# Statistics

## Discrete and continuous distributions

### Exponentials:

**To predict the amount of waiting time until the next event (i.e., success, failure, arrival, etc.).**

For example, we want to predict the following:

* The amount of time until the customer finishes browsing and actually purchases something in your store (success).
* The amount of time until the hardware on AWS EC2 fails (failure).
* The amount of time you need to wait until the bus arrives (arrival).

* Memoryless property( Weibull is used to beat for it has increasing hazard rate)

### Poisson Vs Exponential

If the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution.

Assuming that the time between events is not affected by the times between previous events (i.e., they are independent), then the number of events per unit time follows a Poisson distribution with the rate **λ** **=** **1/μ**.

### Gamma:

The exponential distribution predicts the wait time until the **\*very first\*** event. The gamma distribution, on the other hand, predicts the wait time until the **\*k-th\*** event occurs.

### T-test

It assumes that situations produce normal data that differ only in the sense that the average outcome in one situation is different from the average outcome of the other situation.

That being said, if we apply the t-test to data drawn from a non-normal distribution, we are probably increasing the risk of errors. Per the [Central Limit Theorem](https://en.wikipedia.org/wiki/Central_limit_theorem) (CLM), the t-test becomes more robust as the control/treatment groups become sufficiently large.

### KS test

It is a non-parametric and distribution-free test: It makes no assumption about the distribution of data. The KS test can be used to compare a sample with a reference probability distribution, or to compare two samples.

Suppose we have observations x1, x2, …xn that we think come from a distribution P. The KS test is used to evaluate:

* Null Hypothesis: The samples do indeed come from P
* Alternative Hypothesis: The samples do not come from P

Ref: <[*https://towardsdatascience.com/when-to-use-the-kolmogorov-smirnov-test-dd0b2c8a8f61*](https://towardsdatascience.com/when-to-use-the-kolmogorov-smirnov-test-dd0b2c8a8f61)>

A **non**-**homogeneous Poisson process** is similar to an ordinary **Poisson process**, except that the average rate of arrivals is allowed to vary with time. Many applications that generate random points in time are modeled more faithfully with such **non**-**homogeneous processes**.

From <[*https://www.google.com/search?q=what+is+non+homogeneous+poisson+process&rlz=1C1GCEU\_enIN883IN883&oq=what+is+homogeneous+proc&aqs=chrome.2.69i57j0l3.11672j1j1&sourceid=chrome&ie=UTF-8*](https://www.google.com/search?q=what+is+non+homogeneous+poisson+process&rlz=1C1GCEU_enIN883IN883&oq=what+is+homogeneous+proc&aqs=chrome.2.69i57j0l3.11672j1j1&sourceid=chrome&ie=UTF-8)>

### Logistic Regression

**Impact using LR**

The *odds* of an event is the probability of it happens over the probability of it doesn’t happen. For example, if the probability of an event is 0.8, the odds of the event occurring is 0.8/0.2 = 4, and that is also to say that the event will occur 4 times for every time the event does not occur and this event is 300% more likely to happen than not.

Ref:

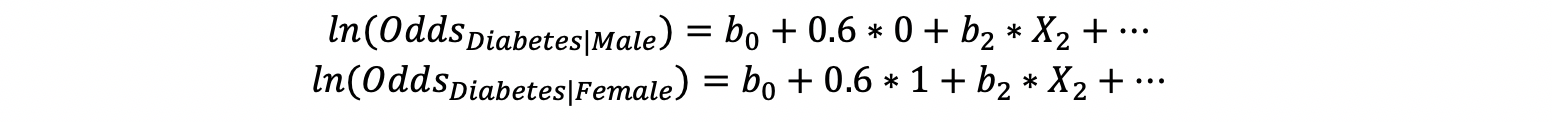
<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

<http://logisticregressionanalysis.com/817-understanding-logistic-regression-output-part-3-assessing-the-effects-of-the-x-variables/>

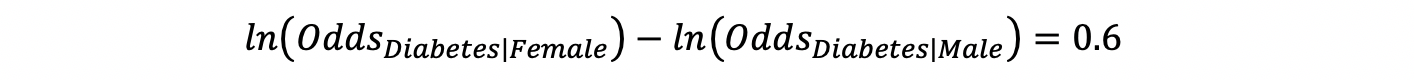
<https://medium.com/ro-data-team-blog/interpret-the-impact-size-with-logistic-regression-coefficients-5eec21baaac8>

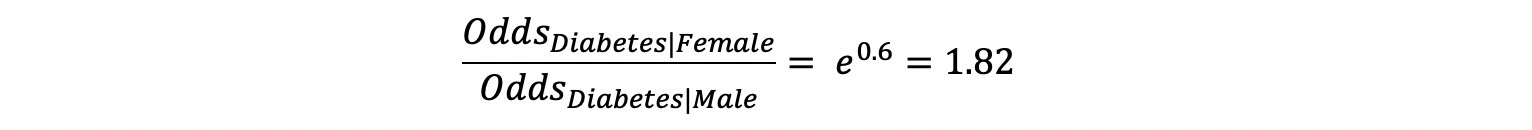
For **categorical**:

e.g. coefficient of gender is 0.6



We subtract the two equations to cancel effect:

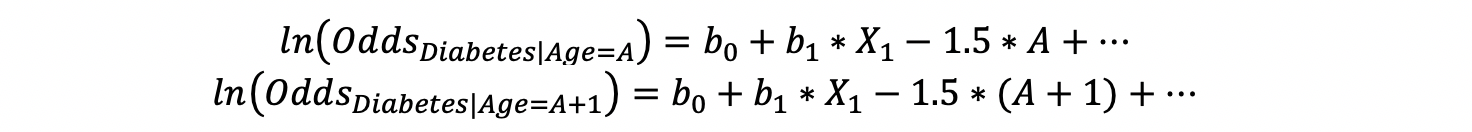


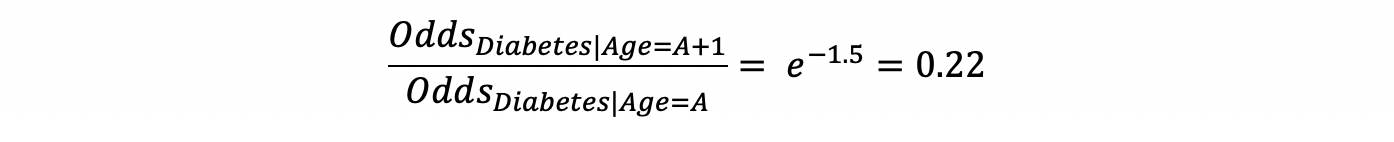


Odds of females getting diabetes over the odds of males getting diabetes is 1.82 with all the other variables fixed. In terms of percentage change, **the odds for females getting diabetes are 82% higher than the odds for male getting diabetes.**

For **continuous**:

e.g. coefficient of age is -1.5





**Holding all the other variables fixed, by increasing one year of age we expect to see the odds of getting diabetes reduce by about 78%**

## Conditional probability and Bayes theorem

## Hypothesis testing and confidence interval estimation

### Normal Table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Range** | **Expected fraction of population inside range** | **Approx Expected Frequency** | **Approximate frequency for daily event** |
| mu ±0.5sig | 0.382 924 922 548 026 | 2 in 3 | Four times a week |
| mu ±1sig | 0.682 689 492 137 086 | 1 in 3 | Twice a week |
| mu ±1.5sig | 0.866 385 597 462 284 | 1 in 7 | Weekly |
| mu ±2sig | 0.954 499 736 103 642 | 1 in 22 | Every three weeks |
| mu ±2.5sig | 0.987 580 669 348 448 | 1 in 81 | Quarterly |
| mu ±3sig | 0.997 300 203 936 740 | 1 in 370 | Yearly |
| mu ±3.5sig | 0.999 534 741 841 929 | 1 in 2149 | Every six years |
| mu ±4sig | 0.999 936 657 516 334 | 1 in 15 787 | Every 43 years (twice in a lifetime) |
| mu ±4.5sig | 0.999 993 204 653 751 | 1 in 147 160 | Every 403 years (once in the modern era) |
| mu ±5sig | 0.999 999 426 696 856 | 1 in 1 744 278 | Every 4776 years (once in recorded history) |
| mu ±5.5sig | 0.999 999 962 020 875 | 1 in 26 330 254 | Every 72 090 years (thrice in history of modern humankind) |
| mu ±6sig | 0.999 999 998 026 825 | 1 in 506 797 346 | Every 1.38 million years (twice in history of humankind) |
| mu ±6.5sig | 0.999 999 999 919 680 | 1 in 12 450 197 393 | Every 34 million years (halfivay since the extinction of dinosaurs) |
| mu ±7sig | 0.999 999 999 997440 | 1 in 390 682 215 445 | Every 1.07 billion years (a quarter of Earth's history) |

## Central Limit Theorem

If we take infinite samples from normal distribution, then distribution of the sample means is bell shaped with a mean equal to the population mean. **Sampling distribution of sample mean is always normal.**

## P Value

How do we know that the difference in two treatments is not just by chance.

It is the probability that it is as extreme as measure value. P( X>|measured value|) if the test is two sides otherwise P( X>measured value) for Ha = u>uo and P( X<measured value) for Ha = u<uo

## Random variables

## Simple Linear Regression

**Gauss Markov Assumptions**

* **Linearity:**

To detect nonlinearity one can inspect plots of observed vs. predicted values or residuals vs. predicted values.

* **Normality**
* **Independence**
* **Homoscedasticity (equal variance) of residuals**

Ref: <[*https://towardsdatascience.com/verifying-the-assumptions-of-linear-regression-in-python-and-r-f4cd2907d4c0*](https://towardsdatascience.com/verifying-the-assumptions-of-linear-regression-in-python-and-r-f4cd2907d4c0)>

Interaction in models:

* How to interpret interaction in linear regression

Y = bo+b1\*a+b2\*b +b1,2\*ab

e.g. a = Smoker/ Non Smoker; b = Asbestos exposure ; y = Stay Legth

Mean value of y will increase by b1 **for Smoker relative to Non-Smoker**

Given **They're Not Exposed to Asbestos**

Mean value of y will increase by b1+ b12 **for Smoker relative to Non-Smoker**

Given **They're Exposed to Asbestos**

**Assumptions/ In order to include interaction term into a model:**

1) It should make sense conceptually:

It should be reasonable to believe that the effect of Asbestos exposure would be different for Smokers and Non-Smokers

2) Interaction term should be statistically significant:

Can be judged by p-value, confidence interval or other forms of significance tests

Ref : <https://www.coursera.org/lecture/linear-regression-business-statistics/interaction-effects-in-a-regression-an-introduction-oLRvl>

## Generalized Linear Models

## Non-Parametric Methods

## Causal Modelling

# Foundational:

## Linear Algebra(PCA, SVD, NMF)

## Data Structures and Algorithms

## Stacks,

## heaps,

## queue,

## linked-list,

## hashing,

## tree and graph search

## Distributed Programming

## Object Oriented Programming

## Object-oriented programming (OOP) is **a programming paradigm based on the concept of "objects"**, which can contain data and code: data in the form of fields (often known as attributes or properties), and code, in the form of procedures (often known as methods).

## there are four fundamental concepts of Object-oriented programming – **Inheritance, Encapsulation, Polymorphism, and Data abstraction**.

# Supervised

## Linear Regression

## Decision Tree (CHAID, C4.5, CART)

## SVM

## Boosting

## Bagging

# Unsupervised

## Clustering

## K-Means,

## hierarchical clustering,

## EM clustering,

## Density-based clustering

## Factor Analysis

# Time Series

## Standard Forecasting(Exponential Smoothing)

## ARIMA

## FbProphet

## Forecast XGB

## VAR (Vector Autoregressive Model)

## Panel Regression

## DTW (Dynamic Time Warping)

Project EDA:

1. Seasonality Shrinkage using Fourier analysis
2. Mean Shift Analysis
3. Neighborhood analysis : Correlation amongst neighboring countries incorporated
4. Velocity Resolution(Vr) is used for People Tracking and Counting" configuration:  L / (2Tf) where L is the wavelength and Tf is the frame period.
   1. Maximum velocity ( V(m) ) = L / (4Tc) where Tc is the duration from one chirp to the other
   2. And Velocity resolution should be: ( V(r) ) = V(m) / numLoop

Ref : <https://e2e.ti.com/support/sensors-group/sensors/f/sensors-forum/978625/iwr6843-maximum-velocity-and-velocity-resolution-calculation-in-mixed-chirp-frames>

1. As market events might have diminishing effects we need to take care of them accordingly
2. Tease out effect of holidays on the data
3. Product Effects:
   1. Change in Prices,
   2. If there are any rebates,
   3. Time from product launch/ Loss of Exclusivity
   4. Market Share
   5. Stocking patterns
   6. Sales Effort (Lagged Effect)
   7. Contracting laws (with payers, distributors)

### Metrics:

1. Mean Forecast error
   1. It is known as bias and it’s value other than zero mean the tendency of the model to over or under forecast
2. Mean Absolute Error
3. Mean Squared Error:
   1. Squaring the forecast error forces, them to be positive thereby giving more weight on large errors. However in models with large forecast errors, the score will provide worst performance.
4. Root Mean squared Error

Ref : <https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/>

### Deep Learning in Time Series

Why DL to time series:

Deep learning is a subset of machine learning algorithms that learn to extract these features by representing input data as vectors and transforming them with a series of clever linear algebra operations into a given output

1. They can learn from **arbitrary mappings** from inputs to outputs
   1. Neural networks can be useful for time series forecasting problems by **eliminating the immediate need for massive feature engineering processes,** data scaling procedures, and the need for making the data stationary by differencing. Avoids differencing to make data stationary
   2. Eliminate massive data engineering
2. They support **multiple inputs and outputs**
3. They can **automatically extract patterns** in input data that spans over long sequences

\*\* Classical forecasting methods **assume** that a linear relationship and a ﬁxed temporal dependence exist among variables of a data set, and this assumption by default excludes the possibility of exploring more complex (and probably more interesting) relationships among variables.

|  |  |
| --- | --- |
| **Direct** | **Recursive** |
| a separate model is developed to forecast each forecast lead time | a single model is developed to make one-step forecasts, and the model is used recursively where prior forecasts are used as input to forecast the subsequent lead time |
| when forecasting discontiguous lead times | when forecasting a short contiguous block of lead times, |
| The direct approach may be more appropriate when we need to forecast a mixture of multiple contiguous and discontiguous lead times over a period of a few days; such is the case, for example, with air pollution forecasting problems or for anticipatory shipping forecasting, used to predict what customers want and then ship the products automatically. |  |

### Types of data:

1) univariate data, such as **lag observations** from the target variable that is being forecasted;

2) multivariate data, such as **lag observations from other variables** (for example, weather and targets in case of air pollution forecasting problems);

3) metadata, such as **data about the date or time being forecast**. Data can be drawn from across all chunks, providing a rich data set for learning a mapping from inputs to the target forecast lead time

**Basic Things**

1. CNN used for spatial data while RNN for sequence data
   1. CNN can be stacked in deep neural network
2. ML uses algorithms to parse data while DL structures algorithms in layers to create artificial NN
3. Pooling: Progressively reduces the spatial size of the representation to reduce computation

<https://www.oreilly.com/ideas/3-reasons-to-add-deep-learning-to-your-time-series-toolkit>

# Deep Learning

## Linear Perceptron Models, Back Propagation and levenberg marquardt rules

## CNN

## RNN, LSTM, GRU

## GANs

## Attention Mechanism

## Auto Encoder

# NLP

## Tokenization, stemming, stop words, lemmatization

## Information Retrieval - POS tagging, NER, TF\*IDF \*

## Text Similarity Measures (Fuzzy matching, Edit Distance, etc.) \*

## Search applications (indexing, facets, annotations) \*

## Topic detection (such as Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI))

## Text Summarization

## Deep Learning Methods - word2vec, lda2vec, doc2vec

## Speech2Text

## BERT, Transformers, GPT Models

## Language Mode Training

## QnA Chatbots

# Optimisation

## Linear Programming

## Simplex Method and Mixed Integer Programming

## Lagrange Multiples and Duality

## Gradient Descent and SGD

## Advanced optimization methods (Stochastic optimization, Robust optimization, Non-linear Programming)

## Evolutionary (Genetic Algorithms, Simulated Annealing, Tabu search, Ant Colony, ...)

## BFGS, Conjugate gradient descent, quasi-newton methods

# Graph Mode and Simulations

## Graph Clustering( modularity community detection)

## Monte Carlo Simulation

## Probabilistic Graphical Models - Markov models, Hierarchical Bayesian networks, CRF \*

## Page Rank, HITS \*

## Graph based CNN \*

## Probabilistic Circuits \*

## Graph Based Representation Learning \*

# Recommender Systems

## Memory based collaborative filtering (KNN-based) \*

## Model based collaborative filtering (Matrix factorization) \*

## Hybrid collaborative filtering \*

# Semi Supervised Learning

## Pseudo-label generation based \*

## PU learning \*

## Co-training Models \*

# Reinforcement Learning

## Model Based Learning

## Model Free learning

## Planning Network

# Solutions

## Container Based Solutions (Docker, Kubernetes, etc.)

## MLOps Frameworks (MLFlow, Seldon, Azure, etc.) \*

## CUDA programming

## Dash, streamlit

# Technologies

## Cloud Technologies (AWS, Google Cloud Platform, etc) \*

## Tableau

## Web Scraping (beautiful soup, Selenium)

## Dataiku (DSS)

## Altreyx

## HTML, CSS

## Flask

## Django

## Pytorch, FastAI

## Tensor Flow Keras

# External Tools

## Ablation Analysis

## Language Interpretability Tool (LIT) \*

## Automated Feature Engineering Feature Tools

## Automated Model Selection Auto ML

## Explanation Models (SHAP, LIME)

## Project Summaries

|  |  |
| --- | --- |
| **Business Problem** | Enhance existing demand planning team for better forecasting |
| **USP** | Created ML engine integrated with automatic dataflow from/to database, customized algorithms and error metrics  Predict SKU or Segment level demand using certain driver metrics (volatility, medical benefit etc.) and techniques such as ARIMA or AI/ML Optimize the inventory using a multi-tiered objective function with a forward looking projection derived from the SKU demand prediction |
| **Business/Client Impact** | Enhanced current process via   * automation, (automated choice of best features and algorithms) * Max accuracy (through cross validation, ensembles and multivariate model) * reduced turnaround time and * increased efficiency |
| **Data** |  |
| **Skills Learnt** |  |