This work derives from the works of Michael Halls-Moore on Quantstart and Quantopian Lecture Series.

# Expected Value and Arithmetic mean

The expected value of a random variable is the probability-weighted average of all possible values. When these probabilities are equal, the expected value is the same as arithmetic mean, defined as the sum of the observations divided by the number of observations:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

where $X_1, X_2, \ldots, X_N$ are our observations.

For example, if a dice is rolled repeatedly many times, we expect all numbers from 1 - 6 to show up an equal number of times. So the expected value in rolling a six-sided die is 3.5.

In [1]:
```
from __future__ import print_function
from auquanToolbox.dataloader import load_data_nologs
import numpy as np
import scipy.stats as stats

# Let's say the random variables x1 and x2 have the following values
x1 = [10,9,8,5,6,7,4,3,2]
x2 = x1 + [100]

print ('Mean of x1:', sum(x1), '/', len(x1), '=', np.mean(x1))
print ('Mean of x2:', sum(x2), '/', len(x2), '=', np.mean(x2))
```

```
Mean of x1: 54 / 9 = 6.0
Mean of x2: 154 / 10 = 15.4
```

When the probabilities of different observations are not equal, i.e a random variable $X$ can take value $X_1$ with probability $p_1$, $X_2$ with probability $p_2$, and so on, the expected value of X is the same as *weighted* arithmetic mean. The weighted arithmetic mean is defined as

$$\sum_{i=1}^{n} p_i X_i$$

where $\sum_{i=1}^{n} p_i = 1$

Therefore, the expected value is the average of all values obtained you perform the experiment it represents many times. This follows from the law of large numbers - the average of the results obtained from a large number of repetitions of an experiment should be close to the expected value, and will tend to become closer as more trials are performed.

## Some properties of expected values that are handy:

- The expected value of a constant is equal to the constant itself $E[c] = c$
- The expected value is linear, i.e $E[aX + bY] = aE[X] + bE[Y]$
- If $X \leq Y$ , then $E[X] \leq E[Y]$
- The expected value not multiplicative, i.e. $E[XY]$ is not necessarily equal to $E[X]E[Y]$. The amount by which they differ is called the covariance, covered in a later notebook. $Cov(X, Y) = E[XY] - E[X]E[Y]$ If X and Y are uncorrelated, $Cov(X, Y) = 0$

## Other measures of centrality that are commonly used are:

- Median

Number which appears in the middle of the list when it is sorted in increasing or decreasing order, i.e. the value in $(n+1)/2$ when $n$ is odd and the average of the values in $n/2$ and $(n+2)/2$ positions when $n$ is even. One advantage of using median in describing data compared to the mean is that it is not skewed so much by extremely large or small values

The median uses the value that splits the data set in half, but not how much smaller or larger the other values are.

In [2]:
```
print('Median of x1:', np.median(x1))
print('Median of x2:', np.median(x2))
```

```
Median of x1: 6.0
Median of x2: 6.5
```

- Mode

Most frequently occuring value in a data set. The mode of a probability distribution is the value x at which its probability distribution function takes its maximum value.

In [3]:
```
def mode(l):
    # Count the number of times each element appears in the list
    counts = {}
    for e in l:
        if e in counts:
            counts[e] += 1
        else:
            counts[e] = 1

    # Return the elements that appear the most times
    maxcount = 0
    modes = {}
    for key in counts:
        if counts[key] > maxcount:
            maxcount = counts[key]
            modes = {key}
        elif counts[key] == maxcount:
            modes.add(key)

    if maxcount > 1 or len(l) == 1:
        return list(modes)
    return 'No mode'

print ('All of the modes of x1:', mode(x1))
```

```
All of the modes of x1: No mode
```

- Geometric mean

It is the central tendency of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the nth root of the product of n numbers:

$$G = \sqrt[n]{X_1 X_1 \dots X_n}$$

for observations $X_i \geq 0$. We can also rewrite it as an arithmetic mean using logarithms:

$$\ln G = \frac{\sum_{i=1}^{n} \ln X_i}{n}$$

The geometric mean is always less than or equal to the arithmetic mean (when working with nonnegative observations), with equality only when all of the observations are the same.

In [4]:
```
# Use scipy's gmean function to compute the geometric mean
print ('Geometric mean of x1:', stats.gmean(x1))
print ('Geometric mean of x2:', stats.gmean(x2))
```

```
Geometric mean of x1: 5.35627121246
Geometric mean of x2: 7.1775512683
```

If we have stocks returns $R_1, \ldots, R_T$ over different times, we use the geometric mean to calculate average return $R_G$ so that if the rate of return over the whole time period were constant and equal to $R_G$, the final price of the security would be the same as in the case of returns $R_1, \ldots, R_T$.

$$R_G = \sqrt[T]{(1 + R_1) \ldots (1 + R_T)} - 1$$

- Harmonic mean

The harmonic mean is less commonly used than the other types of means. It is defined as

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

As with the geometric mean, we can rewrite the harmonic mean to look like an arithmetic mean. The reciprocal of the harmonic mean is the arithmetic mean of the reciprocals of the observations:

$$\frac{1}{H} = \frac{\sum_{i=1}^{n} \frac{1}{X_i}}{n}$$

The harmonic mean for nonnegative numbers $X_i$ is always at most the geometric mean (which is at most the arithmetic mean), and they are equal only when all of the observations are equal.

In [5]:
```
print ('Harmonic mean of x1:', stats.hmean(x1))
print ('Harmonic mean of x2:', stats.hmean(x2))
```

```
Harmonic mean of x1: 4.66570664472
Harmonic mean of x2: 5.15738201465
```

The harmonic mean can be used when the data can be naturally phrased in terms of ratios.

# Variance and Standard Deviation

Variance and Standard Deviation are measures of dispersion of dataset from the mean.

We can define the mean absolute deviation as the average of the distances of observations from the arithmetic mean. We use the absolute value of the deviation, so that 5 above the mean and 5 below the mean both contribute 5, because otherwise the deviations always sum to 0.

$$MAD = \frac{\sum_{i=1}^{n} |X_i - \mu|}{n}$$

where $n$ is the number of observations and $\mu$ is their mean.

Instead of using absolute deviations, we can use the squared deviations, this is called **variance** $\sigma^2$ : the average of the squared deviations around the mean:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

**Standard deviation** is simply the square root of the variance, $\sigma$, and it is the easier of the two to interpret because it is in the same units as the observations.

Note that variance is additive while standard deviation is not.

In [7]:
```
print('Variance of x1:', np.var(x1))
print('Standard deviation of x1:', np.std(x1))
print('Variance of x2:', np.var(x2))
print('Standard deviation of x2:', np.std(x2))
```

```
Variance of x1: 6.66666666667
Standard deviation of x1: 2.58198889747
Variance of x2: 801.24
Standard deviation of x2: 28.3061830701
```

Standard deviation indicates the amount of variation in a set of data values. A low standard deviation indicates that the data points tend to be close to the expected value, while a high standard deviation indicates that the data points are spread out over a wider range of values.

## Some properties of standard deviation that are handy:

- The standard deviation of a constant is equal to 0
- Standard deviations cannot be added. Therefore, $\sigma(X+Y) \neq \sigma(X) + \sigma(Y)$
- However, variance, can be added. Infact, $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + Cov(X,Y)$
- If X and Y are uncorrelated, $Cov(X,Y) = 0$ and $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$

# Volatility

If an experiment is performed daily and the results of an experiment on one day do not affect the on their results any other day, daily observation are uncorrelated. If we measure daily standard deviation as $\sigma_i$ then we can calculate the standard deviation for an year, also called annualized standard deviation as:

$$\sigma_{ann} = \sqrt{\sum_{i=1}^{T} \sigma_i^2}$$

In finance, we sum over all trading days and this annualized standard deviation is called **Volatility**.

# These are Only Estimates

It is important to remember that when we are working with a subset of actual data, these computations will only give you sample statistics, that is mean and standard deviation of a sample of data. Whether or not this reflects the current true population mean and standard deviation is not always obvious, and more effort has to be put into determining that. This is especially problematic in finance because all data are time series and the mean and

variance may change over time. In general do not assume that because something is true of your sample, it will remain true going forward.