Stats and Probability Information

Rob Hayward

May 26, 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.

$$Pr(X = x) = \sum_y Pr(X = x, Y = y) = \sum_y Pr(X = x|Y = y)Pr(Y = y) \tag{1}$$

From Wikipedia
For the continuous case

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

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

0.1 Markov Chain Monte Carlo Methods

This comes from Dave Miles. There are four sections.

- Introduction
- Showing the MCMC works
- Example to extract the marginal posterior distribution
- Use R to implement MCMC

This is an update of the code by Economics by Simulation
The Gibbs sampler exploits the characteristics of the Markov chain. With two parameters θ_1 and θ_2, $p(\theta_1, \theta_2)$ is the prior pdf and $L(\theta_1, \theta_2|y) = p(y|\theta_1, \theta_2$

is the likelihood function. Using Bayes theory, the posterior pdf for the parameters is

$$p(\theta_1, \theta_2|y) \propto p(\theta_1, \theta_2)L(\theta_1, \theta_2|y) \tag{3}$$

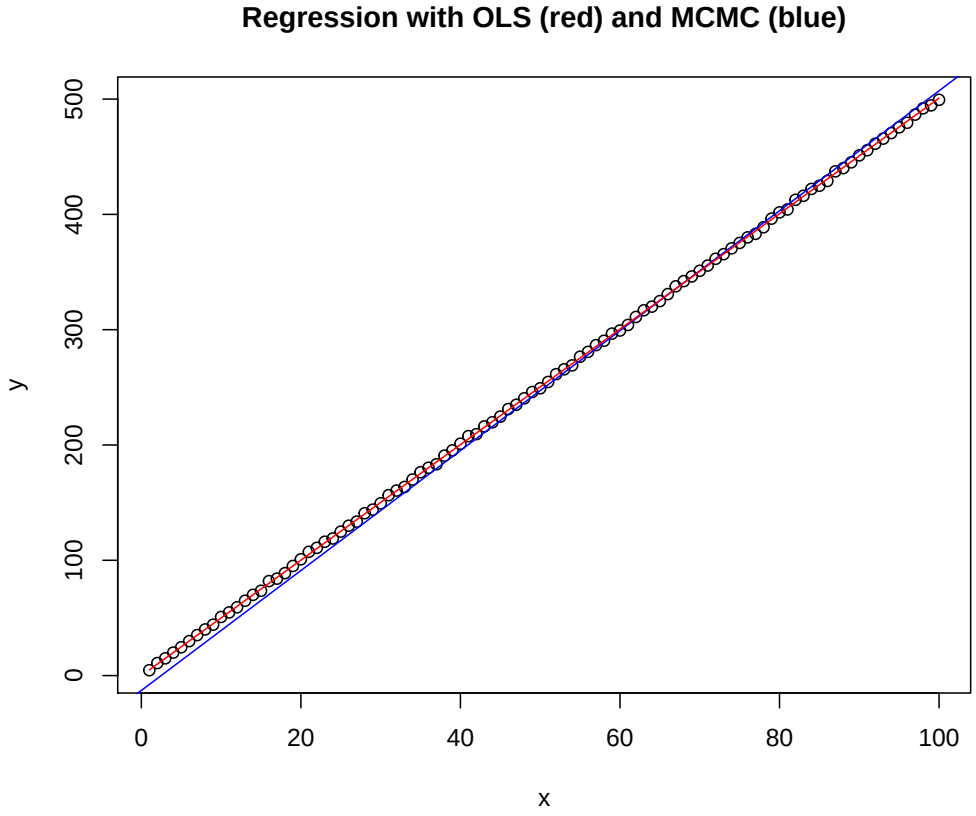
There are a numbrer of steps.

1. Assign initial values to $\theta_1^{(0)}$ and $\theta_2^{(0)}$
2. Draw a random value $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, y)$
3. The draw a random value $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, y)$
4. Draw a random value $\theta_1^{(2)}$ from $p(\theta_1|\theta_2^{(1)}, y)$
5. Repeat items 3 and 4

We end up a with two series that are Markov chains so the initial values do not matter and after many replications the chains start to behave as if they were random draws from the *marginal* posterior distibution $p(\theta_1|y)$ and $p(\theta_2|y)$ rather than the *conditional* posterior distribution $p(\theta_1|\theta_2, y)$ $p(\theta_1|\theta_2, y)$. The early values are part of the *burn in* period and should be discarded.

Here is a very simple version of the sampler applied to linear regression. A draw for the slope coefficient is made conditional upon the given intercept and then a new intercept is drawn conditional on the slope. This is repeated. I think that they should be random draws from the data.

```
# Create a simple MCMC sampler

x <- seq(1, 100, by = 1)
e <- rnorm(length(x))
y <- 0.5 + 5 * x + e
# prior
a <- rep(NA, times = length(x))
b <- rep(NA, times = length(x))
a[1] = 1
b[1] = 1
for (i in 2:100) {
    b[i] <- (y[i] - a[i - 1] - rnorm(1))/x[i]
    a[i] <- y[i] - b[i] * x[i] - rnorm(1)
}
plot(y ~ x, type = "p", main = "Regression with OLS (red) and MCMC (blue)")
abline(a = mean(a), b = mean(b), col = "blue")
lines(fitted(lm(y ~ x)), col = "red")
```



Once there is a large sample of $p(\theta_1|y)$ and the burn-in have been discarded, the mean, variance, median or mode of the marginal posterior can be calculated. The process can be expanded to more parameters.

0.1.1 Gibbs Sampler

The Gibbs sampler is a special case of the Metropolis-Hasting algorithm. The R code from Dave Giles in the file DaveGilesMCMCcoes.R has better version than this. I am not sure how we got from $\rho = 0.5$ to $sd = 1 - \rho^2$.

This example is based on two random variables Y_1 and Y_2, with a mean vector of (μ_1, μ_2) a correlation of ρ, the variances of Y_1 and Y_2 are σ_1^2 and σ_2^2 respectively and the covariance between Y_1 and Y_2 is $\rho\sigma_1\sigma_2$. [1]

The conditional distribution of Y_1 given Y_2 is

$$p(Y_1|Y_2) \sim N\left(\mu_1 + \frac{\rho\sigma_1(Y_2 - \mu_2)}{\sigma_2}, \sigma_1^2(1-\rho^2)\right) \tag{4}$$

[1] As $\rho = \frac{\sigma_{Y_1,Y_2}}{\sigma_1\sigma_2}$. Some additional reading on covariance http://www.cogsci.ucsd.edu/ desa/109/trieschmarksslides.pdf

and the conditional distribution of Y_2 given Y_1 is

$$p(Y_2|Y_1) \sim N\left(\mu_2 + \frac{\rho\sigma_2(Y_1 - \mu_1)}{\sigma_2}, \sigma_1^2(1-\rho^2)\right) \tag{5}$$

This can be used to find the marginal distribution. It is known that the marginal distribution for Y_1 is

$$p(Y_1) \sim N(\mu_1, \sigma_1^2) \tag{6}$$

and for Y_2 is

$$p(Y_2) \sim N(\mu_2, \sigma_2^2) \tag{7}$$

However, to test the MCMC draw from the conditional to find the marginal (that we already know).

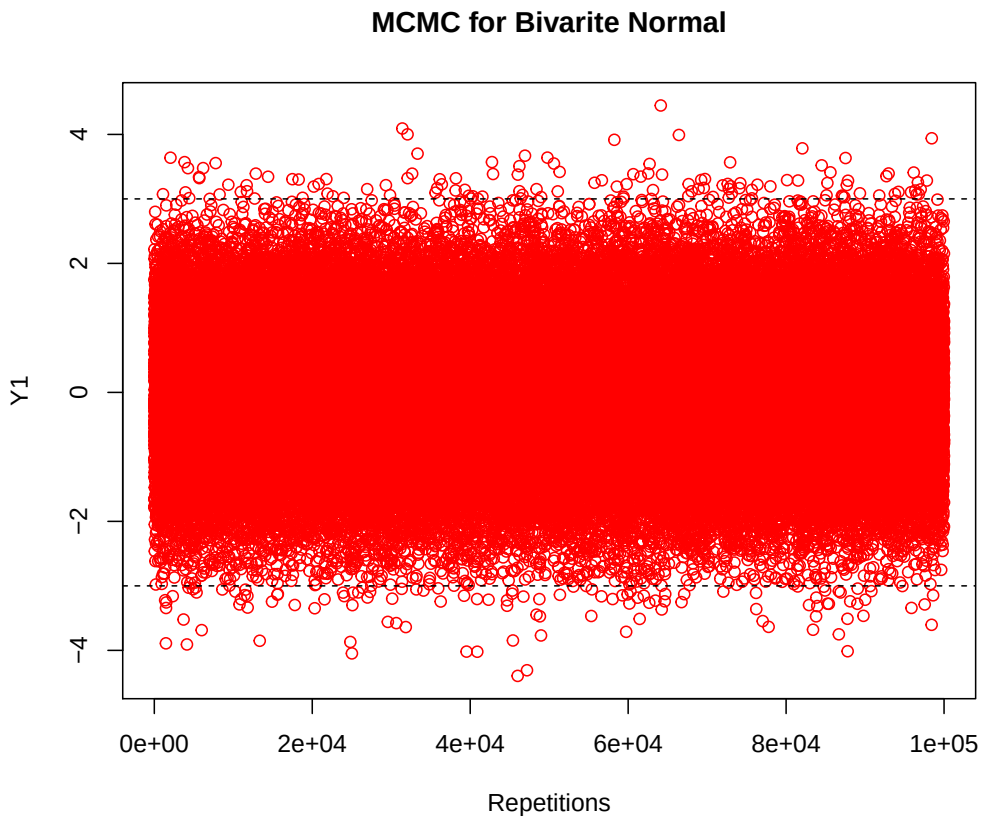
Set $\mu_1 = 0$ and $\mu_2 = 0$ and $\sigma_1 = 1$ and $\sigma_2 = 1$

The steps for the Gibbs sampler are

1. Chose an initial value for Y_1 called $Y_1^{(0)}$
2. Next, generate a random Y_2 value ($Y_2^{(0)}$) from $p(Y_2|Y_1^{(0)})$
3. Then, generte a new random Y_1 value (say $Y_1^{(1)}$) from $p(Y_1|Y_2^{(0)})$
4. Repeat steps 2 and 3 many times saving the strings of Y_1 and Y_2 values.
5. Throw away the first thousand or so values as they are draws from the *conditional distribution*
6. Now the values from the marginal distribution of interest

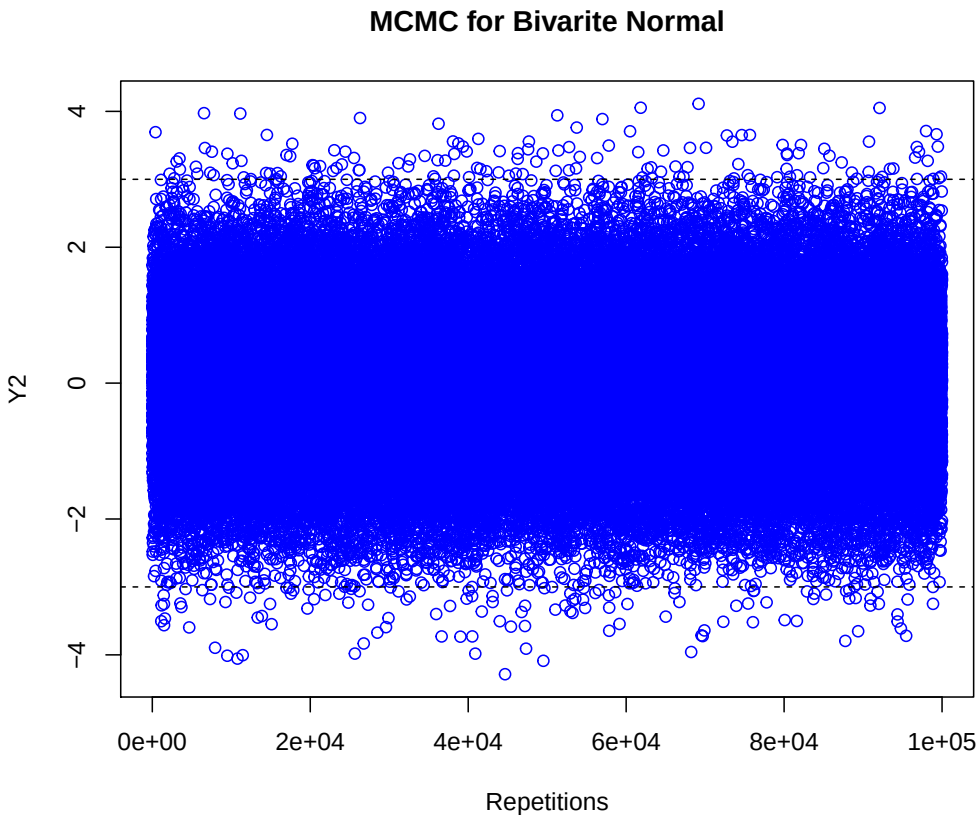
```
set.seed(123)
nreps <- 105000
nb <- 5000
yy1 <- array(, nreps)
yy2 <- array(, nreps)
rho <- 0.5
sd <- sqrt(1 - rho^2)
y1 <- rnorm(1, 0, sd)
for (i in 1:nreps) {
    y2 <- rnorm(1, 0, sd) + rho * y1
    y1 <- rnorm(1, 0, sd) + rho * y2
    yy1[i] <- y1
```

```
    yy2[i] <- y2
}
nb1 <- nb + 1
yy1b <- yy1[nb1:nreps]
yy2b <- yy2[nb1:nreps]
plot(yy1b, col = 2, main = "MCMC for Bivarite Normal", xlab = "Repetitions",
    ylab = "Y1")
abline(h = 3, lty = 2)
abline(h = -3, lty = 2)
```



The values are centered around zero as they should be.

```
plot(yy2b, col = 4, main = "MCMC for Bivarite Normal", xlab = "Repetitions",
    ylab = "Y2")
abline(h = 3, lty = 2)
abline(h = -3, lty = 2)
```

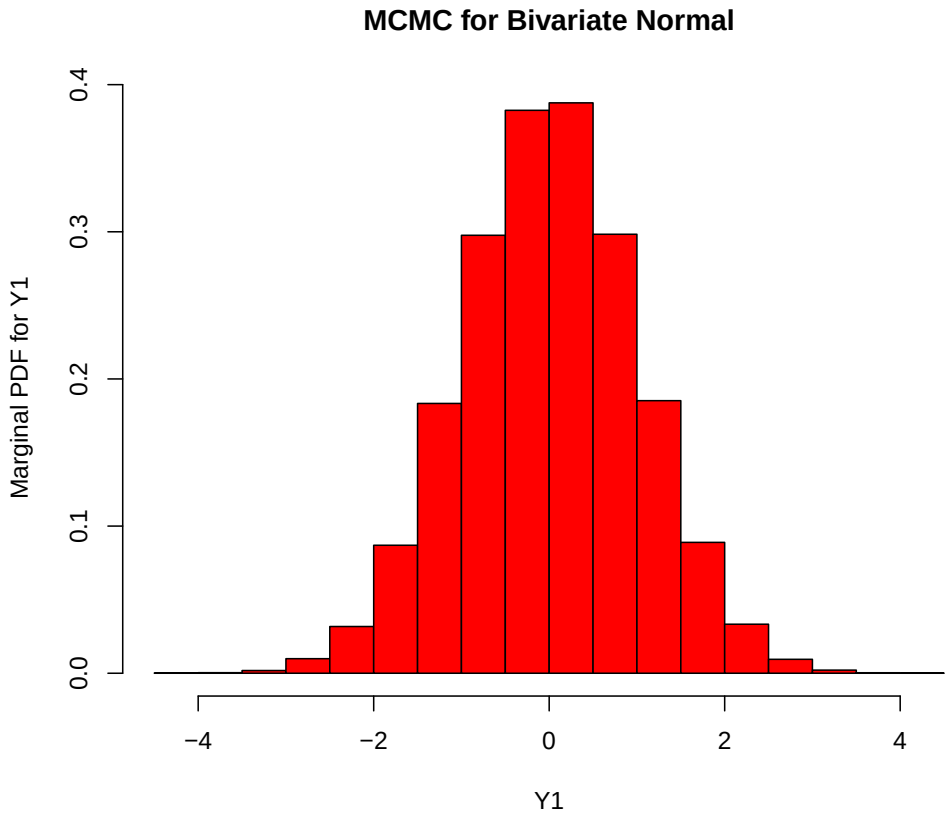


The summary statistics are below.

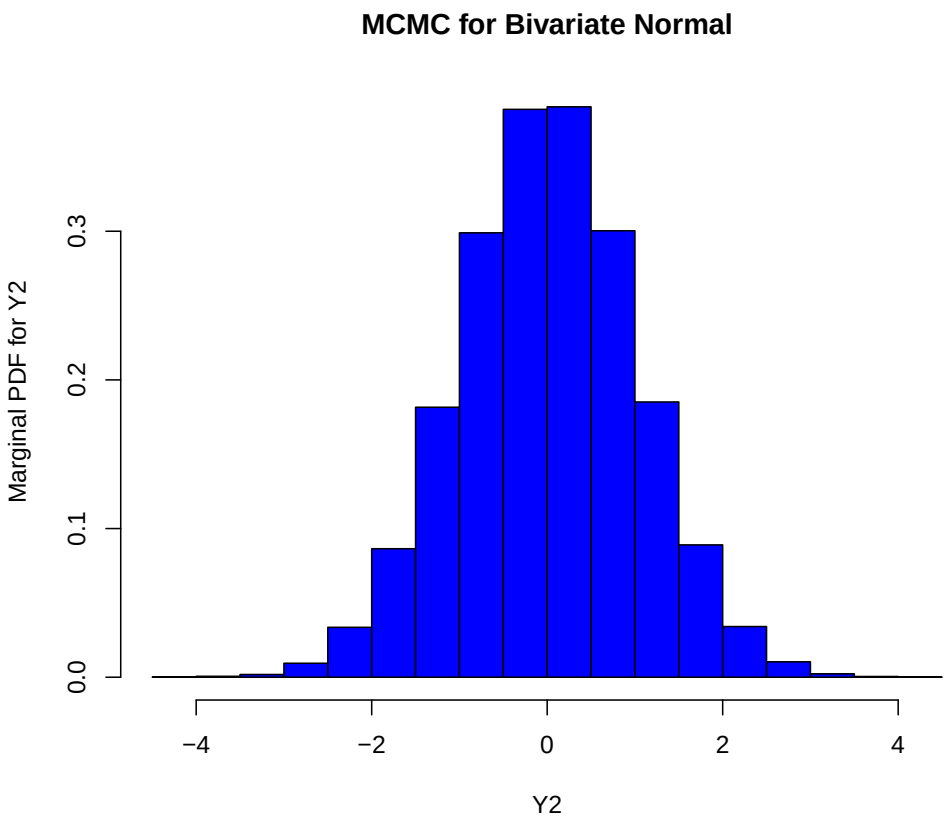
```
summary(yy1b)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.400  -0.671   0.008   0.005   0.678   4.450

var(yy1b)

## [1] 0.9954

summary(yy2b)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.280  -0.668   0.007   0.007   0.681   4.110

var(yy2b)

## [1] 1.004
```

The means are zero and the variances are close to unity.

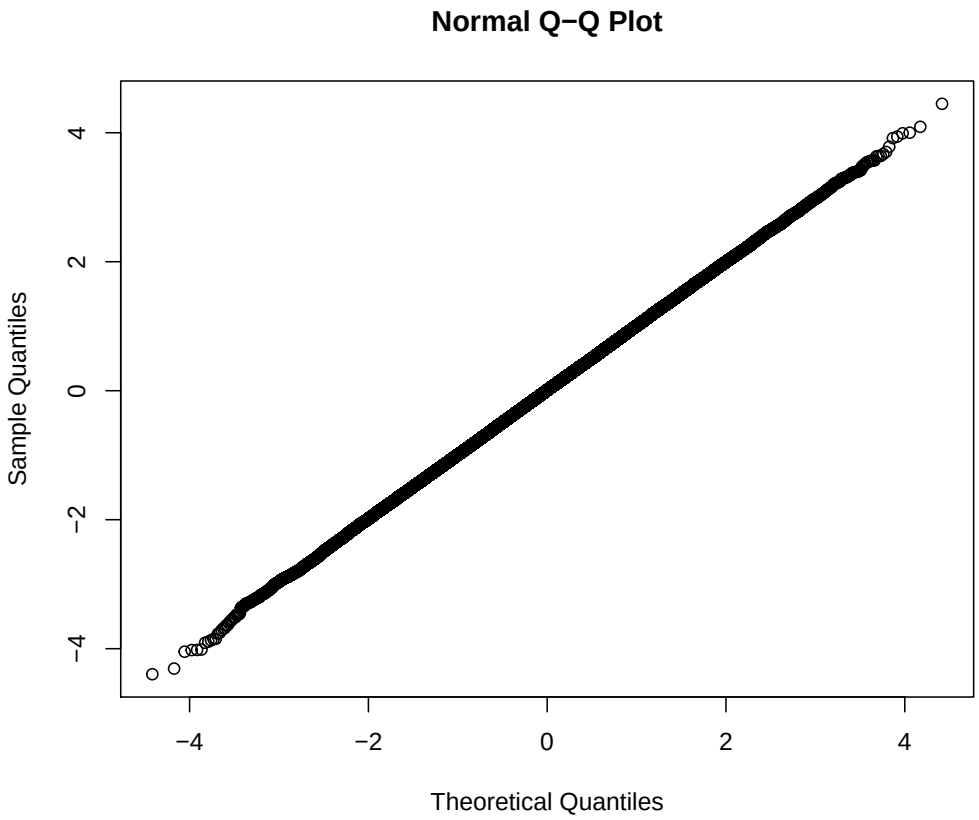
```
hist(yy1b, prob = T, col = 2, main = "MCMC for Bivariate Normal", xlab = "Y1",
    ylab = "Marginal PDF for Y1")
```



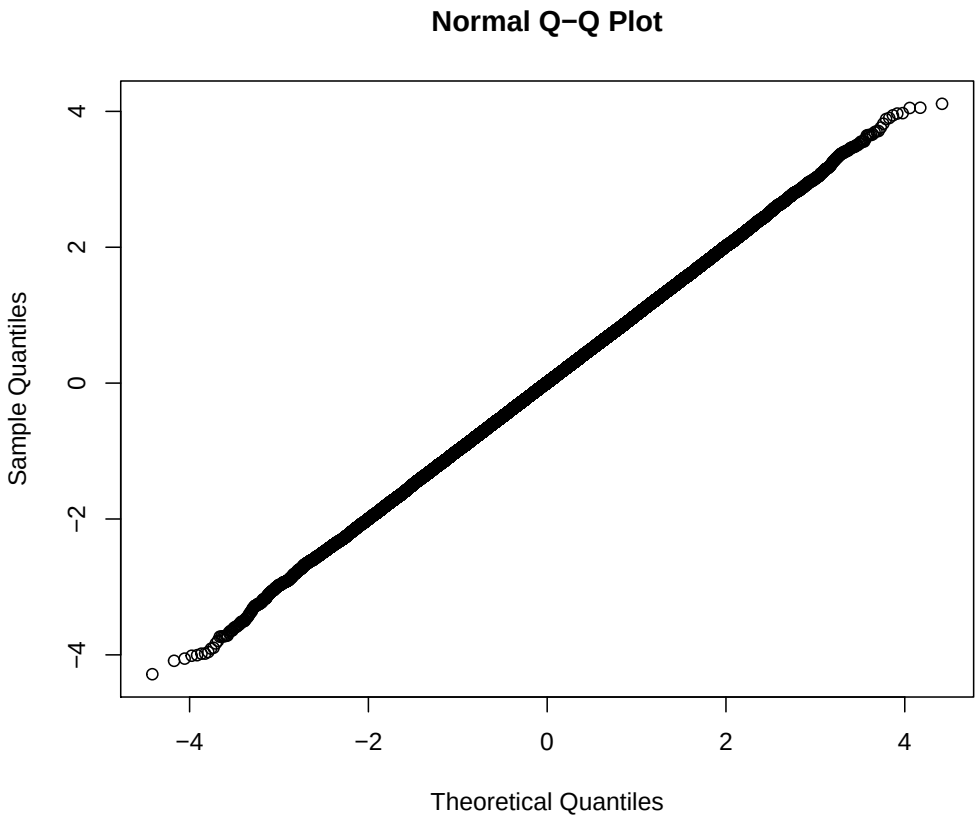

```
hist(yy2b, prob = T, col = 4, main = "MCMC for Bivariate Normal", xlab = "Y2",
    ylab = "Marginal PDF for Y2")
```


MCMC for Bivariate Normal
Marginal PDF for Y2
0.0
0.1
0.2
0.3
−4
−2
0
2
4
Y2

```
qqnorm(yy1b)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4

```
qqnorm(yy2b)
```




```
library(tseries)
```

```
## Error:  there is no package called 'tseries'
```

```
jarque.bera.test(yy1b)
```

```
## Error:  could not find function "jarque.bera.test"
```

```
jarque.bera.test(yy1b)
```

```
## Error:  could not find function "jarque.bera.test"
```

0.2 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 comonents 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 varibles 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 the 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 ($\theta_{i=1\ldots K}$); mixture weight ($\Phi_{i=1\ldots K}$); $\mathbf{\Phi}$ K-dimensional vector summing to 1; $F(x|\theta)$ probability distribution of observations parameterised on θ; α shared hyperparameter for component weights; β shared hyperparameter for mixture weights; $H(\theta|\alpha)$ prior probability distribution of component parameters;

0.3 Regression

This is a series of sessions that introduce regression with R. The first page is here. Linear Regression: Step-by-step. This is essentially a version of the Andrew Ng Machine Learning Course.

If Y is the dependent variable, $X_1, X_2, \ldots X_n$ are the explanatory variables and Θ are unknown parameters, regression takes can be carried out in a number of ways.

0.3.1 OLS

$$h_\Theta(x) = \Theta_0 + \Theta_1 x \tag{8}$$

The aim is to find values of Θ so that the model will fit the data. The *cost function* will assess the difference between the predicted and the actual values. One version of the cost function is

$$J(\Theta_0, \Theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(X^{(i)} - y^{(i)}))^2 \tag{9}$$

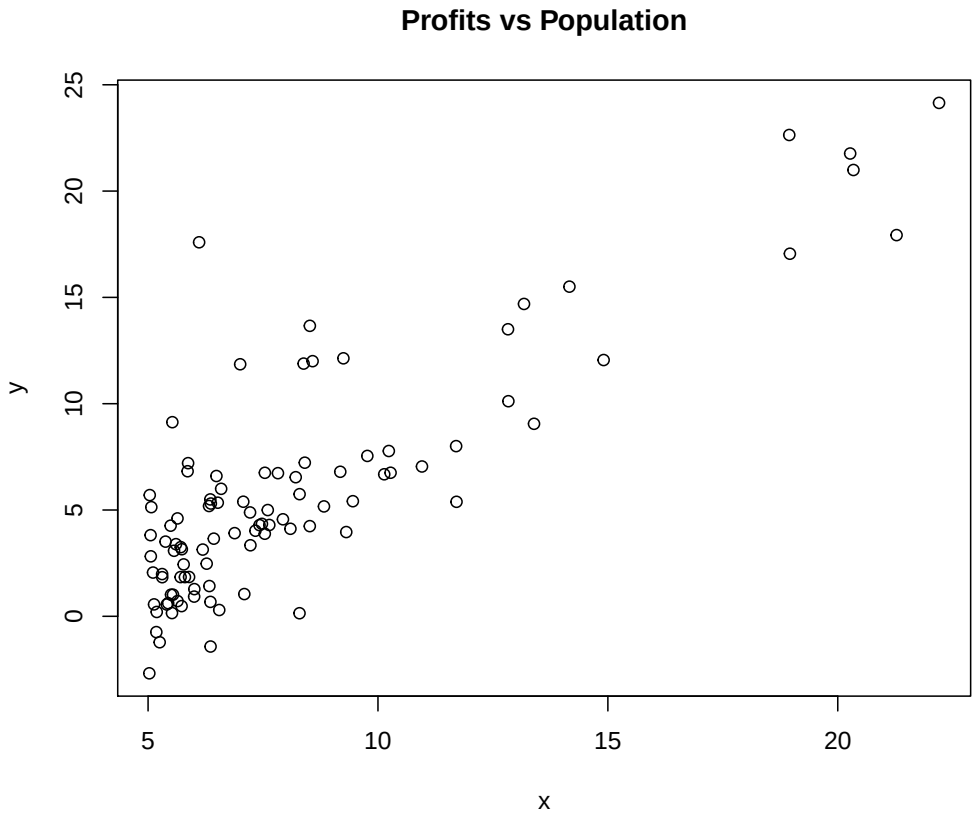
Change the values of Θ to minimise the cost. The method used is gradient decent.

The partital derivative of the cost function with respect to the Θ are

$$\Theta_0 = \Theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)}) \tag{10a}$$

$$\Theta_1 = \Theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)}) x^{(i)} \tag{10b}$$

```
data <- read.csv("../Data/Reg.csv")
y <- data$profit
x <- data$population
plot(y ~ x, main = "Profits vs Population")
```



The objective of the regression is to minimise the cost function.

$$J(\Theta_0, \Theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(X^{(i)} - y^{(i)}))^2 \tag{11}$$

Where the hypothesis is given by

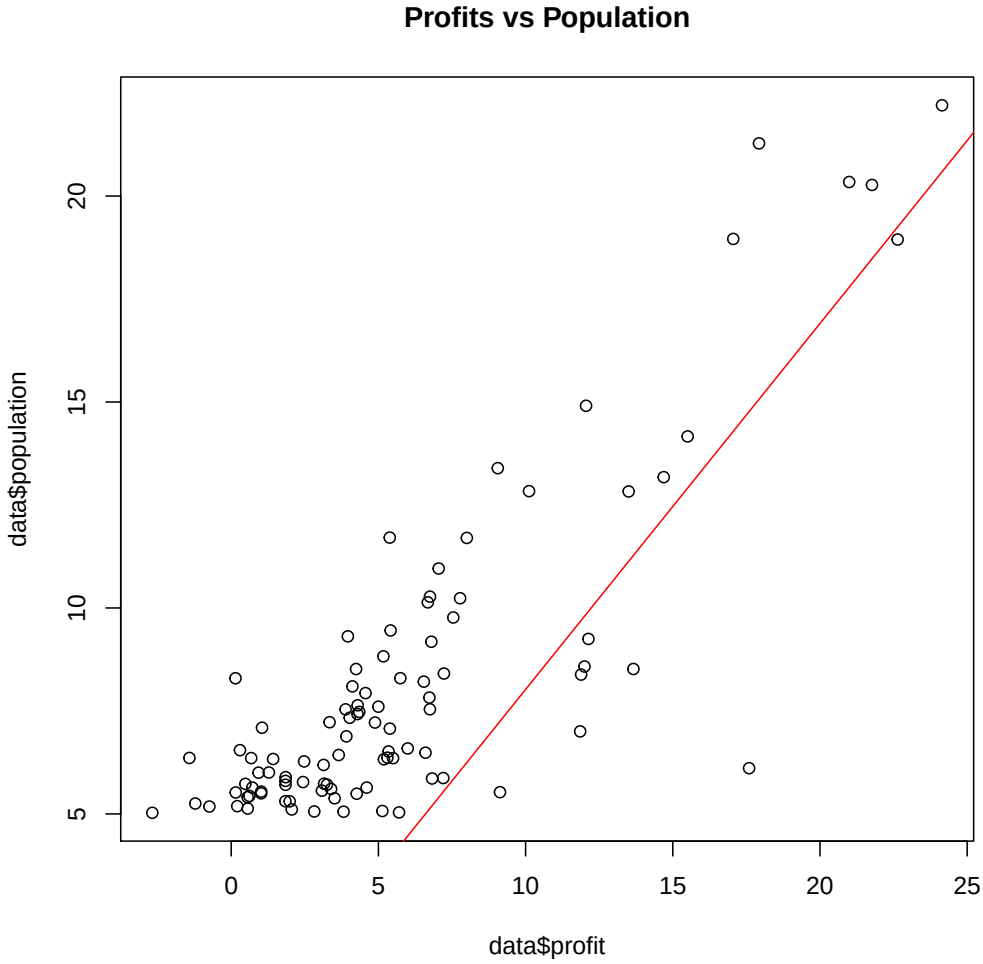
$$h_\Theta(x) = \Theta^T x = \Theta_0 + \Theta_1 x_1 \tag{12}$$

To take account of the intercept Θ_0 an additional column of constants is added to x.

```
# Add ones to x
x <- cbind(1, x)
# Initialise theta vector
theta <- c(0, 0)
# Number of observations
m <- nrow(x)
# Calculate the cost
cost <- sum(((x %*% theta) - y)^2)/(2 * m)
cost

## [1] 32.07
```

The initial value is 32.07 (how do I take this from the chunk?). Now

```
# Set learning parameter
alpha <- 0.001
# Number of iterations
iterations <- 1500
# updating thetas using gradient update
for (i in 1:iterations) {
    theta[1] <- theta[1] - alpha * (1/m) * sum(((x %*% theta) - y))
    theta[2] <- theta[2] - alpha * (1/m) * sum(((x %*% theta) - y) * x[, 2])
}
# Predict for areas of the 35,000 and 70,000 people
predict1 <- c(1, 3.5) %*% theta
predict2 <- c(1, 7) %*% theta
predict1

##       [,1]
## [1,] 2.247

predict2

##       [,1]
## [1,] 5.356

plot(data$population ~ data$profit, main = "Profits vs Population")
abline(theta, col = "red")
```



I think it is right, but....Needs to be looked at.

0.4 Adjusted R squared

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

$$R_A^2 = 1 - \left[\frac{(n-1)}{(n-k)}\right][1 - R^2] \tag{13}$$

Adding a regressor to the equation will increase (reduce)) the $R_A^"$ 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 $R_A^"$ 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

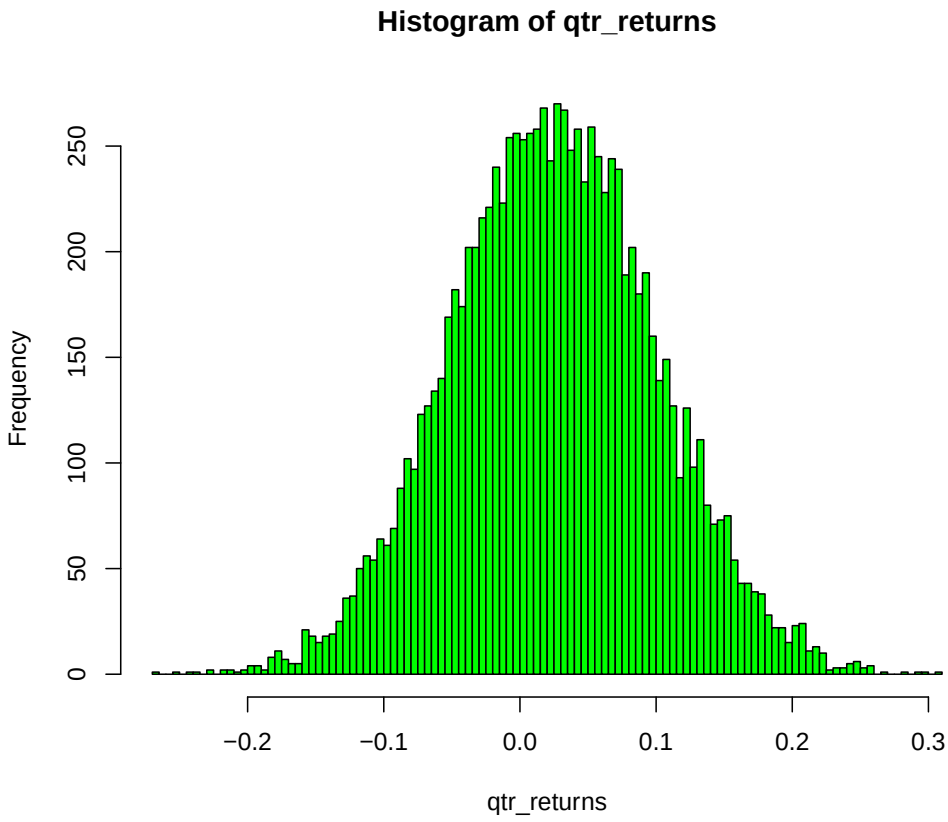
0.5 Monte Carlo Simulation

This comes from Revoluitionary Analytics. The analysis is in annual terms.

$$\mu\Delta t + \sigma Z\sqrt{\Delta t} \tag{14}$$

where μ is the drift or average annual return, Z is a standard Normal random variable, t is measured in years so for monthly returns Δt 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 μ and σ inputs as well as a transition matrix of the probabilities that there is a switch from one regime to another.

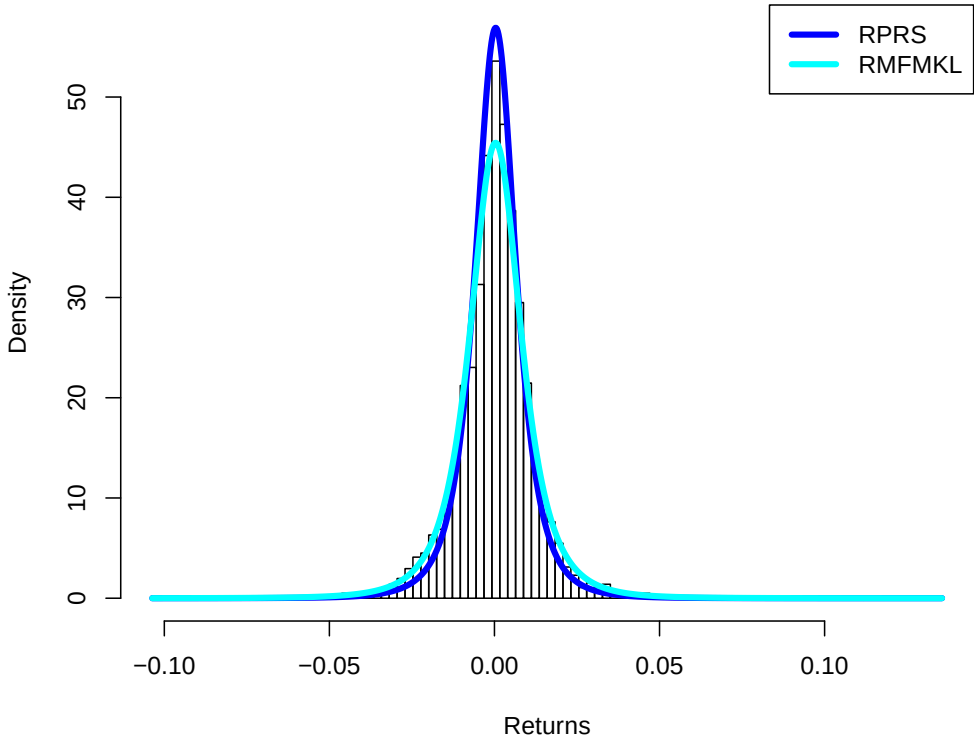
0.6 Generalised Lambda Distribution

This is from Revolutionary Analytics. The four parameters λ_1, λ_2, λ_3 and λ_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 generrate simualted 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 =
    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 FMKL 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
```

0.7 Lasso method

Rob TibshiraniCancer 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 cencerous 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.

0.8 Standard Error of the estimated mean

The standard error of the estimate of the mean is

$$SD_x = \frac{\sigma}{\sqrt{n}} \tag{15}$$

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

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

0.9 Logistic Regression

This comes from Wim-Vector - logistic regression.

```
CarData <- read.table(url("http://archive.ics.uci.edu/ml/machine-learning-databa
    sep = ",", col.names = c("buying", "maintenance", "doors", "persons", "lug_b
        "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 ***
## maintenancevhigh   -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 adde together to get the overall rating score.

The error results of the model can be examined.

```
table(CarData$rating, predict(logisticModel, type = "response") >= 0.5)
```

```
##
##         FALSE TRUE
##   acc      32  352
##   good      0   69
##   unacc  1166   44
##   vgood     0   65
```

To be completed.

0.10 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).

$$\Delta F = \frac{SS_{REG} - SS_B}{MS_W} \tag{16}$$

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.'"