Arithmetic Mean

An arithmetic mean is simpler than a geometric mean as it averages out the numbers (i.e., it adds all the numbers and then divides the sum by the frequency of those numbers). Take, for example, the grades of ten students who appeared in a mathematics test.

78, 65, 89, 93, 87, 56, 45, 73, 51, 81

Calculating the arithmetic mean will mean

$$mean = \frac{78+65+89+93+87+56+45+73+51+81}{10} = 71.8$$

Hence the arithmetic mean of scores taken by students in their mathematics test was 71.8. Arithmetic mean is most suitable in situations when the observations (i.e., math scores) are independent of each other. In this case it means that the score of one student in the test won't affect the score that another student will have in the same test.

Geometric Mean

As we saw earlier, arithmetic mean is calculated for observations which are independent of each other. However, this doesn't hold true in the case of a geometric mean as it is used to calculate mean for observations that are dependent on each other. For example, suppose you invested your savings in stocks for five years. Returns of each year will be invested back in the stocks for the subsequent year. Consider that we had the following returns in each one of the five years:

60%, 80%, 50%, -30%, 10%

Are these returns dependent on each other? Well, yes! Why? Because the investment of the next year is done on the capital garnered from the previous year, such that a loss in the first year will mean less capital to invest in the next year and vice versa. So, yes, we will be calculating the geometric mean. But how? We will do so as follows:

$$[(0.6 + 1) * (0.8 + 1) * (0.5 + 1) * (-0.3 + 1) * (0.1 + 1)]^{1/5} - 1 = 0.2713$$

Hence, an investment with these returns will yield a return of 27.13% by the end of the fifth year. Looking at the calculation above, you can see that at first we first converted percentages into decimals. Next we added 1 to each of them to nullify the effects brought on by the negative terms. Then we multiplied all terms among themselves and applied a power to the resultant. The power applied was 1 divided by the frequency of observations (i.e., five in this case). In the end we subtracted the result by 1. Subtraction was done to nullify the effect introduced by an addition of 1, which we did initially with each term. The subtraction by 1 would not have been done had we not added 1 to each of the terms (i.e., yearly returns).

Median

Median is a measure of central location alongside mean and mode, and it is less affected by the presence of outliers in your data. When the frequency of observations in the data is odd, the middle data point is returned as the median.

In this chaapter we will use **statistics.median(data)** to calculate the median. This returns the median (middle value) of numeric data if frequency of values is odd and otherwise mean of the middle values if frequency of values is even using "mean of middle two" method. If data is empty, StatisticsError is raised.

Mode

Mode is suitable on data which is discrete or nominal in nature. Mode returns the observation in the dataset with the highest frequency. Mode remains unaffected by the presence of outliers in data.

Variance

Variance represents variability of data points about the mean. A high variance means that the data is highly spread out with a small variance signifying the data to be closely clustered.

1. Symbol: σ_x^2
2. Formula:
   a. $\frac{\sum(X-\bar{X})^2}{n-1}$
   b. $\sigma_x^2 = \sum(x_i - \mu_x)^2 p_i$
3. Why n-1 beneath variance calculation? *The sample variance averages out to be smaller than the population variance; hence, degrees of freedom is accounted for as the conversion factor.*
4. Rules of variance:
   i. $\sigma_{a+bx}^2 = b^2\sigma_x^2$
   ii. $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$ (If X and Y are independent variables)

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2$$

   iii. $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\boldsymbol{r}\sigma_x\sigma_y$ (if X and Y have correlation r)

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\boldsymbol{r}\sigma_x\sigma_y$$

We will be incorporating **statistics.variance(data, xbar=None)** to calculate variance in our coding exercises. This will return the sample variance across at least two real-valued numbered series.

Standard Deviation

Standard deviation, just like variance, also captures the spread of data along the mean. The only difference is that it is a square root of the variance. This enables it to have the same unit as that of the data and thus provides convenience in inferring explanations from insights. Standard deviation is highly affected by outliers and skewed distributions.

- Symbol: σ
- Formula: $\sqrt{\sigma^2}$

We measure standard deviation instead of variance because

- *It is the natural measure of spread in a Normal distribution*
- *Same units as original observations*

Changes in Measure of Center Statistics due to Presence of Constants

Let's evaluate how measure of center statistics behave when data is transformed by the introduction of constants. We will evaluate the outcomes for mean, median, IQR (interquartile range), standard deviation, and variance. Let's first start with what behavior each of these exhibits when a constant "a" is added or subtracted from each of these.

Addition: *Adding a*

- $x'_{new} = a + x'$
- $median_{new} = a + median$
- $IQR_{new} = a + IQR$
- $s_{new} = s$
- $\sigma^2_{x\,new} = \sigma^2_x$

Adding a constant to each of the observations affected the mean, median, and IQR. However, standard deviation and variance remained unaffected. Note that the same behavior will come through when observations within the data are subtracted from a constant. Let's see if the same behavior will repeat when we multiply a constant (i.e., "b") to each observation within the data.

Multiplication: *Multiplying b*

- $x'_{new} = bx'$
- $median_{new} = bmedian$
- $IQR_{new} = bIQR$
- $s_{new} = bs$
- $\sigma^2_{x\,new} = b^2\sigma^2_x$

Wow! Multiplying a constant to each observation within the data changed all five measures of center statistics. Do note that you will achieve the same effect when all observations within the data are divided by a constant term.

After going through the description of center of measures, Nancy was interested in understanding the trip durations in detail. Hence Eric came up with the idea to calculate the mean and median trip durations. Moreover, Nancy wanted to determine the station from which most trips originated in order to run promotional campaigns for existing customers. Hence Eric decided to determine the mode of 'from_station_name' field.

■ **Note** Determining the measures of centers using the statistics package will require us to transform the input data structure to a list type.

Listing 1-15. Determining the Measures of Center Using Statistics Package

```
trip_duration = list(data['tripduration'])
station_from = list(data['from_station_name'])
print 'Mean of trip duration: %f'%statistics.mean(trip_duration)
print 'Median of trip duration: %f'%statistics.median(trip_duration)
print 'Mode of station originating from: %s'%statistics.mode(station_from)
```

Output

```
Mean of trip duration: 1202.612210
Median of trip duration: 633.235000
Mode of station originating from: Pier 69 / Alaskan Way & Clay St
```

The output of Listing 1-15 revealed that most trips originated from Pier 69/Alaskan Way & Clay St station. Hence this was the ideal location for running promotional campaigns targeted to existing customers. Moreover, the output showed the mean to be greater than that of the mean. Nancy was curious as to why the average (i.e., mean) is greater than the central value (i.e., median). On the basis of what she had read, she realized that this might be either due to some extreme values after the median or due to the majority of values lying after the median. Eric decided to plot a distribution of the trip durations (see Listing 1-16) in order to determine which premise holds true.

Listing 1-16. Plotting Histogram of Trip Duration

```
data['tripduration'].plot.hist(bins=100, title='Frequency distribution of
Trip duration')
plt.show()
```

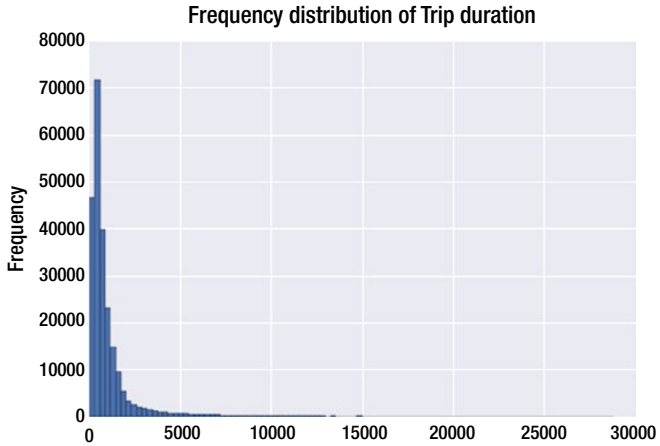

Figure 1-11. *Frequency distribution of trip duration*

The distribution in Figure 1-11 has only one peak (i.e., mode). The distribution is not symmetric and has majority of values toward the right-hand side of the mode. These extreme values toward the right are negligible in quantity, but their extreme nature tends to pull the mean toward themselves. Thus the reason why the mean is greater than the median.

The distribution in Figure 1-11 is referred to as a normal distribution.

The Normal Distribution

Normal distribution, or in other words Gaussian distribution, is a continuous probability distribution that is bell shaped. The important characteristic of this distribution is that the mean lies at the center of this distribution with a spread (i.e., standard deviation) around it. The majority of the observations in normal distribution lie around the mean and fade off as they distance away from the mean. Some 68% of the observations lie within 1 standard deviation from the mean; 95% of the observations lie within 2 standard deviations from the mean, whereas 99.7% of the observations lie within 3 standard deviations from the mean. A normal distribution with a mean of zero and a standard deviation of 1 is referred to as a standard normal distribution. Figure 1-12 shows normal distribution along with confidence intervals.

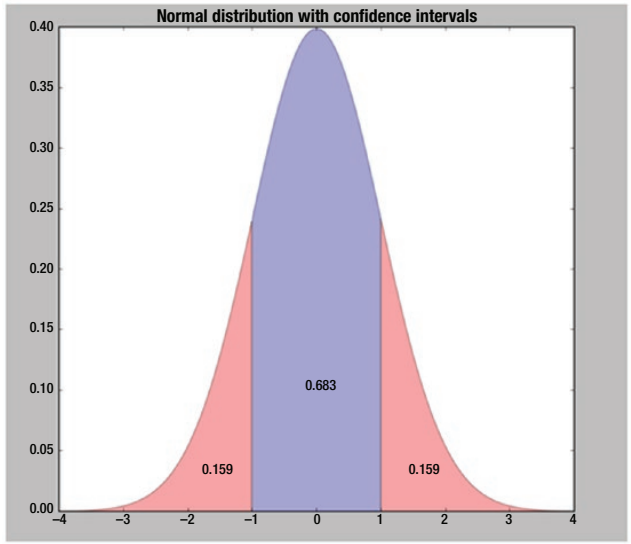

Figure 1-12. *Normal distribution and confidence levels*

These are the most common confidence levels:

Confidence level	Formula
68%	Mean ± 1 std.
95%	Mean ± 2 std.
99.7%	Mean ± 3 std.

Skewness

Skewness is a measure of the lack of symmetry. The normal distribution shown previously is symmetric and thus has no element of skewness. Two types of skewness exist (i.e., positive and negative skewness).

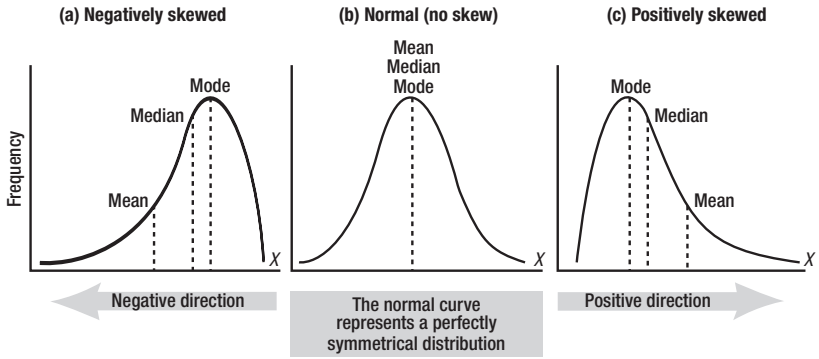

Figure 1-13. *Skewed and symmetric normal distributions*

As seen from Figure 1-13, a relationship exists among measure of centers for each one of the following variations:

- Symmetric distributions: *Mean* = *Median* = *Mode*
- Positively skewed: *Mean* < *Median* < *Mode*
- Negatively skewed: *Mean* > *Median* > *Mode*

Going through Figure 1-12 you will realize that the distribution in Figure 1-13(c) has a long tail on its right. This might be due to the presence of outliers.

Outliers

Outliers refer to the values distinct from majority of the observations. These occur either naturally, due to equipment failure, or because of entry mistakes.

In order to understand what outliers are, we need to look at Figure 1-14.

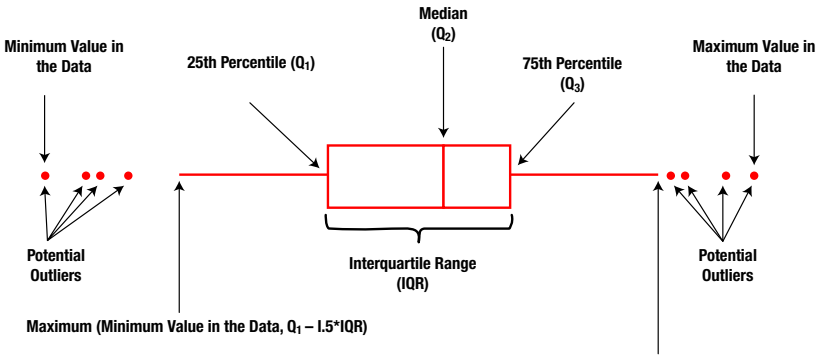

Figure 1-14. *Illustration of outliers using a box plot*

From Figure 1-14 we can see that the observations lying outside the whiskers are referred to as the outliers.

Listing 1-17. Interval of Values Not Considered Outliers

```
[Q1 - 1.5 (IQR) ,  Q3 + 1.5 (IQR) ] (i.e. IQR = Q3 - Q1)
```

Values not lying within this interval are considered outliers. Knowing the values of Q1 and Q3 is fundamental for this calculation to take place.

Is the presence of outliers good in the dataset? Usually not! So, how are we going to treat the outliers in our dataset? Following are the most common methods for doing so:

- *Remove the outliers:* This is only possible when the proportion of outliers to meaningful values is quite low, and the data values are not on a time series scale. If the proportion of outliers is high, then removing these values will hurt the richness of data, and models applied won't be able to capture the true essence that lies within. However, in case the data is of a time series nature, removing outliers from the data won't be feasible, the reason being that for a time series model to train effectively, data should be continuous with respect to time. Removing outliers in this case will introduce breaks within the continuous distribution.
- *Replace outliers with means:* Another way to approach this is by taking the mean of values lying with the interval shown in Figure 1-14, calculate the mean, and use these to replace the outliers. This will successfully transform the outliers in line with the valid observations; however, this will remove the anomalies that were otherwise present in the dataset, and their findings could present interesting insights.
- *Transform the outlier values:* Another way to cop up with outliers is to limit them to the upper and lower boundaries of acceptable data. The upper boundary can be calculated by plugging in the values of Q3 and IQR into Q3 + 1.5IQR and the lower boundary can be calculated by plugging in the values of Q1 and IQR into Q1 – 1.5IQR.
- *Variable transformation*: Transformations are used to convert the inherited distribution into a normal distribution. Outliers bring non-normality to the data and thus transforming the variable can reduce the influence of outliers. Methodologies of transformation include, but are not limited to, natural log, conversion of data into ratio variables, and so on.

Nancy was curious to find out whether outliers exist within our dataset—more precisely in the `tripduration` feature. For that Eric decided to first create a box plot (see Figure 1-15) by writing code in Listing 1-18 to see the outliers visually and then checked the same by applying the interval calculation method in Listing 1-19.

Listing 1-18. Plotting a Box plot of Trip Duration

```
box = data.boxplot(column=['tripduration'])
plt.show()
```

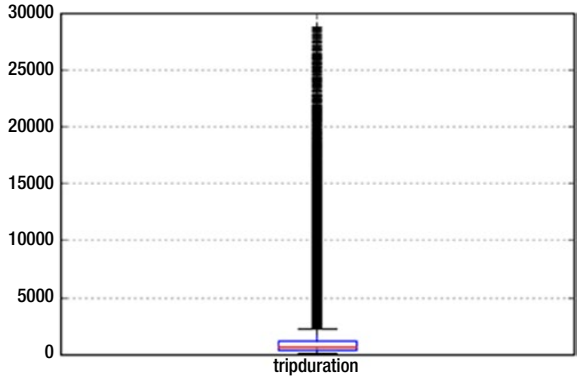

Figure 1-15. *Box plot of trip duration*

Nancy was surprised to see a huge number of outliers in trip duration from the box plot in Figure 1-15. She asked Eric if he could determine the proportion of trip duration values which are outliers. She wanted to know if outliers are a tiny or majority portion of the dataset. For that Eric wrote the code in Listing 1-19.

Listing 1-19. Determining Ratio of Values in Observations of tripduration Which Are Outliers

```
q75, q25 = np.percentile(trip_duration, [75 ,25])
iqr = q75 - q25
print 'Proportion of values as outlier: %f percent'%(
(len(data) - len([x for x in trip_duration if q75+(1.5*iqr)
>=x>= q25-(1.5*iqr)]))*100/float(len(data)))
```

Output

```
Proportion of values as outlier: 9.548218 percent
```


Eric explained the code in Listing 1-19 to Nancy as follows:

> *As seen in Figure 1-14, Q3 refers to the 75th percentile and Q1 refers to the 25th percentile. Hence we use the* `numpy.percentile()` *method to determine the values for Q1 and Q3. Next we compute the IQR by subtracting both of them. Then we determine the subset of values by applying the interval as specified in Listing 1-18. We then used the formula to get the number of outliers.*

Listing 1-20. Formula for Calculating Number of Outliers

```
Number of outliers values = Length of all values - Length of all non
outliers values
```

In our code, `len(data)` determines *Length of all values and Length of all non outliers* values is determined by `len([x for x in trip_duration if q75+(1.5*iqr) >=x>= q25-(1.5*iqr)]))`.

Hence then the formula in Listing 1-20 was applied to calculate the ratio of values considered outliers.

Listing 1-21. Formula for Calculating Ratio of Outlier Values

```
Ratio of outliers = ( Number of outliers values / Length of all values ) * 100
```

Nancy was relieved to see only 9.5% of the values within the dataset to be outliers. Considering the time series nature of the dataset she knew that removing these outliers wouldn't be an option. Hence she knew that the only option she could rely on was to apply transformation to these outliers to negate their extreme nature. However, she was interested in observing the mean of the non-outlier values of trip duration. This she then wanted to compare with the mean of all values calculated earlier in Listing 1-15.

Listing 1-22. Calculating z scores for Observations Lying Within tripduration

```
mean_trip_duration = np.mean([x for x in trip_duration if q75+(1.5*iqr)
>=x>= q25-(1.5*iqr)])
upper_whisker = q75+(1.5*iqr)
print 'Mean of trip duration: %f'%mean_trip_duration
```

Output

```
Mean of trip duration: 711.726573
```

The mean of non-outlier trip duration values in Listing 1-22 (i.e., approximately 712) is considerably lower than that calculated in the presence of outliers in Listing 1-15 (i.e., approximately 1,203). This best describes the notion that mean is highly affected by the presence of outliers in the dataset.

Nancy was curious as to why Eric initialized the variable `upper_whisker` given that it is not used anywhere in the code in Listing 1-22. Eric had a disclaimer for this: "upper_whisker is the maximum value of the right (i.e., positive) whisker i.e. boundary uptill which all values are valid and any value greater than that is considered as an outlier. You will soon understand why we initialized it over here."

Eric was interested to see the outcome statistics once the outliers were transformed into valid value sets. Hence he decided to start with a simple outlier transformation to the mean of valid values calculated in Listing 1-22.

Listing 1-23. Calculating Mean Scores for Observations Lying Within tripduration

```
def transform_tripduration(x):

    if x > upper_whisker:
        return mean_trip_duration
    return x

data['tripduration_mean'] = data['tripduration'].apply(lambda x: transform_
tripduration(x))

data['tripduration_mean'].plot.hist(bins=100, title='Frequency distribution
of mean transformed Trip duration')
plt.show()
```

Eric remembers walking Nancy through the code in Listing 1-23.

> *We initialized a function by the name of `transform_tripduration`. The function will check if a trip duration value is greater than the upper whisker boundary value, and if that is the case it will replace it with the mean. Next we add `tripduration_mean` as a new column to the data frame. We did so by custom modifying the already existing `tripduration` column by applying the `transform_tripduration` function.*

Nancy was of the opinion that the transformed distribution in Figure 1-16 is a positively skewed normal distribution. Comparing Figure 1-16 to Figure 1-10 reveals that the skewness has now decreased to a great extent after the transformation. Moreover, the majority of the observations have a `tripduration` of 712 primarily because all values greater than the upper whisker boundary are not converted into the mean of the non-outlier values calculated in Listing 1-22. Nancy was now interested in understanding how the center of measures appear for this transformed distribution. Hence Eric came up with the code in Listing 1-24.

Listing 1-24. Determining the Measures of Center in Absence of Outliers

```
print 'Mean of trip duration: %f'%data['tripduration_mean'].mean()
print 'Standard deviation of trip duration: %f'%data['tripduration_mean'].std()
print 'Median of trip duration: %f'%data['tripduration_mean'].median()
```

Output

```
Mean of trip duration: 711.726573
Standard deviation of trip duration: 435.517297
Median of trip duration: 633.235000
```

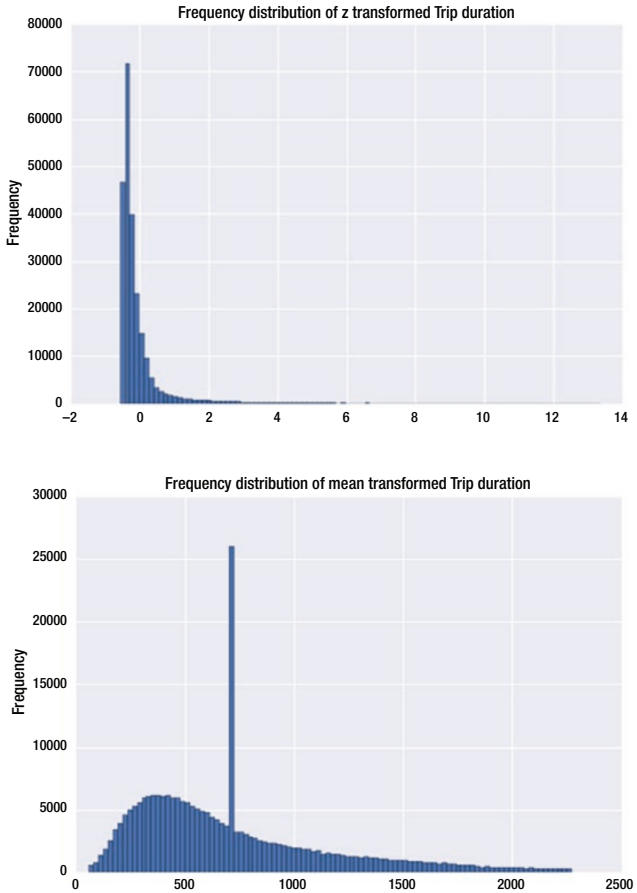

Figure 1-16. *Frequency distribution of mean transformed trip duration*

Nancy wasn't sure if the techniques they applied were obsolete or have applications in real world applications as well. Hence Eric compiled the following list of applications that his friends from the industry use to bring Statistics and Probability into live.

Applications of Statistics and Probability

Applications of statistics and probability are vibrant in several fields of study.

Actuarial Science

Actuaries use the concepts of mathematics, probability theory, statistics, finance, economics, computer science, and modeling to evaluate and price risks. Their application cases exist in the domains of insurance, consulting firms, and government.

Biostatistics

There are applications of statistics in various branches of biology. This encompasses the design of biological experiments and making inferences from them. Diving deeper into biostatistics reveals examples in which subjects (patients, cells, etc.) exhibit variation in response to some stimuli (e.g., medicine). Biostatisticians use inferential statistics to give meaning to these anomalies.

Astrostatistics

Astrostatistics is an amalgam of statistical analysis, astrophysics, and data mining. Data collected from automatic scanning of cosmos is used to make deductions.

Business Analytics

Business analytics uses operational and statistical theories to make predictive models. It also incorporates optimization techniques to garner effective insights for customers and business executives.

These insights enable companies to automate and optimize their business processes. Business intelligence differs from business analytics in that business intelligence helps us answer what happened whereas business analytics helps us understand the reason for this anomaly (i.e., why it happened in the first place and the chances of it happening again). These analytics are used in various business areas such as enterprise optimization, fraud analytics, pricing analytics, supply chain analytics, and so on.

Econometrics

The application of statistical methods for estimating economic relationships constitutes econometrics. Some of the examples include measuring the effect of divorce laws on divorce and marriage rates, change in wages of native workers from impact on immigration policies, or forecasting macroeconomic variables of inflation rates, interest rates, or gross domestic product.

Machine Learning

Several machine learning algorithms are based on statistical theories or an advanced version of the same. An example of this is the Bayesian theory which is commonly used.

Statistical Signal Processing

Past corpus of speeches is used to determine the highest probability of spoken words. Moreover, statistical signal processing is used in the following applications:

- Game theory
- Estimation and filtering
- Signal processing
- Linear systems

Elections

Campaign managers use the results of the polls to infer wins in the coming elections for their political parties.