

# Stats and Probability Information

May 11, 2014

## Continuous and marginal distributions

The marginal distribution of  $x$  in a two-variable distribution is equal to the sum of the joint distribution over  $y$ .

$$Pr(X = x) = \sum_y Pr(X = x, Y = y) = \sum_y Pr(X = x|Y = y)Pr(Y = y) \quad (1)$$

From [Wikipedia](#)

For the continuous case

$$p_X(x) = \int_y p_{X,Y}(x, y)dy = \int_y p_{X|Y}(x|y)p_Y(y)dy \quad (2)$$

There are three related distributions: the marginal, the joint and the conditional.

## 0.1 Mixture Model

This is a probabilistic model that relates some random variables to some other variables. The model has sub-populations. The properties of the sub-population are different from those of the parent. The sub-populations may not be observable. For example, the distribution of returns may be different in different sub-population or regime.

A *mixture distribution* is the probability distribution of a random variable whose values are derived from an underlying set of random variables. The *mixture components* are individual distributions with *mixture weights*. Even in cases where the mixture components have a normal distribution, the mixture distribution is likely to be non-normal. Mixture models are used to understand the sub-population when there is only access to the information about the pooled population.

The mixture model will be comprised of N random variables distributed according to K components, with each component belonging to the same distribution. The k mixture weights sum to one. Each component will have parameters (mean and variance in the case of normal distribution).

The method will try to estimate all the parameters of the model from the data. The underlying data is known ( $x_i$ ); the number of mixture components is set ( $K$ ); the parameters of the distribution of each mixture component ( $\theta_{i=1...K}$ ); mixture weight ( $\Phi_{i=1...K}$ );  $\Phi$  K-dimensional vector summing to 1;  $F(x|\theta)$  probability distribution of observations parameterised on  $\theta$ ;  $\alpha$  shared hyperparameter for component weights;  $\beta$  shared hyperparameter for mixture weights;  $H(\theta|\alpha)$  prior probability distribution of component parameters;

## 0.2 Adjusted R squared

**Adjusted R squared** applied a penalty to the basic R squared to account for additional variables. The equation is

$$R_A^2 = 1 - \left[ \frac{(n-1)}{(n-k)} \right] [1 - R^2] \quad (3)$$

Adding a regressor to the equation will increase (reduce) the  $R_A^2$  when the absolute value of the t-statistic is greater (less) than one. Adding a group of regressors to the model will reduce (increase) the  $R_A^2$  when the absolute value of the F-statistic is greater than one.

Proof <http://davegiles.blogspot.com/2014/04/proof-of-result-about-adjusted.html>

## 0.3 Monte Carlo Simulation

This comes from **Revolutionary Analytics**. The analysis is in annual terms.

$$\mu\Delta t + \sigma Z\sqrt{\Delta t} \quad (4)$$

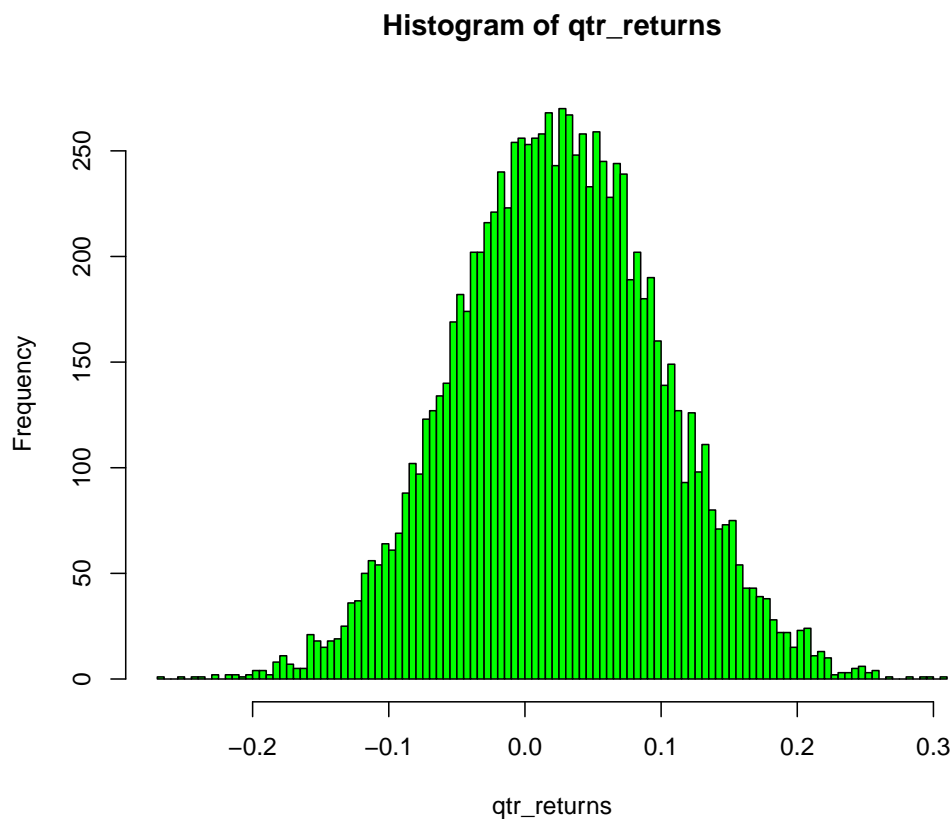
where  $\mu$  is the drift or average annual return, Z is a standard Normal random variable, t is measured in years so for monthly returns  $\Delta t$  equals  $\frac{1}{12}$ .

```
n <- 10000
# Fixing the seed gives us a consistent set of simulated returns
set.seed(106)
z <- rnorm(n) # mean = 0 and sd = 1 are defaults
mu <- 0.1
sd <- 0.15
```

```

delta_t <- 0.25
# apply to expression (*) above
qtr_returns <- mu * delta_t + sd * z * sqrt(delta_t)
hist(qtr_returns, breaks = 100, col = "green")

```



Now the descriptive statistics can be uncovered from the simulated results.

```

stats <- c(mean(qtr_returns) * 4, sd(qtr_returns) * 2) # sqrt(4)
names(stats) <- c("mean", "volatility")
stats

##      mean volatility
## 0.09901    0.14976

```

This is the basic model. It would also be possible to simulate two variables and to include some relationship between the two in the analysis. It would also be possible to simulate an asset in two different regimes. A Monte-Carlo

Markov Model (MCMM) would require another set of  $\mu$  and  $\sigma$  inputs as well as a transition matrix of the probabilities that there is a switch from one regime to another.

## 0.4 Generalised Lambda Distribution

This is from [Revolutionary Analytics](#). The four parameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  indicate the location, scale, skew and kurtosis of the distribution.

```
require(GLDEX)
require(quantmod)

getSymbols("SPY", from = "1994-02-01")

## [1] "SPY"

SPY.Close <- SPY[, 4] # Closing prices

SPY.vector <- as.vector(SPY.Close)

# Calculate log returns
sp500 <- diff(log(SPY.vector), lag = 1)
sp500 <- sp500[-1] # Remove the NA in the first position
# Set normalise='Y' so that kurtosis is calculated with reference to
# kurtosis = 0 under Normal distribution
fun.moments.r(sp500, normalise = "Y")

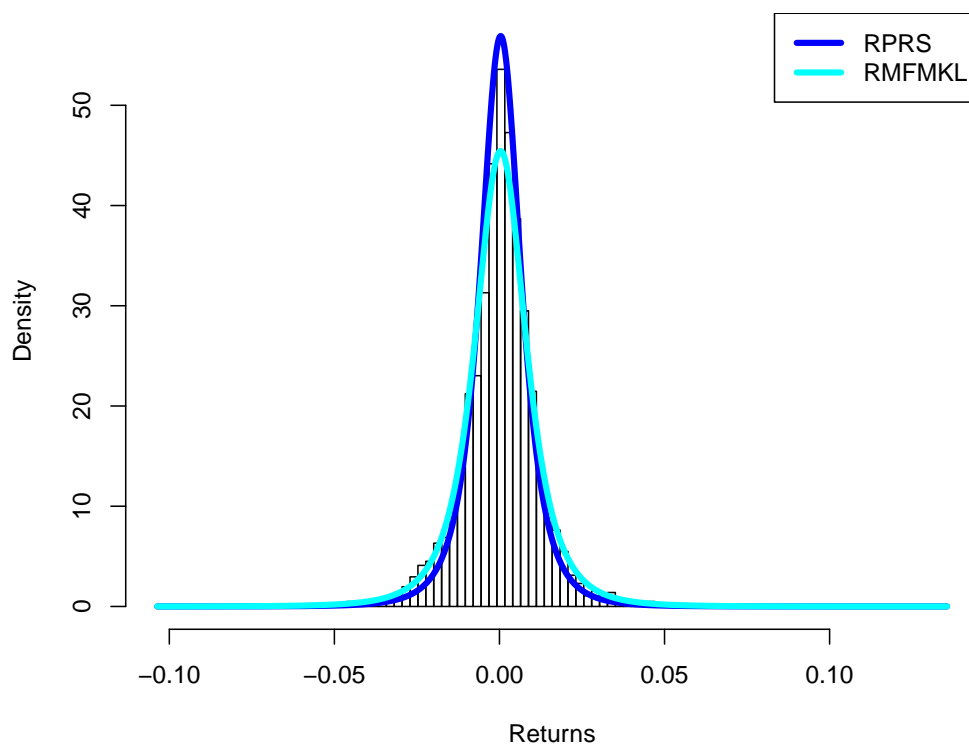
##          mean    variance    skewness    kurtosis
## 0.0002664 0.0001528 -0.0969262 9.5433262
```

Now fit the GLD with the function `fun.data.fit.mm`. There are warnings but these can be ignored.

```
spLambdaDist = fun.data.fit.mm(sp500)
spLambdaDist

##          RPRS      RMFMKL
## [1,] 3.884e-04 3.244e-04
## [2,] -4.238e+01 2.039e+02
## [3,] -1.677e-01 -1.699e-01
## [4,] -1.641e-01 -1.616e-01
```

```
fun.plot.fit(fit.obj = spLambdaDist, data = sp500, nclass = 100, param = c("rs"  
"fmkl"), xlab = "Returns")
```



Now it is possible to generate simulated results using the function `rgl()`. Lambdas need to be identified.

```
lambda_params_rs <- spLambdaDist[, 1]  
lambda1_rs <- lambda_params_rs[1]  
lambda2_rs <- lambda_params_rs[2]  
lambda3_rs <- lambda_params_rs[3]  
lambda4_rs <- lambda_params_rs[4]  
  
lambda_params_fmkl <- spLambdaDist[, 2]  
lambda1_fmkl <- lambda_params_fmkl[1]  
lambda2_fmkl <- lambda_params_fmkl[2]  
lambda3_fmkl <- lambda_params_fmkl[3]  
lambda4_fmkl <- lambda_params_fmkl[4]
```

Now generate simulations of each variety.

There are problems with the `rgl` function. I am not sure what this does. It is 10 million simulations. I think that the `rgl` just uses extra hardware to make the change. It may be useful to re-do this last section using a different method.

```
require(gld)

## Loading required package: gld

require(GLDEX)

## Loading required package: GLDEX
## Loading required package: cluster

# RS version:
set.seed(100) # Set seed to obtain a reproducible set
rs_sample <- rgl(n = 1e+07, lambda1 = lambda1_rs, lambda2 = lambda2_rs, lambda3 = lambda3_rs,
  lambda4 = lambda4_rs, param = "rs")

# Moments of simulated returns using RS method:
fun.moments.r(rs_sample, normalise = "Y")

##          mean    variance  skewness  kurtosis
## 2.664e-04  9.759e-05 -1.052e-01  9.987e+00

# Moments calculated from market data:
fun.moments.r(sp500, normalise = "Y")

##          mean    variance  skewness  kurtosis
## 0.0002664  0.0001528 -0.0969262  9.5433262

# FKML version:
set.seed(100) # Set seed to obtain a reproducible set
fmkl_sample <- rgl(n = 1e+05, lambda1 = lambda1_fmkl, lambda2 = lambda2_fmkl,
  lambda3 = lambda3_fmkl, lambda4 = lambda4_fmkl, param = "fmkl")

# Moments of simulated returns using FKML method:
fun.moments.r(fmkl_sample, normalise = "Y")
```

```
##          mean    variance    skewness    kurtosis
## 0.0002434 0.0001542 -0.0870419  8.6061100

# Moments calculated from market data:
fun.moments.r(sp500, normalise = "Y")

##          mean    variance    skewness    kurtosis
## 0.0002664 0.0001528 -0.0969262  9.5433262
```

Compare the moments to the S&P500 market data

```
fun.moments.r(rs_sample, normalise = "Y")

##          mean    variance    skewness    kurtosis
## 2.664e-04 9.759e-05 -1.052e-01  9.987e+00

fun.moments.r(sp500, normalise = "Y")

##          mean    variance    skewness    kurtosis
## 0.0002664 0.0001528 -0.0969262  9.5433262

fun.moments.r(fmkl_sample, normalise = "Y")

##          mean    variance    skewness    kurtosis
## 0.0002434 0.0001542 -0.0870419  8.6061100

fun.moments.r(sp500, normalise = "Y")

##          mean    variance    skewness    kurtosis
## 0.0002664 0.0001528 -0.0969262  9.5433262
```

## 0.5 Lasso method

**Rob Tibshirani** Cancer example that requires identification of appropriate cell. There are 20 cases that are being used as a training set. Train a classifier to identify whether the cells are cancerous or not. There are 11,000 features. It would be useful to use as few of the features as possible. Therefore, also want to know which features are important for the classification.

Sparcity means that the features are reduced by only using those that pass a particular level of significance. [More here.](#)

## 0.6 Standard Error of the Regression

The standard error of the estimate of the mean is

$$SD_x = \frac{\sigma}{\sqrt{n}} \quad (5)$$

This can be derived from the variance of the sum of independent random variables

- If  $X_1, X_2, \dots, X_n$  are independent observations from a population with a mean  $\mu$  and a standard deviation  $\sigma$ ,
- Variance of the Total =  $T = (X_1, X_2, \dots, X_n)$  is  $n\sigma^2$
- the variance of  $T/n$  is  $\frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$
- the standard deviation of  $T/n$  must be  $\frac{\sigma}{\sqrt{n}}$
- $T/n$  is the mean  $\bar{x}$

## 0.7 Logistic Regression

This comes from [Wim-Vector - logistic regression](#).

```
CarData <- read.table(url("http://archive.ics.uci.edu/ml/machine-learning-databases/car/"),
  sep = ",", col.names = c("buying", "maintenance", "doors", "persons", "lug_boot",
    "safety", "rating"))
logisticModel <- glm(rating != "unacc" ~ buying + maintenance + doors + persons +
  lug_boot + safety, family = binomial(link = "logit"), data = CarData)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(logisticModel)

##
## Call:
## glm(formula = rating != "unacc" ~ buying + maintenance + doors +
##     persons + lug_boot + safety, family = binomial(link = "logit"),
##     data = CarData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.216   0.000   0.000   0.026   2.434
```



```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -28.426   1257.526  -0.02    0.98
## buyinglow      5.048     0.567    8.90 < 2e-16 ***
## buyingmed      3.922     0.484    8.10 5.5e-16 ***
## buyingvhigh   -2.066     0.375   -5.51 3.5e-08 ***
## maintenancelow  3.406     0.469    7.26 3.9e-13 ***
## maintenancemed  3.406     0.469    7.26 3.9e-13 ***
## maintenancenvhigh -2.825     0.415   -6.82 9.4e-12 ***
## doors3         1.856     0.404    4.59 4.4e-06 ***
## doors4         2.482     0.428    5.80 6.6e-09 ***
## doors5more     2.482     0.428    5.80 6.6e-09 ***
## persons4       29.965   1257.526    0.02    0.98
## personsmore    29.584   1257.526    0.02    0.98
## lug_bootmed    -1.517     0.376   -4.04 5.4e-05 ***
## lug_bootsmall  -4.448     0.475   -9.36 < 2e-16 ***
## safetylow     -30.505   1300.343   -0.02    0.98
## safetymed     -3.004     0.358   -8.40 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2110.47  on 1727  degrees of freedom
## Residual deviance:  339.36  on 1712  degrees of freedom
## AIC: 371.4
##
## Number of Fisher Scoring iterations: 21
```

The variable levels and values are joined together. The model will provide an estimate of the effect of each category on the rating. The values for each category are added together to get the overall rating score.

To be completed.

## 0.8 F-Tests

This is an example that is based on testing for fabricated evidence. The full documentation is [Deborah Mayo](#). As the paper says

“In each experiment, participants (undergraduate students)

were randomly assigned to three groups, and each group was given a different intervention. All participants were then tested on some outcome measure.“

The accusation is that the results are **too** linear. The relationships are too perfect compared to other studies. The method used to asses this is called *delta-F*. The odds of seeing such linear trends are calculated based on the assumption that the trend is linear.

Unless there is a huge sample, the probability of obtaining a linear trend is very low becuase there is noise. The amount of noise is evident in the *within group variance*. For a given sample size and a given level of within group variance, the odds of obtaining a linear trend are calculated as the sum of squares accounted for by a linear model and a none-linear model (one-way ANOVA) divided by the mean-square error (within group variance).

$$\Delta F = \frac{SS_{REG} - SS_B}{MS_W} \quad (6)$$

There is one degree of freedom in the numerator and  $3(n - 1)$  degrees of freedom in the denominator.

If the difference between the two models (linear and non-linear) is small, it means that there is an underlying linear relationship. Assuming that the relationship is linear, this delta-F metric should follow an F distribution.

This is more or less the same method that is used to test whether a simple model fits the data as well as a complex model. In that case, the null hypothesis is that the simple model is the correct version and the objective is to determine if the difference between the two is unlikely given the null.

From the paper

“But here the whole thing is turned on its head. Random noise means that a complex model will sometimes fit the data better than a simple one, even if the simple model describes reality. In a conventional use of F-tests, that would be regarded as a false positive. But in this case its the absence of those false positives thats unusual.“