1 Overview

This library implements a number of functions for mathematical operations, it has dependencies upon the scala-nlp-breeze matrix library, and its purpose is to explore the implementation and useage of a variety of tools and techniques starting with some statistical tools. This document is structured with the main headings corresponding to the subject areas implemented in the library, and minor headings corresponding to the major package names in the project. Each subsection below the minor headings is also decomposed into individual subject areas, and the lowest level heading represents a class.

2 Methods of Counting

2.1 package au.id.cxd.math.count

The count package contains a series of modules dedicated to methods of counting.

2.1.1 Factorial

The factorial operation is provided as $n!$ implementing:

$$\prod_{i=1}^{n-1}(n-i)$$

The Factorial implementation will memoize results, allowing for efficient reuse during runtime.

2.1.2 Choose

The choose module implements $\binom{n}{m}$, how many ways can m items be selected with replacement from a set of n items.
Determined as:

$$\frac{n!}{m!n!}$$

2.1.3 Permutation

The method of selecting m ordered items from a set of n ordered items $P\binom{n}{m}$.

$$\frac{n!}{(n-m)!}$$

3 Probability

3.1 package au.id.cxd.math.probability

This package provides a series of modules that support operations for inference via probability, and for estimation of distributions.

3.1.1 Inequalities

The class TchebysheffInequality implements a simple estimation of a pdf using the inequality rule:

$$P(\mu - k\sigma < Y < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Which can be restated as:

$$P(lower < Y < upper) \geq 1 - \frac{1}{k^2}$$

The value of k is derived from either upper and lower bounds since:

$$k = \frac{upper - \mu}{\sigma}$$

After determining the value of k the probability can be estimated by substituting

$$p = 1 - \frac{1}{k^2}$$

3.1.2 Discrete Distributions

The discrete distributions packages contains the following.

Binomial

The binomial distribution (class name Binomial) has the parameters p for the prior proportion of successes and n for the total number of trials and calculates the probability of y successes

$$P(y;n;p) = \sum_{i=1}^{n} \binom{n}{y_i} p_i^y (1-p)^{n-y_i}$$

The properties of the binomial are:

Mean: $\mu = np$
Variance: $\sigma^2 = np(1-p) = npq$

Geometric Distribution

The geometric distribution (class name Geometric) with 1 parameter for probability p and $y \geq 1$. The variable represents the nth trial where the success occurs (for instance if y=2 then the trial was successful on the 2nd attempt). The parameter p represents the probability of success. The probability function calculates the probability of success at the nth trial.

$$P(y;p) = p(1-p)^{y-1}$$

The simple properties of the distribution are:

Mean: $\mu = \frac{1}{p}$
Variance: $\sigma^2 = \frac{1-p}{p^2}$

Hyper Geometric Distribution

The Hypergeometric (class name HyperGeometric) distribution represents the probability of choosing y number of events of the same kind from a subset of r like items within a population of all N possible items (of different kinds) for the sample of size n containing the mixed items. The constraints are such that $r \leq n \leq N$ and $y \leq r \leq n$. The parameters are y, r, n, N. The probability distribution is defined as follows.

$$P(y;r,n,N) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

The simple properties of the distribution are:

Mean: $\mu = \frac{nr}{N}$
Variance: $\sigma^2 = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$

Negative Binomial Distribution

The Negative Binomial Distribution (class name NegativeBinomial) provides the probability of the nth success or potentially nth failure of a bernoulli trial. The parameters are r representing the (r -1) initial trials that where the successful and y the total number of trials before the next success r occurs. The distribution is calculated as follows:

$$P(y;r) = \binom{y-1}{r-1} p^r q^{y-r}$$

where $y = r, r+1, ...$

The simple properties of the distribution are:

Mean: $\mu = \frac{r}{p}$
Variance: $\sigma^2 = \frac{r(1-p)}{p^2}$

The Poisson Distribution

The Poisson Distribution (class name Poisson) provides the probability of an event occurring a certain number of times within an interval. It is commonly used to model a number of events occuring in a certain period of time. We can use the two parameters lambda and y to represent the number of events, and period of time respectively. The distribution is defined as:

$$P(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where $y \geq 0$ and $\lambda > 0$
The simple properties of the distribution are:

Mean: $\mu = \lambda$
Variance: $\sigma^2 = \lambda$

3.1.3 Functions

Package au.id.cxd.math.function

The package au.id.cxd.math.function contains a set of supporting functions.

The Gamma Function

The gamma function (class name GammaFn) provides the integral for the following:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

It is equivalent to:

$$\Gamma(\alpha) = (\alpha - 1)!$$

In order to support real numbers, it can be approximated using:

$$\Gamma(z+1) = \sqrt{2\pi}\left(z+\gamma+\frac{1}{2}\right)^{z+\frac{1}{2}} e^{-\left(z+\gamma+\frac{1}{2}\right)}\left[c_0 + \sum_{i=1}^{N}\frac{C_i}{z+i}\right]$$

Note that the approximation is minimized when $\gamma = 5$ and $N = 6$

The implementation of the gamma function is taken from Grant Palmer "Technical Java : Developing Scientific and Engineering Applications", Prentice Hall 2007 It is also described in "Numerical Recipes The Art of Scientific Computing" pp213..215. Both define c_0 and $C_{1..6}$ as constants.

The Beta Function

The beta function (class name BetaFn) supports the beta distribution and is built upon the gamma function in the following manner:

$$\beta(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

and is implemented as:

$$\beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

3.1.4 Continuous Distributions

Package au.id.cxd.math.probability.continuous

The package au.id.cxd.math.probability.continuous contains a set of classes that can be used to model continuous distributions and their simple properties.

Uniform Distribution

The uniform distribution (class name Uniform) provides a uniform output for y between a *min* and *max* and is implemented as follows.

$$f(y;min,max) = \begin{cases} \frac{1}{max-min} & \text{if } min \leq y \leq max \\ 0 & \text{otherwise} \end{cases}$$

The distribution has the following simple properties:
Mean: $\mu = \frac{max+min}{2}$
Variance: $\sigma^2 = \frac{(max-min)^2}{12}$

Gamma Distribution

The Gamma distribution (class name Gamma) is suitable for modelling distributions that are left skewed and is useful for things such as time measurements (similar to the Poisson distribution for discrete events). The distribution is defined as:

$$f(y;\alpha,\beta) = \begin{cases} \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & \text{for } 0 \leq y \leq \infty \\ 0 & \text{elsewhere} \end{cases}$$

The simple properties of the distribution are:
Mean: $\mu = \alpha\beta$
Variance: $\sigma^2 = \alpha\beta^2$
The Gamma distribution also has an important relationship with the chi-square distribution where ν is the degrees of freedom, if the gamma distribution has the parameters $\alpha = \nu/2$ and $\beta = 2$ then it is also a χ^2 distribution with ν degrees of freedom.

Exponential Distribution

The Exponential distribution (class name Exponential) is an instance of the Gamma distribution where the α parameter equals 1, leaving the parameter β to be provided. The distribution can be used to model processes with "lifetimes" or decay. The distribution is defined as:

$$f(y;\beta) = \begin{cases} \frac{1}{\beta}e^{-y/\beta}, & 0 \leq y \leq \infty \\ 0 & \textit{elsewhere} \end{cases}$$

The exponential distribution has the following simple properties:
Mean: $\mu = \beta$
Variance: $\sigma^2 = \beta^2$

Beta Distribution

The Beta distribution (class name Beta) is often used to model proportions over the range of 0 to 1. It has two parameters, $\alpha > 0$ and $\beta > 0$, the domain of the dependent variable is $0 \leq y \leq 1$, however it is possible to scale variables to fit within this range for use with the beta distribution. The probability density function is modelled as:

$$f(y;\alpha,\beta) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\beta(\alpha,\beta)} & 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Note the $\beta(\alpha,\beta)$ is the Beta function described in "The Beta Function" earlier in this document. The distribution has the following simple properties:
Mean: $\mu = \frac{\alpha}{\alpha+\beta}$
Variance: $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Normal Distribution

The normal distribution is commonly used for modelling natural processes and due to the central limit theorem is useful for inferences with a large enough sample and

population. It accepts two parameters μ for the mean and σ^2 for the variance. The normal distribution is defined as follows:

$$f(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \quad -\infty \leq y \leq \infty$$

The simple properties of the distribution are as follows:
Mean: $E[Y] = \mu$
Variance: $Var[Y] = \sigma^2$

Chi-square Distribution

The chi-square χ^2 distribution with degree of freedom df is implemented as a Gamma distribution in the class "ChiSquare" with the distribution having the parameters $\alpha = df/2$ and $\beta = 2$:

$$\chi^2_{df} = f(y;\alpha = df/2, \beta = 2) = \begin{cases} \frac{y^{df/2-1}e^{-y/2}}{2^{\alpha}\Gamma(df/2)} & \text{for } 0 \leq y \leq \infty \\ 0 & \text{elsewhere} \end{cases}$$

F Distribution

The F distribution is the ration of two chi-square distributions divided by their respective degrees of freedom.

$$f(x;d_1;d_2) = \frac{\chi^2_{d_1}/d_1}{\chi^2_{d_2}/d_2}$$

It is defined as:

$$f(x;d_1,d_2) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right) d_1^{d_1/2} d_2^{d_2/2}}{\Gamma(d_1/2)\Gamma(d_2/2)} \frac{x^{d_1/2-1}}{(d_1+d_2x)^{(d_1+d_2)/2}}$$

$$= \frac{d_1^{d_1/2} d_2^{d_2/2} x^{d_1/2-1}}{(d_2+d_1x)^{(d_1+d_2)/2} B(d_1/2,d_2/2)}$$

$$= \frac{\sqrt{d_2^{d_2}}\sqrt{d_1^{d_1}}\sqrt{x^{d_1-2}}}{\sqrt{(d_2+d_1x)^{d_1+d_2}}\beta(d_1/2,d_2/2)}$$

where β is the beta function (from the gamma function) and $x > 0$
The mean of the distribution is defined where denominator $df > 2$.

$$\mu = \frac{d_2}{d_2-2}$$

And the variance of the distribution is defined where the denominator $df > 4$

$$\sigma^2 = \frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$$

The f distribution is most often used in testing hypothesis about an unknown variance.

3.1.5 Inference

Analysis of Variance (Anova)

The ratio of Mean Square of Treatment (MST) and Mean Square of Error (MSE) is used in the test of the hypothesis that the means of k normally distributed populations are equal $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ assuming that variances are also equal, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$. The alternate hypothesis states that at least one of the population means are not equal to the others. The test statistic is calculated as

$$F = \frac{MST}{MSE}$$

and the rejection region is given for the critical value of the F distribution at $k-1$ and $n-k$ degrees of freedom.

$$F > F_{\alpha,(k-1),(n-k)}$$

The procedure of calculating the test statistic involves the calculation of the Total Sum of Square Treatment and the Sum of Square Error.
Firstly the correction for the mean is calculated (CM).

$$CM = \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)^2 = n\bar{Y}^2$$

The total sum of squares is calculated.

$$TotalSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - CM$$

The Sum of square treatment is calculated as:

$$SST = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet}-\bar{Y})^2 = \sum_{i=1}^{k}\frac{Y_{i.}^2}{n_i} - CM$$

where $Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y}_{i\bullet} = \left(\frac{1}{n_i}\right)\sum_{j=1}^{n_i} Y_{ij} = \left(\frac{1}{n_i}\right) Y_{i\bullet}$.
The Sum of square error is calculated from the Sum of Square treatment:

$$SSE = TotalSS - SST$$

Note that the SSE is also equal to the pooled sum of squares for all samples:

$$SSE = \sum_{i=1}^{k}(n_i-1)S_i^2$$

where

$$S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2$$

The mean squared error is the pooled variance defined as:

$$S^2 = MSE = \frac{SSE}{n-k}$$

and the mean sum of square treatment is the sum of square treatment divided by the degree of freedom $k-1$.

$$MST = \frac{SST}{k-1}$$

The implementing class "Anova" performs the calculation for a multivariate set of k samples by breaking the calculations into their components as described above. Note that the operation is performed on a matrix, hence for samples of unequal sizes, a sparse matrix is used where empty values for samples of smaller sizes are padded with 0*s*, the total number of rows is the maximum n_i of the k samples.

3.1.6 Regression