perform the following regression. We regress the returns across a large sample of stocks during a given period—say a month—on the attribute values of the stocks as of the beginning of that month. Then, we repeat this regression over many different periods. If the coefficients of the attribute values are not zero and are significant in a sufficiently high number of the regressions, we conclude that differences in return across the stocks relate to the differences in their attribute values.

**Panel regression example**

Using state level data on murder rates and unemployment rates in 1987, 1990, and 1993, we want to estimate the effect of unemployment on murder rate.

Panel data is useful because:

* With observations that span both time and individuals in a cross-section, more information is available, giving more efficient estimates.
* The use of panel data allows empirical tests of a wide range of hypotheses.
* With panel data we can control for : – Unobserved or unmeasurable sources of individual heterogeneity that vary across individuals but do not vary over time – omitted variable bias

1. **What are the key features of asset return time series?**

***One-Period Simple Return***: Holding the asset from one period from date t − 1 to date t would result in a simple gross return.

***Multi-period simple return***: Holding the asset for k periods between dates t − k and t gives a k-period simple gross return:

**Time interval**: Actual time interval is important in discussing and comparing returns (e.g., monthly, annual). If the time interval is not given, it is implicitly assumed to be one year.

***Continuously Compounded Returns***: The natural log of the simple gross return of an asset is called the continuously compounded return or log return.

***Portfolio Return***: The simple net return of a portfolio consisting of N assets is a weighted average of the simple net returns of the assets involved, where the weight on each asset is the fraction of the portfolio’s value investment in that asset:

1. **The following table gives annual, end of year prices of a bond and of the consumer prices index**
   1. Calculate the simple returns

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Bond value | CPI value | Simple return (Rt) |
| 2011 | 36.9 | 108.0 | - |
| 2012 | 39.8 | 110.3 | 0.0786 |
| 2013 | 42.4 | 113.6 | 0.0653 |
| 2014 | 38.1 | 116.1 | -0.1014 |
| 2015 | 36.4 | 118.4 | -0.0446 |
| 2016 | 39.2 | 120.9 | 0.0769 |
| 2017 | 44.6 | 123.2 | 0.1378 |
| 2018 | 45.1 | 125.4 | 0.0112 |